

UC Merced

UC Merced Previously Published Works

Title

Long-read metagenomics of soil communities reveals phylum-specific secondary metabolite dynamics

Permalink

<https://escholarship.org/uc/item/4zv9w8s8>

Authors

Van Goethem, Marc W

Osborn, Andrew R

Bowen, Benjamin P

et al.

Publication Date

2021

DOI

10.1101/2021.01.23.426502

Peer reviewed

1 **Long-read metagenomics of soil communities reveals phylum-specific secondary**
2 **metabolite dynamics**

3 Marc W. Van Goethem¹, Andrew R. Osborn¹, Benjamin P. Bowen¹, Peter F. Andeer¹, Tami L.
4 Swenson¹, Alicia Clum², Robert Riley², Guifen He², Maxim Koriabine², Laura Sandor², Mi Yan²,
5 Chris G. Daum², Yuko Yoshinaga², Thulani P. Makhalanyane³, Ferran Garcia-Pichel^{4,5}, Axel Visel²,
6 Len A. Pennacchio², Ronan C. O'Malley^{1,2}, Trent R. Northen^{1,2*}

7 ¹ Molecular EcoSystems Biology Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd,
8 Berkeley, CA, 94720, USA

9 ² DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley,
10 CA, 94720, USA

11 ³ Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genomics and
12 Microbiology, University of Pretoria, Lynnwood Rd, Hatfield. Pretoria, 0028, South Africa

13 ⁴ Center for Fundamental and Applied Microbiomics, Biodesign Institute, Arizona State University,
14 Tempe, Arizona, USA

15 ⁵ School of Life Sciences, Arizona State University, Tempe, Arizona, USA

16 * Correspondence and requests for materials should be addressed to T.R.N. (email:
17 TRNorthen@lbl.gov)

18 **Abstract** [175 words]

19 Microbial biosynthetic gene clusters (BGCs) encoding secondary metabolites are thought to impact
20 a plethora of biologically mediated environmental processes, yet their discovery and functional
21 characterization in natural microbiomes remains challenging. Here we describe deep long-read
22 sequencing and assembly of metagenomes from biological soil crusts, a group of soil communities
23 that are rich in BGCs. Taking advantage of the unusually long assemblies produced by this
24 approach, we recovered nearly 3,000 BGCs for analysis, including 695 novel, full-length BGCs.
25 Functional exploration through metatranscriptome analysis of a 3-day wetting experiment
26 uncovered phylum-specific BGC expression upon activation from dormancy, elucidating distinct
27 roles and complex phylogenetic and temporal dynamics in wetting processes. For example, a
28 pronounced increase in BGC transcription occurs at night in cyanobacteria but not in other phyla,
29 implicating BGCs in nutrient scavenging roles and niche competition. Taken together, our results
30 demonstrate that long-read metagenomic sequencing combined with metatranscriptomic analysis
31 provides a direct view into the functional dynamics of BGCs in environmental processes and
32 suggests a central role of secondary metabolites in maintaining phylogenetically conserved niches
33 within biocrusts.

34 **Keywords:** Long-read metagenomics, secondary metabolism, metatranscriptomics, biological soil
35 crust, soil microbiome

36 **Main** [2573 words]

37 A fundamental challenge in understanding the ecological functions of secondary metabolites (also
38 known as specialized metabolites or natural products) is that most biosynthetic gene clusters
39 (BGCs) are harbored by uncultivated microbes and require specific native contexts for activation ¹.
40 The majority of BGCs encoding secondary metabolites are not usually expressed under standard
41 cultivation conditions in the laboratory ² and their products have therefore been termed ‘secondary’
42 metabolites. A universal feature of BGCs is their modular, co-localized gene architecture ³ and
43 large size, frequently spanning tens of thousands of base pairs. Bacterial secondary metabolites
44 play critical ecological roles in mediating communication, antagonistic interactions, nutrient
45 scavenging, and have historically been a primary source for antibiotic drug development ⁴; in fact
46 more than half of registered drugs are based on natural secondary metabolites ⁵. Additionally,
47 secondary metabolites have applications in agriculture ⁶, biomaterials ⁷, biofuels ⁸, and cosmetics ⁹.
48 Previous work has demonstrated the potential for deep “shotgun” metagenomic sequencing to
49 directly characterize BGCs from environmental samples ^{10, 11}, but the assembly of full-length BGCs
50 from short reads is associated with significant limitations ¹². Alternative techniques include the use
51 of clone libraries ¹ or innovative sequence-based analyses ^{13, 14} including the reconstruction of
52 uncultivated microbes as metagenome-assembled genomes (MAGs; reviewed in ¹⁵). However
53 these approaches typically only give access to dominant members of the community, while often
54 omitting members of the ‘rare biosphere’ ¹⁶.
55 We also know remarkably little about the transcription of BGCs in nature or how the environment
56 regulates their production ¹⁷ especially in soils. This information is critical in understanding how
57 often secondary metabolites are produced in natural communities. Biological soil crusts (biocrusts)
58 are the world’s most extensive biofilms and together cover up to 12% of total soil surface area ¹⁸.
59 Initial studies have suggested that they are rich in secondary metabolites ¹⁹. *Cyanobacteria*
60 dominate biocrust communities, specifically *Microcoleus* spp. that drive biocrust establishment by
61 stabilizing the soil surface, both preventing erosion and improving soil fertility through the release
62 of photosynthate ^{20, 21}. In contrast to many other types of soil environments, biocrusts are easily
63 transferable to the laboratory, which allows for controlled interrogation of relevant environmental

64 processes such as wetting dynamics. In native environments, rain events suspend microbial
65 dormancy in biocrust and cause dramatic shifts to community structure and both primary and
66 secondary metabolite release ²². The secondary metabolites produced by microbes upon wetting
67 are known to include antimicrobial compounds thought to provide a selective advantage ²³, yet the
68 majority of secondary metabolites encoded in the genomes of biocrust community members
69 remain unidentified ²⁴. *Cyanobacteria* are known secondary metabolite producers ^{3, 25} but most
70 studies have focused on aquatic cyanobacteria, leaving the secondary metabolites of terrestrial
71 cyanobacteria largely underexplored ^{26, 27}.

72 We combined long- and short-read metagenomic sequencing to produce ultra-large assemblies
73 that enabled BGC discovery. We then mapped time-series metatranscriptomes to gain insight into
74 the environmental cues governing BGC expression in biocrusts. Our results showed that
75 thousands of gene clusters could be extracted from assembled long-read metagenomes which
76 gave insight into the secondary metabolism of both rare and dominant microbial taxa. Coupling
77 these results to metatranscriptomics indicated that most BGCs were transcribed after a simulated
78 rain event, and that cyanobacteria dominated secondary metabolism.

79

80 **Long-read sequencing permits access to ultra-long gene clusters**

81 Biocrust samples were collected from Moab, UT, USA (Fig. 1a), and transported to the JGI in petri
82 dishes that maintain the physical structure of the crust. We then extracted and prepared high-
83 molecular-weight DNA was extracted and prepared from intact biocrust samples for both long- and
84 short-read metagenomic sequencing (Fig. S1). In total, we sequenced eight SMRT cells from three
85 libraries yielding 156.3 Gb from 36.7 million reads, where half of all sequenced bases were
86 contained in reads of 5 kb or longer, while the longest read was 167 kb. The average read length
87 was 3,084 bp while the mean *N50* value was 4,070 bp. Both statistics were augmented by the
88 Sequel II library which comprised 108 Gb of sequence in just 19.1 million reads.

89 The 2 short-read Illumina libraries provided an additional 20 Gb of sequence (Table S1). To obtain
90 an initial phylogenetic profile of the communities under investigation we performed full-length 16S

91 rRNA gene analysis using exact sequence variants (ESVs) which showed that *Cyanobacteria*, and
92 particularly *Microcoleus vaginatus*, were dominant biocrust community members, with major
93 representations of *Actinobacteria* and *Alphaproteobacteria* (Fig. 1b) which is generally consistent
94 with the known community composition of these biocrusts²⁸. Overall, the biocrust are less complex
95 than other desert soil communities²⁹ yet are notably richer in cyanobacteria.

96 To access biosynthetic gene clusters, we individually assembled the biocrust metagenomes into
97 contiguous sequences (contigs). Using both Canu³⁰ and metaFlye³¹ we assembled the long-read
98 ($n=8$ SMRT cells, 74,953 contigs, $N50 = 18.2$ kb) into assemblies that totalled 781 Mb in size, with
99 half of the sequence present in contigs longer than 20 kb. The longest contig was more than 753
100 kb in length assembled from the largest long-read metagenome (Table S2). The two short-read
101 Illumina libraries assembled into ~8 million contigs (3.7 Gb, $N50 = 1$ kb).

102 We also co-assembled the metagenomes to access even more BGC diversity than was
103 permissible from the individual assemblies. We co-assembled the five largest long-read
104 metagenomes which yielded 1.4 Gb of assembled sequence (Table S2) with the longest contig
105 exceeding 1.3 Mb in length ($N50 = 36$ kb). This co-assembly was as large as our hybrid co-
106 assembly of two short-read Illumina libraries and four long-read libraries produced with
107 metaSPAdes³² (1.7 Gb, $N50 = 2.3$ kb). Putative misassemblies identified through MetaQUAST
108 were identified and removed³³. Overall, the long-read assemblies and co-assemblies produced the
109 largest number of ultra-long contigs (>50 kb) and were thus most suited for the investigation of full-
110 length biosynthetic gene clusters. Together they gave unprecedented access to the BGCs
111 encoded by uncultivated microbes including 1,191 BGCs from the long-read co-assembled
112 metagenome.

113 Overall, the long-read metagenomes, and particularly their co-assemblies, offered substantially
114 deeper insight into biocrust secondary metabolism than was possible through short-read
115 sequencing and assembly (Fig. S2). For example, the Sequel II assembly had 548 BGCs including
116 174 full-length BGCs (i.e., the BGC was not truncated on either contig edge), while the short-read
117 assemblies had 359 BGCs between them yet only 9 full-length BGCs. In total, we predict that 712
118 BGCs are full-length clusters.

119 The single largest BGC was identified in the ultra-large co-assembly and was putatively assigned
120 to the genus *Nostoc*. It encodes a novel hybrid transAT-PKS-NRPS of 111 kb length, harboring 6
121 core biosynthetic genes and 8 additional biosynthetic genes. Manual inspection suggests it is full-
122 length, making it one of the longest BGCs to be identified directly from a soil metagenome (Fig.
123 S3). Clearly the co-assembly of multiple long-read metagenomes offers access to a deeper
124 spectrum of BGCs while the diversity of these clusters found here suggests that much secondary
125 metabolic potential remains unrealized in current databases. Moreover, the use of long-read
126 sequencing is central to finding novel full-length gene clusters, an issue that precluded the use of
127 short-read metagenomics previously.

128

129 **Thousands of novel gene clusters recovered from biocrust metagenomes**

130 We performed gene cluster identification and annotation for secondary metabolites³⁴ using all the
131 *de novo* metagenome assemblies owing to their high contiguity in assembly and high proportion of
132 contigs longer than 5 kb (Fig. 1c). This approach recovered 2,988 biosynthetic gene clusters
133 (BGCs) predicted to produce secondary metabolites from uncultivated biocrust microbes. These
134 span all major secondary metabolite classes with terpenes, ribosomally synthesized and post-
135 translationally modified peptides (RiPPs) and non-ribosomal peptide synthetases (NRPSs)
136 particularly well represented. Cyanobacteria were rich in NRPSs and Type 1 polyketide synthases
137 (PKS) and harbored the most BGCs overall encoding some 1,470 BGCs (Fig. 1d; Table S3). Four
138 hundred and twenty of these non-redundant BGCs could be assigned to the genus *Microcoleus* –
139 the pioneer microbial guild of biocrust³⁵. Next, we determined the genetic novelty of our BGCs by
140 evaluating whether previous sequencing efforts had captured the sequence by making queries to
141 the entire NCBI nt sequence database (accessed December 6, 2019³⁶). Using thresholds of 75%
142 sequence identity over 80% of the sequence length³⁷ we identified 175 BGCs that had been
143 sequenced previously. Thus ~94% of BGCs had not been sequenced before. This reaffirms that
144 biocrust are a rich source of BGCs and underscores the potential for long-read metagenomic
145 sequencing in novel BGCs discovery.

146 Of the known clusters, 143 belonged to Cyanobacteria including the late branching genera
147 *Microcoleus*, *Nostoc*, and *Oscillatoria*. BGCs identified from non-cyanobacterial contigs had
148 interesting novel elements. For example, *Planctomycetes* were rich in acyl-amino acids, while
149 *Alphaproteobacteria* had unusually high numbers of the dipeptide N-acetylglutaminyglutamine
150 amides (NAGGN) as well as N-acyl-homoserine lactones that may be involved in quorum sensing
151 ³⁸. Moreover, many terpenes and Type 3 PKS belonged to the dominant heterotrophic phyla
152 *Alphaproteobacteria* and *Actinobacteria* (Fig. 1e). We also found 17 phenazines in our dataset,
153 some of which may have functions in redox balance during anoxia ³⁹, most of which belonged to
154 cyanobacteria.

155

156 **Constitutive transcription of secondary metabolite gene clusters**

157 Desert biocrust communities are sensitive to rain events, as revealed by dramatic changes in
158 microbial community structure ²⁸ and core gene expression by DNA microarray ^{22, 40}. To identify
159 secondary metabolite BGCs involved in these dynamics, we mapped 13 biocrust
160 metatranscriptomes to our metagenome assemblies. The metatranscriptomes are from a simulated
161 rain event in the laboratory using intact biocrust from the same site (Moab, UT, USA) ⁴⁰. They
162 capture microbial transcription following a wetting event for three diurnal cycles at a resolution of
163 10 individual timepoints. Like the metagenomic data, 16S rRNA transcript analysis using ESVs
164 from the metatranscriptomic datasets revealed an abundance of transcripts from *Cyanobacteria*,
165 and especially *Microcoleus vaginatus* at all timepoints (Fig. 2a). We observed a dramatic increase
166 in 16S rRNA transcript copy numbers across all taxa 15 minutes and 1 hour after wetting possibly
167 indicating increased microbial growth on substrates released during cell membrane
168 permeabilization after wetting ⁴¹ or simply ribosome synthesis as microbes emerge from dormancy.

169 The metatranscriptomic data comprised 137 Gb of high-quality sequence in 919 million transcripts
170 from 13 samples (Tables S1, S2, S4). To calculate secondary metabolite gene transcription after
171 wetting we mapped the individual read transcripts to each contig containing a BGC using BMap ⁴²
172 which leveraged our long contigs to profile transcription for almost 3,000 secondary metabolite

173 gene clusters. Remarkably, we found that 395 genes from 240 BGCs were transcribed at all
174 timepoints (using a threshold of at least five mapped transcripts per gene within a cluster at a
175 single time point), which represent some 6% of all secondary metabolic genes in our dataset (Fig.
176 2b). Our results show stark contrast to previous observations that BGC expression in the
177 laboratory is low wherein most secondary metabolites are not transcribed². Their constitutive
178 expression supports the notion that “secondary” metabolites may play critical (and possibly
179 essential) roles in communication or niche occupancy in these ecosystems. Given the relatively
180 high biosynthetic cost of synthesizing secondary metabolites vs. primary metabolites⁴³ this
181 suggests that these compounds provide fitness benefits to their hosts across the wetting event.

182 Next, we investigated how the observed constitutive expression of secondary metabolic genes
183 compared to the transcription of all other genes, i.e., those not involved in secondary metabolism.
184 Of these 966,111 ‘non-secondary’ genes, just 43,139 (some ~4.5%) were constitutively transcribed
185 at all 10 time points. These mapping rates were not artifacts of gene length differences between
186 primary genes and secondary metabolic gene lengths (Supplementary Results and Fig. S4). We
187 then focused on core 46 metabolic bacterial genes that we expect to have high constitutive
188 expression e.g., those encoding DNA-binding or ribosomal subunit proteins (Table S5), and found
189 that indeed many of these core genes were transcribed at eight or more time points and 18% that
190 were constitutively transcribed (5 mapped transcripts at all 10 timepoints; Fig. 2b). This same
191 analysis of secondary metabolic genes showed a more even distribution across the time points
192 with 6% transcribed at all 10 time points (Fig. 2b). Although lower than for core bacterial genes,
193 this represents a higher proportion of constitutive transcription for secondary metabolic genes than
194 was anticipated.

195 While our results show low level constitutive transcription of many BGCs, the highest level of BGC
196 transcriptional activity occurred at night, 11.5 hours after the initial wetting event (Fig. S5a). This
197 enrichment in transcription was mostly underpinned by a surge in transcriptional activity by the
198 *Cyanobacteria* (Fig. S5b) which likely corresponds to gene induction at night when they are not
199 photosynthetically active⁴⁰. Strikingly, 80% of cyanobacterial BGC transcription peaked at night
200 (2,527 of 3,173 genes). This included the significant transcription of two putative siderophore-

201 producing BGCs (DESeq2: $P < 0.05$), while their observed rearrangements were presumably
202 driven by transposases (Fig. 2c, Fig. S8 and Supplementary Results).

203 We next examined phylogenetic conservation of BGC expression among phyla. Here, we analysed
204 a subset of biosynthetic genes individually ($n=12,470$ genes) using t-SNE visualization ⁴⁴. This
205 analysis revealed segregation of biosynthetic gene transcription by taxonomy (Fig. 3a).
206 Cyanobacterial transcription of secondary metabolites was significantly unlike all other phyla
207 (Pearson's $r > 0.8$; adjusted $P < 0.05$). While Cyanobacteria exhibited the highest level of BGCs
208 transcription at night, 11.5 hours after wetting, other bacteria (in this case almost exclusively
209 heterotrophic guilds) showed maximal BGC transcription during the day (Fig. S5). Notably there
210 was a peak of transcriptional activity 72 hours after wetting (during the day, and the point of dry
211 down) which was due to the increased transcription of terpenes and Type3 PKSs by abundant
212 heterotrophic bacteria such as *Deltaproteobacteria* and *Actinobacteria* (Fig. S6, S7b).

213 Given the conserved phylogenetic signal level at the gene level, we also examined phylogenetic
214 conservation at the cluster level. Here we compare the degrees to which biosynthetic gene clusters
215 shared similar transcriptional profiles across phyla using a co-occurrence network based on the
216 average Z-scores of each BGCs transcription ($n=2,988$). This analysis revealed clustering of
217 secondary metabolite transcription of entire BGCs by taxonomy. Namely, the bacterial phyla had
218 distinct temporal signatures of BGC transcription compared to each other over the course of 3 days
219 (Fig. 3b). Cyanobacterial BGC expression was distinct from all other bacterial groups in the
220 biocrust (Fig. 3c; $P < 0.05$). To our knowledge, this is the first such observation of phylum-level
221 differences in microbial BGC transcription in natural communities. This may reflect conservation of
222 life history traits especially niche competition strategies. For example cyanobacteria can grow
223 heterotrophically on diverse dissolved organic components ⁴⁵ and increased BGC expression may
224 reflect increased competition with heterotrophs occurring at night. Thus, at night *Microcoleus* and
225 other cyanobacteria may produce antibiotics to antagonize heterotrophs competing for dissolved
226 organic compounds ⁴⁶.

227 In addition to antagonism, night-time expression of BGC products can facilitate electron and
228 nutrient transport. Redox-active secondary metabolites are known to be produced by microbes

229 under anoxic conditions ³⁹. For example, *Pseudomonas aeruginosa* enhances substrate-level
230 phosphorylation during anoxia through the production of phenazines that facilitate electron
231 transport ⁴⁷. Constitutive expression of the siderophore-producing gene clusters in cyanobacteria
232 may reflect cation import strategies (notably iron scavenging) needed to support photosynthesis
233 and other metabolic activities (Supplementary Results).

234

235 **Conclusion**

236 In this study we show that long-read metagenomic sequencing is a powerful new tool for the
237 examination of secondary metabolite gene clusters directly from complex environmental samples.
238 Integration with metatranscriptomics revealed that ~6% of secondary metabolic genes were
239 constitutively transcribed over 3 days – a higher percentage than other genes. Thus, while
240 conventionally unexpressed under laboratory conditions, our results show that *in situ* BGCs appear
241 to control important life history traits involved in maintaining microbial niches. BGC expression
242 showed strong phylogenetic conservation where *Cyanobacteria*, unlike other phyla, exhibited the
243 highest levels of transcription at night. We speculate that this may reflect the switch from
244 cyanobacteria serving as primary producers during the day to competing with heterotrophs for
245 dissolved organics at night.

246 **Materials and Methods**

247 *Biocrust Sample Collection and DNA isolation*

248 Biological soil crust (biocrust) was collected from Green Butte Site near Moab, UT, USA
249 (38°42'54.1"N, 109°41'27.0"W) in 2014 as described previously²². This field site is part of a long-
250 term ecological research area of scientific interest aimed at exploring climatic changes in arid
251 regions. We sampled early maturity biocrust (*Microcoleus*-dominated) by coring directly into the
252 soil surface with a petri dish (6 cm² by 1 cm in depth). Samples were maintained in petri dishes in a
253 dark desiccator in the laboratory until required for DNA isolation. Metagenomic DNA was isolated
254 using the MoBio Powersoil kit as per the manufacturer's instructions with a minor modification. We
255 extracted DNA from 2 g of crust material by dividing the sample into four separate tubes (0.5 g in
256 each tube). The nucleic acids from each tube were eluted in 50 µl of elution buffer and then pooled
257 these into a final sample containing 200 µl of elution buffer and DNA.

258 *SMRT Sequencing*

259 We sequenced three SMRT cells on the PacBio RS II Single Molecule, Real-Time (SMRT®) DNA
260 Sequencing System (Pacific Biosciences, CA, USA) using two different library inserts: 10 kb
261 AMPure PB library [*n*=2] and a Low input 3 kb PB library [*n*=1] using binding kit P6 v2 with 360-
262 minute and 120-minute movies for the respective libraries. The same libraries were then
263 sequenced on a PacBio Sequel System (Pacific Biosciences) using Sequel Binding Kit 2.1 with a
264 combination of 600- and 1200-minute movies. A third library was made using 10 kb AMPure PB
265 approach with a Blue Pippin size cutoff of 4.5 kb. It was sequenced on PacBio Sequel II System
266 (Pacific Biosciences) using 1.0 template prep kit and a 900-minute movie.

267 To test how well-suited long-read metagenomes are for BGC recovery, we made use of five
268 publicly available PacBio SMRT metagenomes including a biogas reactor library sequenced on the
269 PacBio RS II System with a 2 kb insert length⁴⁸, and four metagenomes obtained from Lake Biwa,
270 Japan that were sequenced on a PacBio Sequel System with a 4 kb insertion length⁴⁹. Raw
271 sequence statistics for each metagenome is provided in Table S1. We analysed the sequencing
272 effort of the metagenomes using Nonpareil v3.30⁵⁰ which relies on read redundancy. We

273 performed a similar comparison using publicly-available long-read metagenomes which also
274 yielded improvements in contig sizes and BGC yield from co-assembled datasets (Supplementary
275 Material).

276 *Illumina Sequencing*

277 Two unamplified 300 bp Illumina libraries were generated and sequenced 2x150 bp on the HiSeq-
278 2500 1TB platform (Illumina).

279 *Taxonomy*

280 We extracted prokaryotic 16S rRNA genes using SortMeRNA 2.1b⁵¹. These 16S rRNA sequences
281 were then analysed using DADA2⁵² to identify exact sequence variants (ESVs) under default
282 parameters with the exceptions of truncLen (150) and maxEE (1). The ESVs were then assigned
283 taxonomy against the entire SILVA 16S rRNA gene reference database⁵³. The taxonomy of the
284 identified gene clusters was inferred by BLAST queries³⁶ against the NCBI nr-database whereby
285 hits were retained with E-values of less than 1×10^{-10} and bit scores greater than 60.

286 *Assembly*

287 We performed read correction, trimming and assembly for the three RS II SMRT cells with Canu
288 v1.8³⁰. Here we included parameters suggested by the developers of Canu for PacBio
289 metagenomes including an estimated mean genome size of 5 Mb (genomeSize=5m). We also
290 changed the following parameters from their default values: corMinCoverage=0,
291 corOutCoverage=all, corMhapSensitivity=high, correctedErrorRate=0.105,
292 corMaxEvidenceCoverageLocal=10 and corMaxEvidenceCoverageGlobal=10.

293 The four larger Sequel metagenomes were assembled using metaFlye v2.4.2 under default
294 settings with an estimated genome size of 5 Mb and the *-meta* option implemented for
295 metagenomic sequence data³¹. All Illumina sequence data were quality trimmed prior to assembly
296 using Prinseq-lite v0.20.4⁵⁴ with *--min_qual_mean* set to 20 and *-ns_max_n* set to 0 which
297 eliminates low quality reads and ambiguous bases (internal N's). We assembled the two biocrust
298 Illumina metagenomes with metaSPAdes v3.13.0³² as recommended for paired-end short read

299 length Illumina libraries⁵⁵. We also co-assembled the four Sequel libraries together (termed Flye
300 co-assembly), and then with the Sequel II library (termed Ultimate co-assembly) using metaFlye.
301 Finally, we co-assembled the four Sequel libraries with the two Illumina metagenomes using
302 metaSPAdes. Open reading frames (ORFs) of core metabolic genes were predicted from the
303 assembled metagenomes using Prodigal⁵⁶ and annotated using Prokka⁵⁷ in KBase
304 (<https://kbase.us/>)⁵⁸. All assemblies were quality-checked using MetaQUAST³³ which precluded
305 the inclusion of misassemblies from our analysis.

306 *Biosynthetic Gene Cluster Analysis*

307 All contigs > 5 kb in length were explored for biosynthetic gene clusters (BGCs) using the
308 antiSMASH v5.0 web server under strict settings³⁴. Next, we consolidated and passed all putative
309 BGCs through BiG-SCAPE v0.0.0r and CORASON in global mode to explore the phylogenomic
310 relationships between the BGCs recovered from the 11 biocrust metagenomic datasets¹³. BiG-
311 SCAPE consolidates both antiSMASH and the MiBIG 2.0 database to support initial antiSMASH
312 predictions and so we included the entire MiBIG 2.0 database in our analysis to place our BGCs
313 among verified clusters⁵⁹.

314 To determine the genetic novelty of our BGCs we performed homology searches against the NCBI
315 nt database (downloaded December 6th, 2019) using NCBI blast+ 2.9. We only retained top hits
316 based on an E-value of 1×10^{-10} . BGCs were non-redundant (not sequenced previously and thus
317 novel) if sequences matched $\leq 80\%$ of the BGC query length and had an average of $\leq 75\%$
318 sequence identity against the database. We corroborated the taxonomic assignments using the
319 Contig Annotation Tool under default settings (CAT, v5.0.4)⁶⁰. Chemical structure predictions were
320 first created by antiSMASH v5.0.

321 *Metatranscriptomic mapping*

322 We made use of metatranscriptomes sequenced from biocrust material collected at the same
323 sampling site in Moab, Utah that were publicly-available on JGI GOLD⁴⁰. The experimental design
324 tracked the transcriptional responses of biocrust communities over two complete diurnal cycles
325 following an artificial wetting event in the laboratory with 12 hours of light followed by 12 hours of

326 dark (Table S1). The time points at which transcripts were collected include: 0 hours (immediately
327 before wet-up), 3 min, 15 min, 1 hour, 9 hours, 11.5 hours, and 18 hours after wet up, then 72
328 hours after wet up (immediately prior to dry down), then 2 hours and 3 days after dry down. The
329 11.5 hours and 18 hours samples also represent transcriptional activity at night-time while all other
330 samples captured transcription during the day.

331 Transcripts were quality-controlled using Prinseq-lite v0.20.4 as described above for the Illumina
332 data. The metatranscriptomes were then assembled using metaSPAdes. The unassembled
333 transcripts were then mapped to contigs containing BGCs using bbmap v38.73⁴². We used
334 SAMtools v1.9⁶¹ for file conversion and sorting. Mapped sequences and associated contigs were
335 then visualized within Geneious⁶². We used DESeq2 v1.28.0⁶³ in the R statistical environment
336 v3.6.3 to test which genes underwent differential expression by explicitly testing expression against
337 the control sample (0 hours). Here we tested two environmental treatments, (i) the diurnal cycling
338 regime (i.e., day to night to day) and, independently, (ii) the influence of wetting and drying.
339 Transcripts were removed that did not map at the phylum level to the 16S data or that had a
340 maximum count less than 20 in any sample. The remaining transcript levels were normalized by
341 the total counts for each sample and then multiplied by the average count across all samples.
342 Duplicate samples at the 15-minute time point and triplicate samples at the 1-hour timepoint were
343 averaged, and z-scores of normalized transcript abundance mapped to each biosynthetic gene to
344 reveal which time points showed highest gene activity. In addition, z-scores were used with t-SNE
345 (T-distributed Stochastic Neighbor Embedding) to visualize the gene transcription patterns in
346 ordinance space⁴⁴. The t-SNE implementation in sklearn (v 0.23.2) manifold module was used
347 with the following parameters: 'angle': 0.5, 'early_exaggeration': 12.0, 'init': 'random',
348 'learning_rate': 200.0, 'method': 'barnes_hut', 'metric': 'euclidean', 'min_grad_norm': 1e-07,
349 'n_components': 2, 'n_iter': 3000, 'n_iter_without_progress': 300, 'perplexity': 40, 'random_state':
350 None, 'verbose': 1.

351 **Acknowledgements**

352 This work was partially supported by funds provided by the Office of Science Early Career
353 Research Program Office of Biological and Environmental Research, of the U.S. Department of

354 Energy and by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science
355 User Facility, supported by the Office of Science of the U.S. Department of Energy under Contract
356 No. DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory. We also wish to
357 acknowledge Simon Roux, Emiley Eloë-Fadrosh and Eoin Brodie for their constructive feedback.

358 **Author contributions**

359 M.W.V.G. conducted the metagenomic analyses, formulated ideas, and wrote the manuscript.
360 A.R.O. predicted biosynthetic gene clusters, performed DESeq2 analysis and wrote the
361 manuscript. B.P.B. conducted the statistical analyses among transcriptional profiles and produced
362 figures. P.F.A. assisted with contig annotation and provided ideas. T.L.S. collected samples and
363 extracted DNA. A.C. provided access to data, collaborated with Pacific Biosciences scientists to
364 produce long-read metagenomes and offered valuable insights into the sequence data. R.R.
365 assembled the large metagenomes and assisted with transcript mapping. G.H. developed
366 technology of high-molecular weight DNA shearing and size selection optimization for library
367 construction in collaboration with Pacific Biosciences scientists. M.K. developed Pacific
368 Biosciences SMRTbell template preparation methods for HiFi 10 kb libraries from sample quality
369 assessment to library preparation including development of multiplex strategy. L.S. and M. Y.
370 optimized sequencing conditions for metagenomic sample libraries on the Pacific Biosciences RS
371 II, Sequel and Sequel II sequencer platforms in collaboration with Pacific Biosciences. C.G.D.
372 contributed sequencing platform development and operation management to enable the
373 sequencing of environmental metagenomic samples on the Pacific Biosciences sequencing
374 instruments in collaboration with Pacific Biosciences. Y.Y. developed long-read technology for
375 environmental samples from design/coordination to execution in collaboration with Pacific
376 Biosciences. T.P.M. offered logistical support and constructive feedback on the manuscript. F.G.P.
377 provided access to *Microcoleus* genomes and assisted with meaningful interpretation of the
378 cyanobacterial transcriptomics data. R.C.O. contributed overseeing the development of long-read
379 technologies including 10kb HiFi sequencing for environmental metagenomics in collaboration with
380 Pacific Biosciences and JGI staff. A.V. wrote the paper and offered substantial advice on project

381 planning. L.A.P. assisted in the interpretation of biological and genomic data. T.R.N. Designed the
382 study and wrote the manuscript.

383

384 **Conflict of interest statement**

385 All authors read and approved the final manuscript. We declare that we have no competing
386 financial interests.

387 **Data availability**

388 Raw data of the long- and short-read biocrust metagenomes can be accessed on the IMG/M
389 website (Submission ID 241874) or on the NCBI website (BioProject: PRNJNA691698). The raw
390 metatranscriptomic data are publicly-available through the JGI GOLD portal (sequence project IDs
391 1010318 – 1022409).

392 **Reference**

- 393 1. Milshteyn, A., Schneider, J.S. & Brady, S.F. Mining the metabiome: identifying novel natural
394 products from microbial communities. *Chemistry & biology* **21**, 1211-1223 (2014).
- 395 2. Rutledge, P.J. & Challis, G.L. Discovery of microbial natural products by activation of silent
396 biosynthetic gene clusters. *Nature reviews microbiology* **13**, 509-523 (2015).
- 397 3. Dittmann, E., Gugger, M., Sivonen, K. & Fewer, D.P. Natural product biosynthetic diversity and
398 comparative genomics of the cyanobacteria. *Trends in microbiology* **23**, 642-652 (2015).
- 399 4. Cragg, G.M., Kingston, D.G. & Newman, D.J. Anticancer agents from natural products. (CRC press,
400 2011).
- 401 5. Cragg, G.M. & Newman, D.J. Natural products: a continuing source of novel drug leads. *Biochimica
402 et Biophysica Acta (BBA)-General Subjects* **1830**, 3670-3695 (2013).
- 403 6. Singh, R., Kumar, M., Mittal, A. & Mehta, P.K. Microbial metabolites in nutrition, healthcare and
404 agriculture. *3 Biotech* **7**, 15 (2017).
- 405 7. Bohlmann, J. & Keeling, C.I. Terpenoid biomaterials. *The Plant Journal* **54**, 656-669 (2008).
- 406 8. Kang, A. & Lee, T.S. in *Biotechnology for Biofuel Production and Optimization* 35-71 (Elsevier, 2016).
- 407 9. Nowruzzi, B., Sarvari, G. & Blanco, S. The cosmetic application of cyanobacterial secondary
408 metabolites. *Algal Research* **49**, 101959 (2020).
- 409 10. Crits-Christoph, A., Diamond, S., Butterfield, C.N., Thomas, B.C. & Banfield, J.F. Novel soil bacteria
410 possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**, 440-444 (2018).
- 411 11. Sharrar, A.M. et al. Bacterial secondary metabolite biosynthetic potential in soil varies with phylum,
412 depth, and vegetation type. *Mbio* **11** (2020).
- 413 12. Libis, V. et al. Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence
414 network analysis of targeted sequences. *Nature communications* **10**, 1-9 (2019).
- 415 13. Navarro-Muñoz, J.C. et al. A computational framework for systematic exploration of biosynthetic
416 diversity from large-scale genomic data. *Biorxiv*, 445270 (2018).
- 417 14. Sugimoto, Y. et al. A metagenomic strategy for harnessing the chemical repertoire of the human
418 microbiome. *Science* **366** (2019).
- 419 15. Katz, M., Hover, B.M. & Brady, S.F. Culture-independent discovery of natural products from soil
420 metagenomes. *Journal of industrial microbiology & biotechnology* **43**, 129-141 (2016).
- 421 16. Lynch, M.D. & Neufeld, J.D. Ecology and exploration of the rare biosphere. *Nature Reviews
422 Microbiology* **13**, 217-229 (2015).
- 423 17. Amos, G.C. et al. Comparative transcriptomics as a guide to natural product discovery and
424 biosynthetic gene cluster functionality. *Proceedings of the National Academy of Sciences* **114**,
425 E11121-E11130 (2017).
- 426 18. Rodriguez-Caballero, E. et al. Dryland photoautotrophic soil surface communities endangered by
427 global change. *Nature Geoscience* **11**, 185-189 (2018).
- 428 19. Reddy, B.V.B. et al. Natural product biosynthetic gene diversity in geographically distinct soil
429 microbiomes. *Applied and environmental microbiology* **78**, 3744-3752 (2012).
- 430 20. Starkenburg, S.R. et al. (Am Soc Microbiol, 2011).
- 431 21. Belnap, J., Weber, B. & Büdel, B. in *Biological soil crusts: an organizing principle in drylands* 3-13
432 (Springer, 2016).
- 433 22. Swenson, T.L., Karaoz, U., Swenson, J.M., Bowen, B.P. & Northen, T.R. Linking soil biology and
434 chemistry in biological soil crust using isolate exometabolomics. *Nature communications* **9**, 1-10
435 (2018).
- 436 23. Adamek, M., Spohn, M., Stegmann, E. & Ziemert, N. in *Antibiotics* 23-47 (Springer, 2017).
- 437 24. Couradeau, E. et al. Bacteria increase arid-land soil surface temperature through the production of
438 sunscreens. *Nature communications* **7**, 1-7 (2016).
- 439 25. Martins, T.P. et al. Chemistry, bioactivity and biosynthesis of cyanobacterial alkylresorcinols.
440 *Natural Product Reports* **36**, 1437-1461 (2019).
- 441 26. Luesch, H., Moore, R.E., Paul, V.J., Mooberry, S.L. & Corbett, T.H. Isolation of dolastatin 10 from the
442 marine cyanobacterium *Symploca* species VP642 and total stereochemistry and biological
443 evaluation of its analogue symprostatin 1. *Journal of Natural Products* **64**, 907-910 (2001).

- 444 27. Chrapusta, E. et al. Microcystins and anatoxin-a in Arctic biocrust cyanobacterial communities.
445 *Toxicon* **101**, 35-40 (2015).
- 446 28. Van Goethem, M.W., Swenson, T.L., Trubl, G., Roux, S. & Northen, T.R. Characteristics of Wetting-
447 Induced Bacteriophage Blooms in Biological Soil Crust. *Mbio* **10** (2019).
- 448 29. Makhalanyane, T.P. et al. Microbial ecology of hot desert edaphic systems. *FEMS microbiology*
449 *reviews* **39**, 203-221 (2015).
- 450 30. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and
451 repeat separation. *Genome research* **27**, 722-736 (2017).
- 452 31. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs.
453 *Nature Methods* **17**, 1103-1110 (2020).
- 454 32. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P.A. metaSPAdes: a new versatile metagenomic
455 assembler. *Genome research* **27**, 824-834 (2017).
- 456 33. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies.
457 *Bioinformatics* **32**, 1088-1090 (2016).
- 458 34. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic*
459 *acids research* **47**, W81-W87 (2019).
- 460 35. Garcia-Pichel, F., Loza, V., Marusenko, Y., Mateo, P. & Potrafka, R.M. Temperature drives the
461 continental-scale distribution of key microbes in topsoil communities. *Science* **340**, 1574-1577
462 (2013).
- 463 36. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool.
464 *Journal of molecular biology* **215**, 403-410 (1990).
- 465 37. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nature biotechnology*, 1-11 (2020).
- 466 38. Fuqua, C. & Greenberg, E.P. Listening in on bacteria: acyl-homoserine lactone signalling. *Nature*
467 *reviews Molecular cell biology* **3**, 685-695 (2002).
- 468 39. Ciemniecki, J.A. & Newman, D.K. The Potential for Redox-Active Metabolites To Enhance or Unlock
469 Anaerobic Survival Metabolisms in Aerobes. *Journal of Bacteriology* **202** (2020).
- 470 40. Rajeev, L. et al. Dynamic cyanobacterial response to hydration and dehydration in a desert
471 biological soil crust. *The ISME journal* **7**, 2178-2191 (2013).
- 472 41. Felde, V.J.M.N.L., Peth, S., Uteau-Puschmann, D., Drahorad, S. & Felix-Henningsen, P. Soil
473 microstructure as an under-explored feature of biological soil crust hydrological properties: case
474 study from the NW Negev Desert. *Biodiversity and conservation* **23**, 1687-1708 (2014).
- 475 42. Bushnell, B. (Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014).
- 476 43. Donia, M.S., Ruffner, D.E., Cao, S. & Schmidt, E.W. Accessing the hidden majority of marine natural
477 products through metagenomics. *ChemBioChem* **12**, 1230-1236 (2011).
- 478 44. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine*
479 *Learning Research* **15**, 3221-3245 (2014).
- 480 45. Baran, R. et al. Exometabolite niche partitioning among sympatric soil bacteria. *Nature*
481 *communications* **6**, 8289 (2015).
- 482 46. Kupriyanova, E.V. et al. Extracellular β -class carbonic anhydrase of the alkaliphilic cyanobacterium
483 *Microcoleus chthonoplastes*. *Journal of Photochemistry and Photobiology B: Biology* **103**, 78-86
484 (2011).
- 485 47. Hernandez, M. & Newman, D. Extracellular electron transfer. *Cellular and Molecular Life Sciences*
486 *CMLS* **58**, 1562-1571 (2001).
- 487 48. Frank, J.A. et al. Improved metagenome assemblies and taxonomic binning using long-read circular
488 consensus sequence data. *Scientific reports* **6**, 1-10 (2016).
- 489 49. Hiraoka, S. et al. Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in
490 an environmental prokaryotic community. *Nature communications* **10**, 1-10 (2019).
- 491 50. Rodriguez-R, L.M., Gunturu, S., Tiedje, J.M., Cole, J.R. & Konstantinidis, K.T. Nonpareil 3: fast
492 estimation of metagenomic coverage and sequence diversity. *MSystems* **3** (2018).
- 493 51. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in
494 metatranscriptomic data. *Bioinformatics* **28**, 3211-3217 (2012).
- 495 52. Callahan, B.J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature*
496 *methods* **13**, 581-583 (2016).

- 497 53. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and
498 web-based tools. *Nucleic acids research* **41**, D590-D596 (2012).
- 499 54. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets.
500 *Bioinformatics* **27**, 863-864 (2011).
- 501 55. Van der Walt, A.J. et al. Assembling metagenomes, one community at a time. *BMC genomics* **18**, 1-
502 13 (2017).
- 503 56. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification.
504 *BMC bioinformatics* **11**, 119 (2010).
- 505 57. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
- 506 58. Arkin, A.P. et al. KBase: the United States department of energy systems biology knowledgebase.
507 *Nature biotechnology* **36**, 566 (2018).
- 508 59. Kautsar, S.A. et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic
509 acids research* **48**, D454-D458 (2020).
- 510 60. von Meijenfeldt, F.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H. & Dutilh, B.E. Robust taxonomic
511 classification of uncharted microbial sequences and bins with CAT and BAT. *Genome biology* **20**,
512 217 (2019).
- 513 61. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079
514 (2009).
- 515 62. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the
516 organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).
- 517 63. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq
518 data with DESeq2. *Genome biology* **15**, 550 (2014).
- 519

520 **Figure Captions**

521

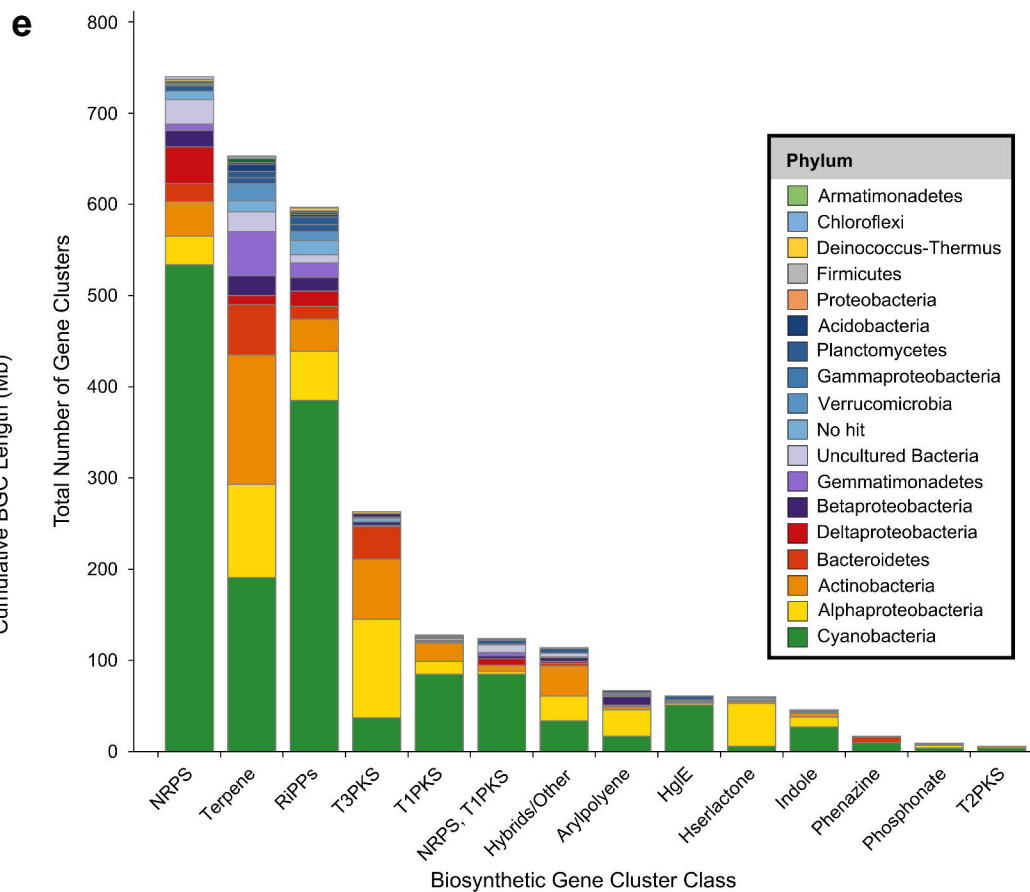
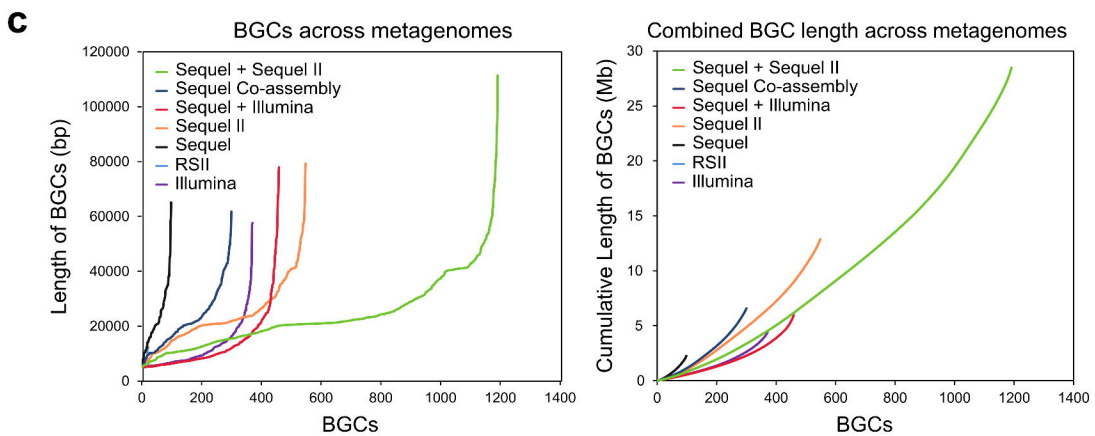
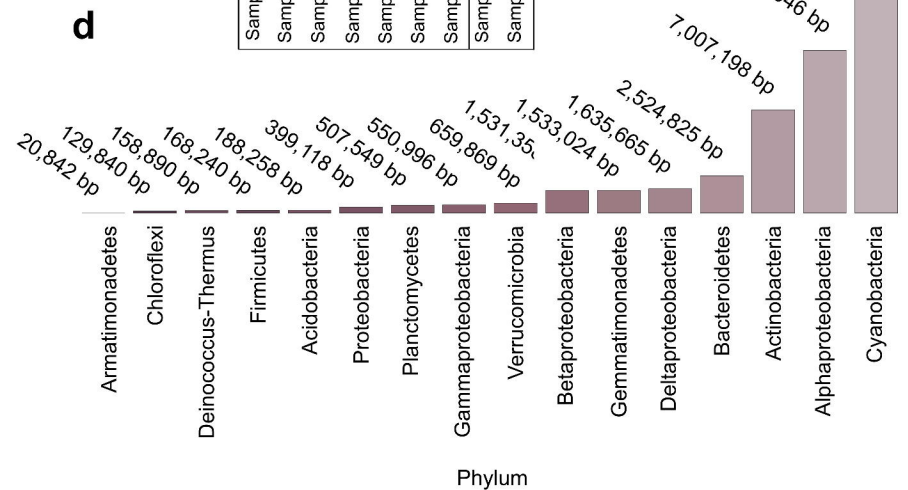
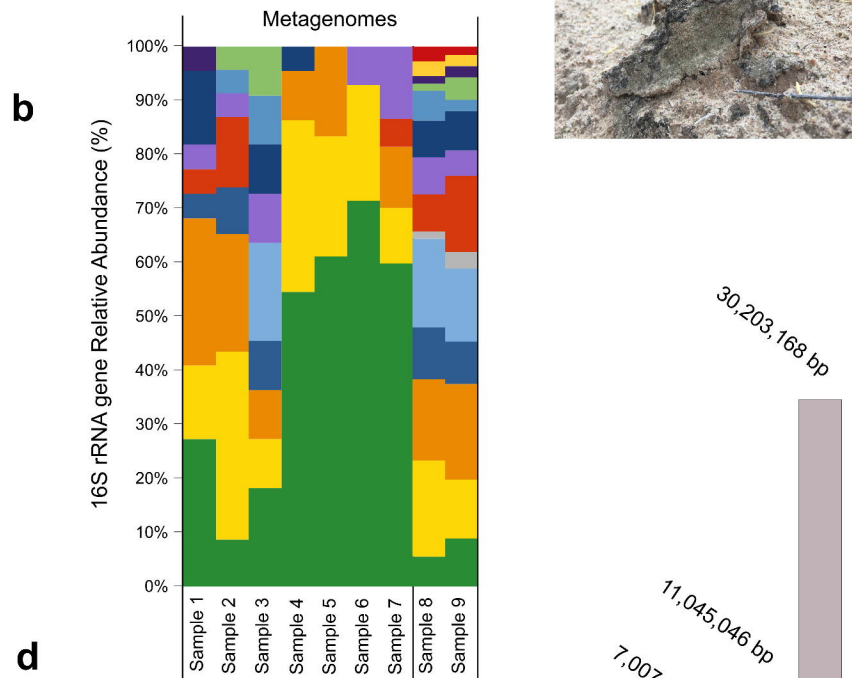
522 **Fig. 1 | Secondary metabolism of biocrust.** **a**, Field sampling location in the Green Butte Site
523 near Canyonlands National Park (Moab, UT) with the biological soil crust inlay showing the
524 characteristic green coloration. **b**, Taxonomic composition of the metagenomes based on exact
525 sequence variants (ESVs) of 16S rRNA genes across sequencing platforms. Relative abundances
526 were calculated after assigning taxonomy against the SILVA reference database. **c**, Left panel
527 shows the number of Biosynthetic Gene Clusters (BGCs) recovered from each assembly, arranged
528 from shortest to longest. Right panel shows the cumulative length of BGCs recovered from each
529 metagenome in Megabases (Mb). **d**, Taxonomic distribution of BGCs in megabase pairs (Mb) at
530 the phylum or class level. **e**, BGCs longer than 5 kb from each major class of secondary
531 metabolism, colored by putative phylum-level assignments.

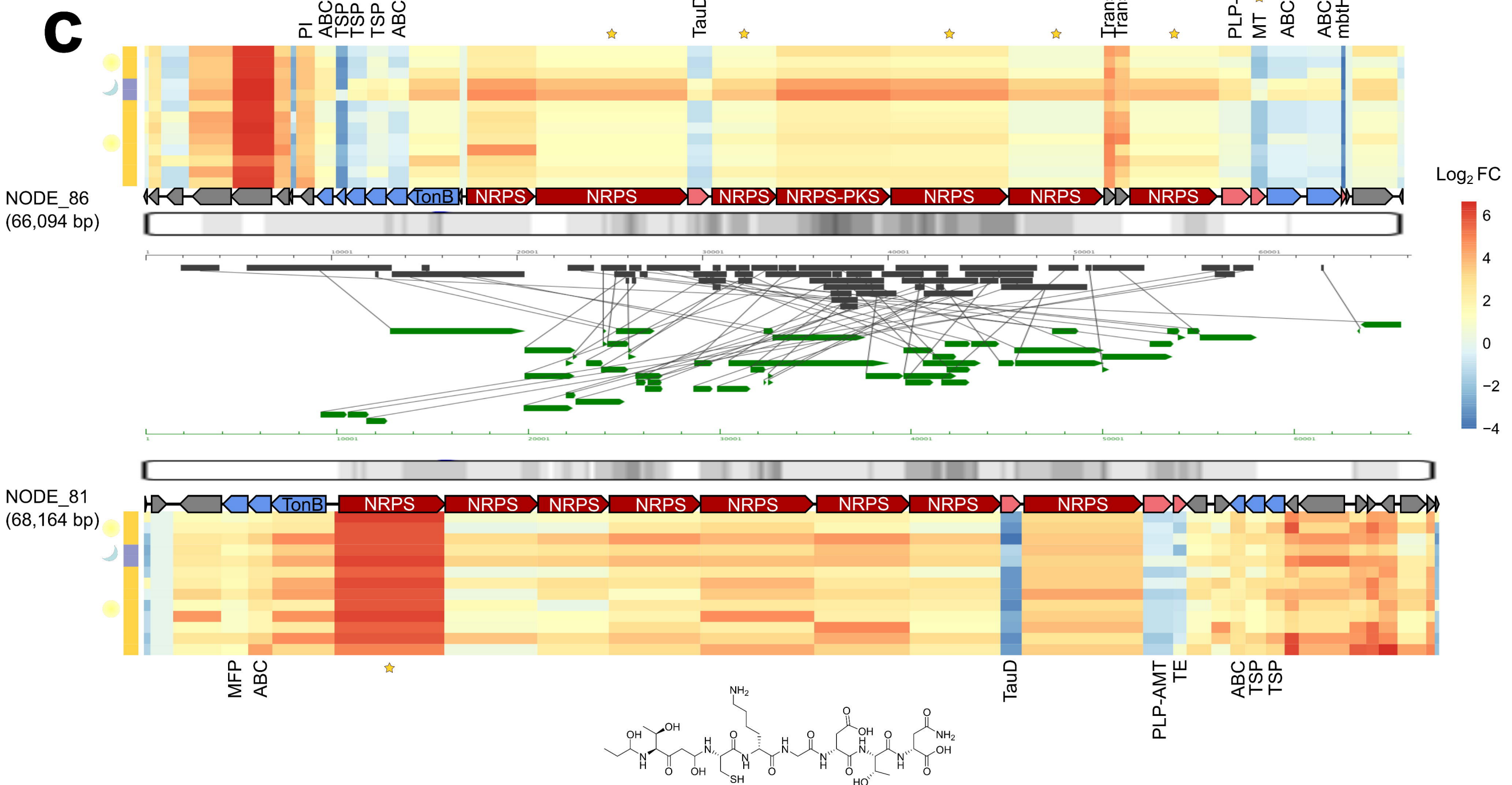
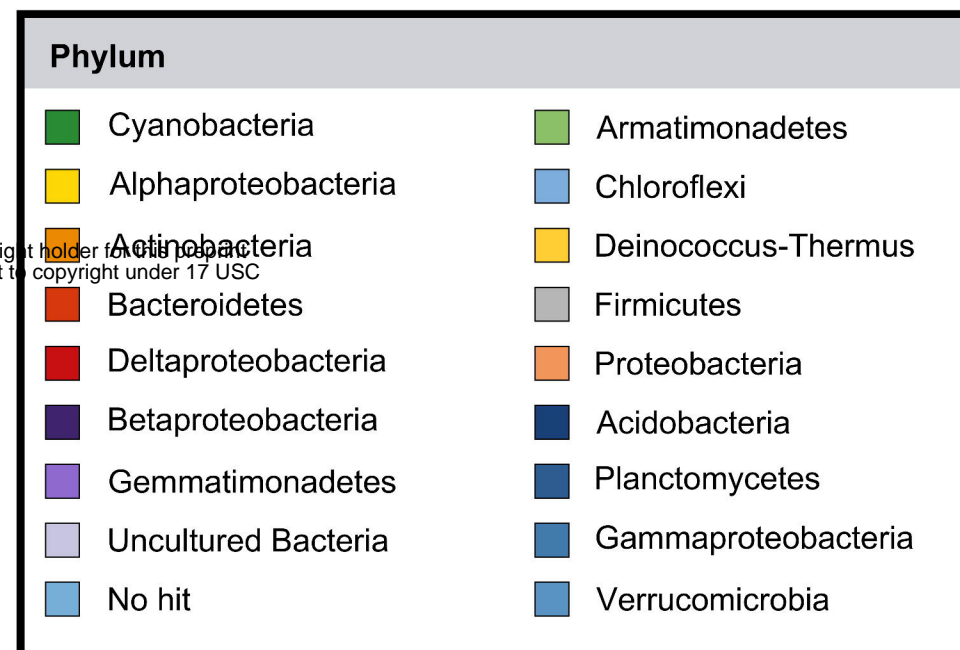
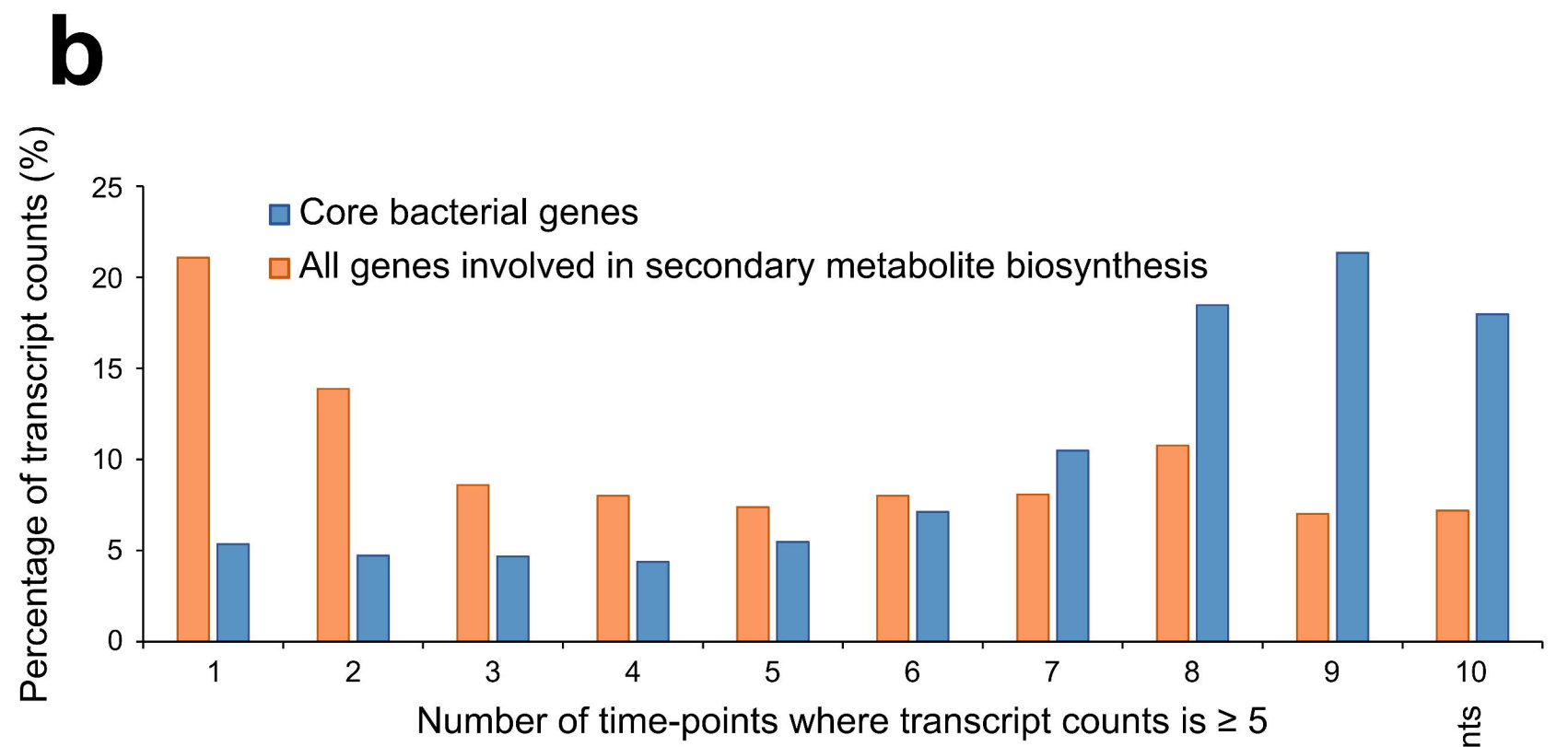
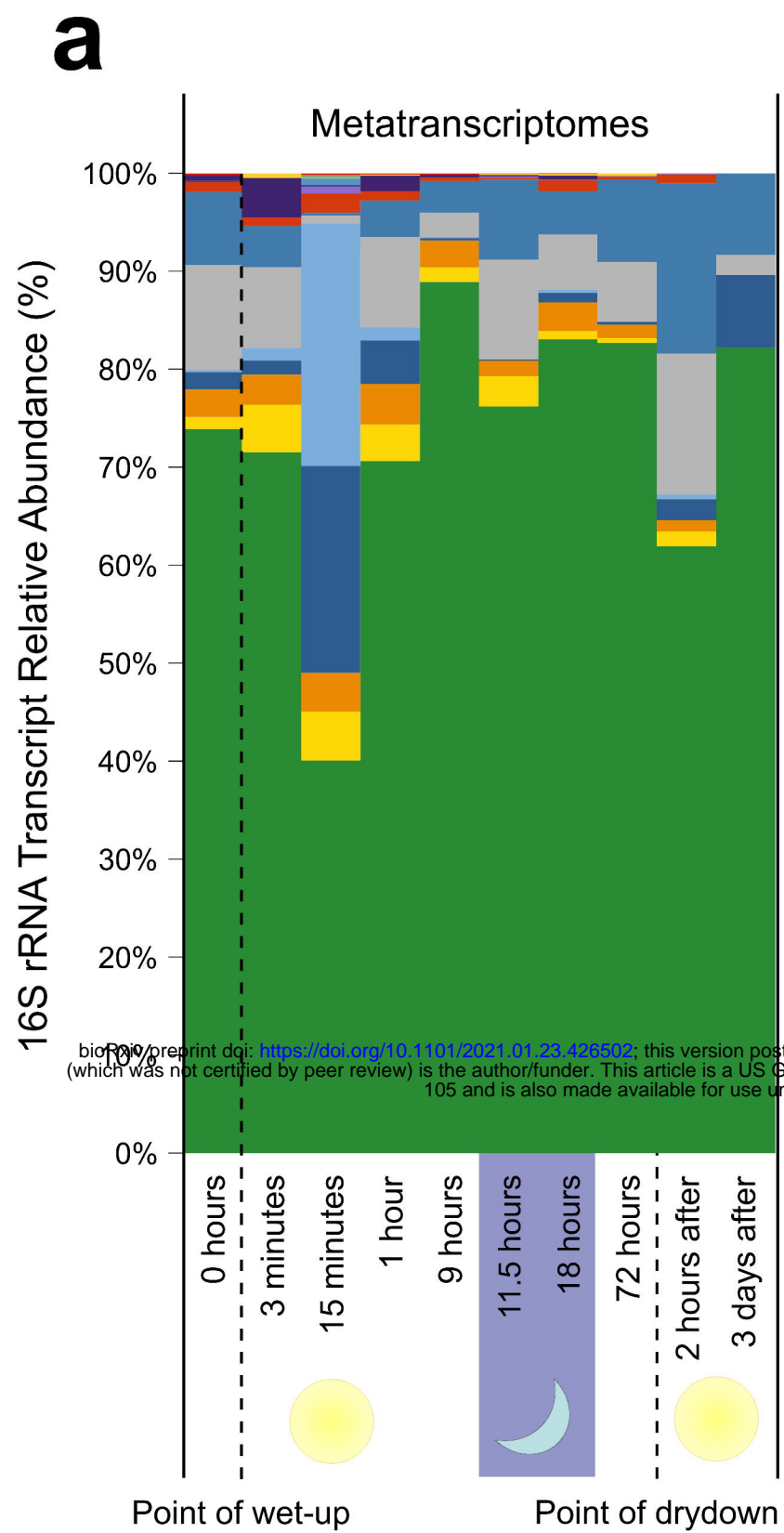
532

533 **Fig. 2. | Transcription of secondary metabolites.** **a**, Taxonomic composition of the
534 metagenomes based on exact sequence variants (ESVs) of 16S rRNA transcripts during a soil
535 wetting experiment. Relative abundances were calculated after assigning taxonomy against the
536 SILVA reference database. **b**, Core bacterial gene transcription ($n=46$ genes including DNA-
537 binding proteins, Large and Small subunit ribosomal proteins) shown in blue compared to
538 secondary metabolite gene transcription (orange). Genes transcribed at all 10 timepoints
539 (rightmost point) are thought to experience constitutive expression. The y -axis indicates the
540 proportion of genes present in each category. **c**, Putative rearranged siderophore-producing gene
541 clusters found in the co-assembled metagenomes that show homology. Transcriptional profiles of
542 gene clusters with differentially expressed genes. Heatmap columns are scaled to the size of the
543 mapped gene, and row order indicates progression across the experiment from 0 hours (bottom
544 row) to 3 days after wetting (top row). Predicted chemical structures are shown on the right.

545

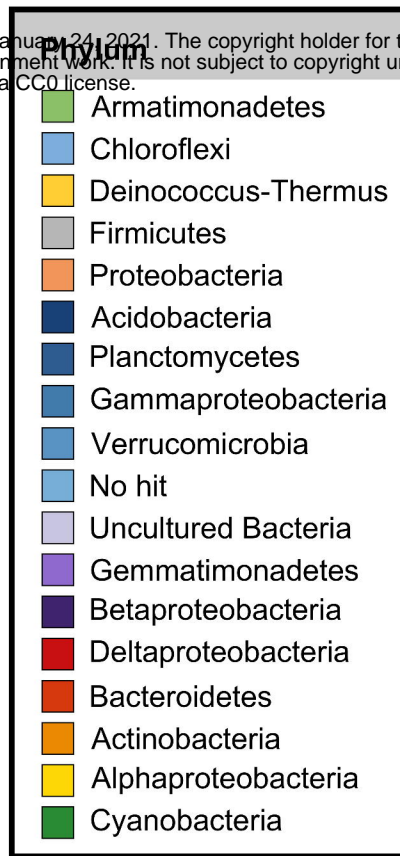
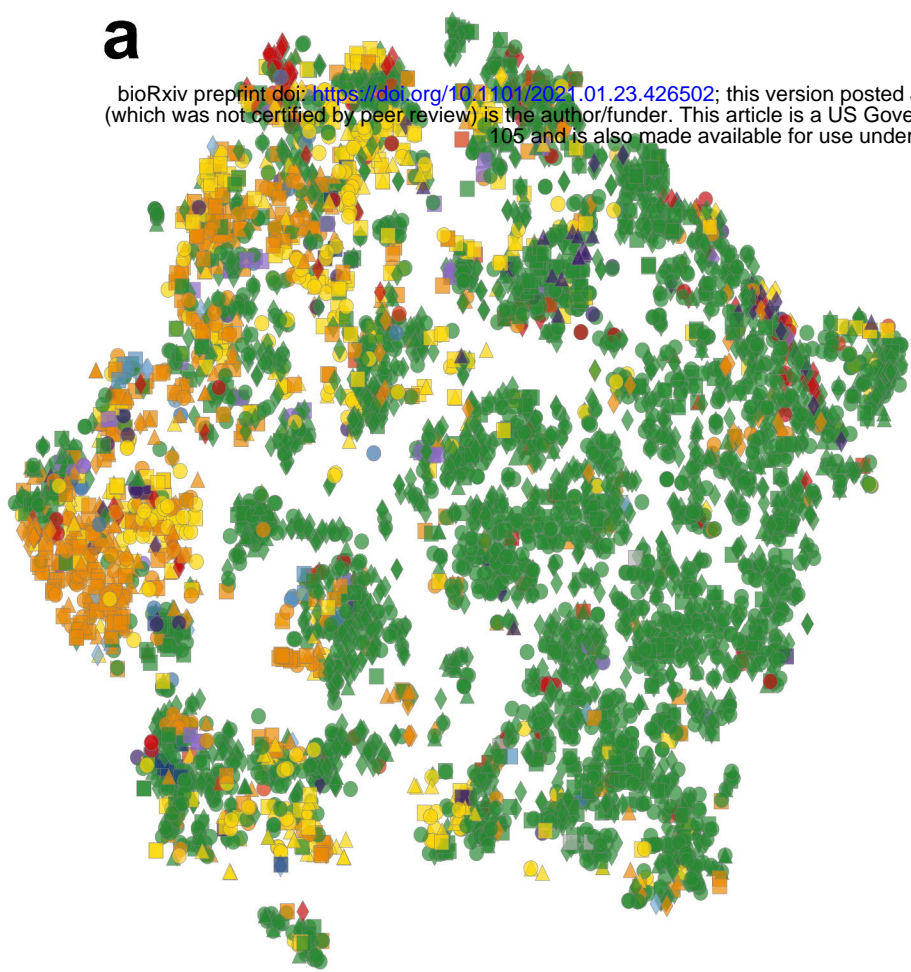
546 **Fig. 3. | Phylum specific transcription of secondary metabolites.** **a**, t-SNE visualization of
547 every individual biosynthetic gene identified. The color of the points indicate the phylum
548 assignment whilst shapes indicate the BGC class. **b**, Co-occurrence network based on Pearson
549 correlations ($r > 0.8$) among entire BGCs ($n=2,988$) based on average z-scores at each time point.
550 Each node is a BGC within a contig that are colored by phylum and shaped by BGC type. Closely
551 clustered nodes share similar transcriptional profiles. **c**, Line plot showing 16S rRNA transcript
552 copy number over time shown by black, dotted lines. Average BGC transcription over time shown
553 by the colored, solid lines. *Cyanobacteria* (green) show a unique night-time upregulation of
554 secondary metabolism. Purple background indicates night-time transcription.



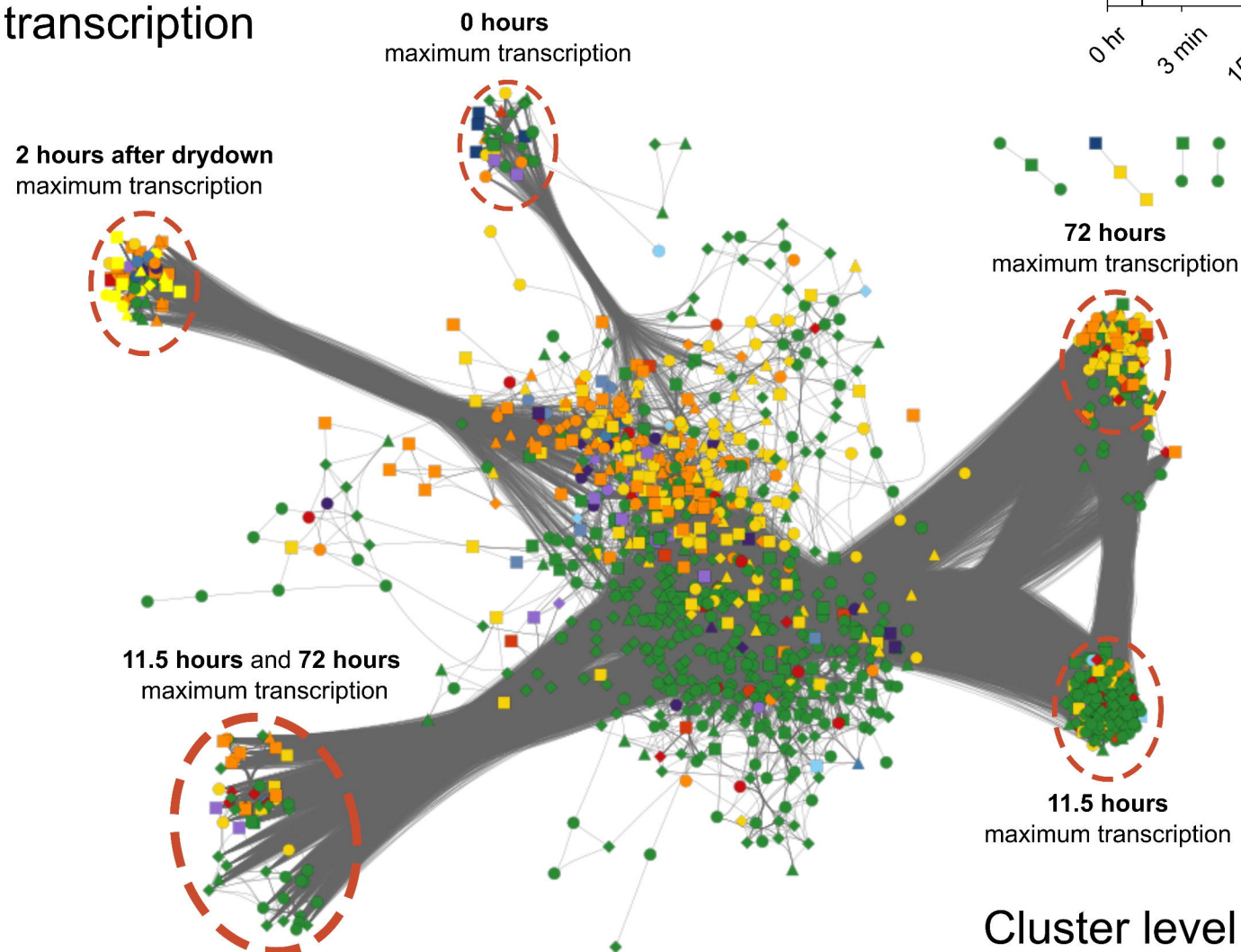
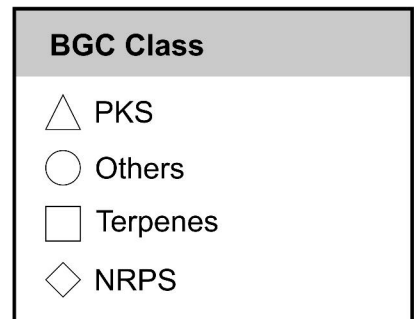
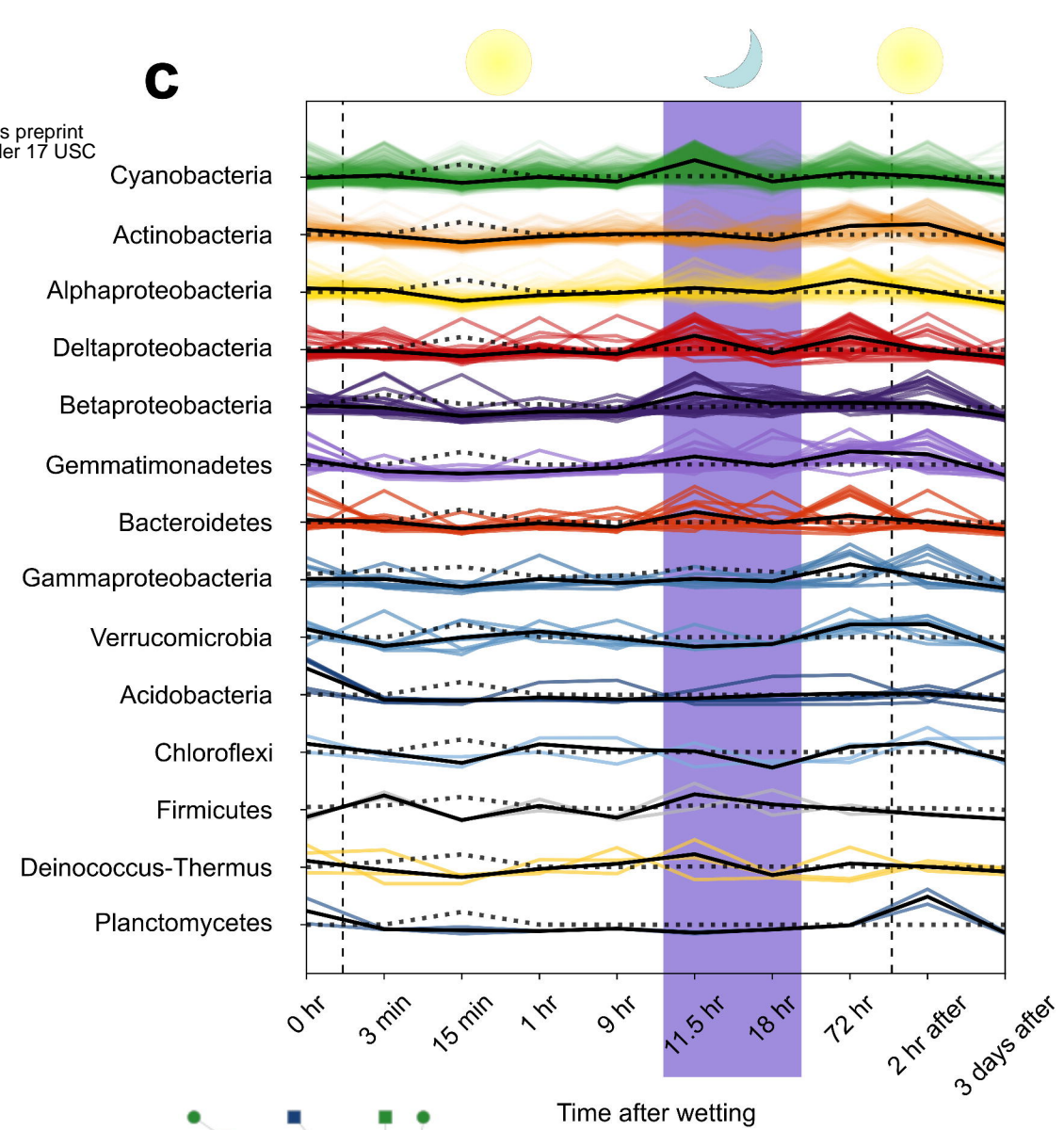


a

bioRxiv preprint doi: <https://doi.org/10.1101/2021.01.23.426502>; this version posted January 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under aCC0 license.



Gene level transcription

b**c**

Cluster level transcription