

1 **Genome mining as a biotechnological tool for the discovery of**
2 **novel biosynthetic genes in lichens**

3

4 **Garima Singh^{1,2,*}, Francesco Dal Grande^{1,2,3}, Imke Schmitt^{1,2,4}**

5 1 Senckenberg Biodiversity and Climate Research Centre (SBIK-F), 60325 Frankfurt am
6 Main, Germany

7 2 LOEWE Center for Translational Biodiversity Genomics (TBG), 60325 Frankfurt am Main,
8 Germany

9 3 Department of Biology, University of Padova, Via U. Bassi, 58/B, 35121 Padova, Italy

10 4 Institute of Ecology, Diversity and Evolution, Goethe University, Frankfurt am Main,
11 Germany

12

13 **Emails:** Garima Singh: garima.singh@senckenberg.de, gsingh458@gmail.com

14 Francesco Dal Grande: francesco.dalgrande@unipd.it

15 Imke Schmitt: imke.schmitt@senckenberg.de

16

17 ***Corresponding author:** Garima Singh

18

19 **Abstract**

20 The ever-increasing demand for novel drugs highlights the need for bioprospecting
21 unexplored taxa for their biosynthetic potential. Lichen-forming fungi (LFF) are a rich source
22 of natural products but their implementation in pharmaceutical industry is limited, mostly
23 because the genes corresponding to a majority of their natural products is unknown.
24 Furthermore, it is not known to what extent these genes encode structurally novel molecules.
25 Advance in next-generation sequencing technologies has expanded the range of organisms
26 that could be exploited for their biosynthetic potential. In this study, we mine the genomes of
27 nine lichen-forming fungal species of the genus *Umbilicaria* for biosynthetic genes, and
28 categorize the BGCs as “associated product structurally known”, and “associated product
29 putatively novel”. We found that about 25-30% of the biosynthetic genes are divergent when
30 compared to the global database of BGCs comprising of 1,200,000 characterized biosynthetic
31 genes from planta, bacteria and fungi. Out of 217 total BGCs, 43 were only distantly related
32 to known BGCs, suggesting they encode structurally and functionally unknown natural
33 products. Clusters encoding the putatively novel metabolic diversity comprise PKSs (30),
34 NRPSs (12) and terpenes (1). Our study emphasizes the utility of genomic data in
35 bioprospecting microorganisms for their biosynthetic potential and in advancing the industrial
36 application of unexplored taxa. We highlight the untapped structural metabolic diversity
37 encoded in the lichenized fungal genomes. To the best of our knowledge, this is the first
38 investigation identifying genes coding for NPs with potentially novel therapeutic properties in
39 LFF.

40

41 **Key words**

42 Natural products, fungi, biosynthetic genes, lichen-forming fungi, secondary metabolites,
43 drug discovery, medicinal fungi, BiG-FAM, BiG-SLiCE

44

45 **Background**

46 Natural products (NPs) are small molecules in nature produced by the organism. Historically,
47 NPs have played a key role in drug discovery due to their broad pharmacological effects
48 encompassing antimicrobial, antitumor, anti-inflammatory properties and against
49 cardiovascular diseases [1,2]. In the past decades about 70% of the drugs were based on NPs
50 or NP analogs [1,2]. The demand for novel drugs however, is ever increasing due to the
51 emergence of antibiotic-resistant pathogens, the rise of new diseases, the existence of diseases
52 for which no efficient treatments are available yet, and the need for replacement of drugs due
53 to toxicity or high side-effects [3,4]. One way to address global health threats and to
54 accelerate NP-based drug discovery efforts is bioprospecting unexplored taxa to assess their
55 biosynthetic potential and identify potentially novel drug leads.

56 Genes involved in the synthesis of a NPs are often grouped together in biosynthetic
57 gene clusters [5–7]. These clusters have a core gene which codes for the backbone structure of
58 the NP and other genes which may be involved in the modification of the backbone or may
59 have a regulatory or transport-related function [5,8–10]. Depending upon the core gene, the
60 BGCs could be grouped into the following major classes: non-ribosomal peptide synthetases
61 (NRPS), polyketide synthases (PKS), NRPS-PKS (hybrid non-ribosomal peptide synthetase-
62 polyketide synthase), terpenes, and RiPP (ribosomally synthesized and post-translationally
63 modified peptide). Conserved motives, especially of the PKS genes, facilitate the
64 bioinformatic detection of the clusters [11–14].

65 Traditionally, a large portion of NP-based drugs have been contributed by a few
66 organisms as the drug discovery was mostly restricted to culturable organisms [15–17]. In the
67 last decades, bioinformatic prediction of biosynthetic gene or biosynthetic gene clusters
68 (group of two or more genes that are clustered together and are involved in the production of
69 a secondary metabolite) has revolutionized NP-based drug discovery as this process is
70 culture-independent and enables rapid identification of entire biosynthetic landscape from so
71 far unexplored NP resources, including silent or unexpressed genes. Two tools have been vital
72 to bioinformatic approach to drug discovery: AntiSMASH [18] and MIBiG [19]. AntiSMASH
73 includes one of the largest BGC database for BGC prediction [18] whereas MIBiG (Minimum
74 Information about a Biosynthetic Gene Cluster) is a data repository allowing functional
75 interpretation of target BGCs by comparison with BGCs with known functions [19]. Recently,
76 efforts have been made to cluster homologous BGCs into gene cluster families (GCFs) and to
77 simultaneously identify novel BGCs [20,21]. Two tools have been introduced to cluster BGCs
78 into GCFs: BiG-FAM clusters structurally and functionally related BGCs into GCFs and
79 identifies structurally most diverse BGCs by comparing the query BGCs to about 1,200,000
80 BGCs of the BiG-FAM database [21]. BiG-SLiCE clusters homologous BGCs of a dataset
81 into GCFs without reference to an external database, to identify unique BGCs in it [20].
82 Bioinformatic prediction and clustering of BGCs allows rapid identification of potentially
83 novel drug leads, reducing the costs and time associated with drug discovery by early
84 elimination of unlikely candidates.

85 Lichens, symbiotic organisms composed of fungal and photosynthetic partners (green
86 algae or cyanobacteria, or both), are suggested to be treasure chests of biosynthetic genes and
87 NPs [22–24]. Although the number of identified NPs per LFF is typically less than 5 [25], the
88 number of BGCs in the genomes of LFF may range from 25-60 [12]. It is not well known
89 how BGCs from LFF relate in structure and function to BGCs from bacteria and non-

90 lichenized fungi, i.e., if a portion of the BGC landscape of LFF is distinct, and might serve as
91 a source of NPs with novel therapeutic properties. Difficulties associated with heterologous
92 expression of LFF genes have so far restricted the application of LFF-derived NPs in the
93 industry. Recently two biosynthetic genes from LFF have been successfully heterologously
94 expressed [9,26]. This, combined with advances in long-read sequencing technology (higher
95 genome quality), and low cost of sequencing provide a promising way forward to discover
96 LFF-derived NPs with pharmacological potential.

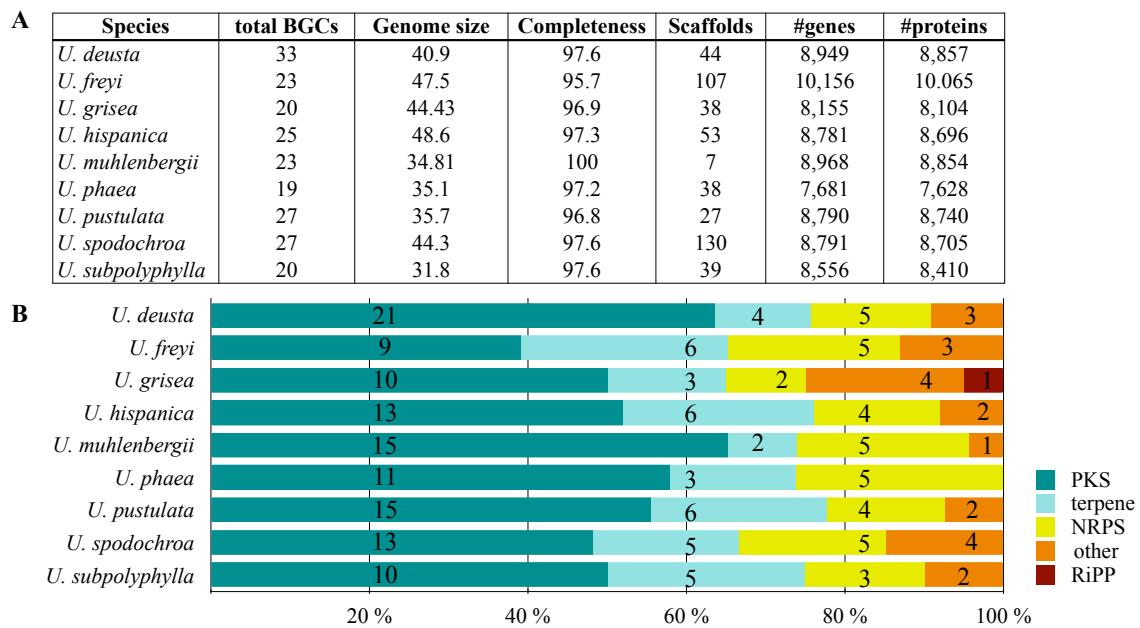
97 Here we employ a long-read sequencing based comparative genomics and genome
98 mining approach to estimate the BGC functional diversity of nine species of the lichenized
99 fungal genus *Umbilicaria*. Specifically, we aim to answer the following questions: (1) What is
100 the functional diversity of BGCs in *Umbilicaria*? and 2) what is the percentage of novel
101 BGCs and species-specific BGCs in *Umbilicaria*?

102

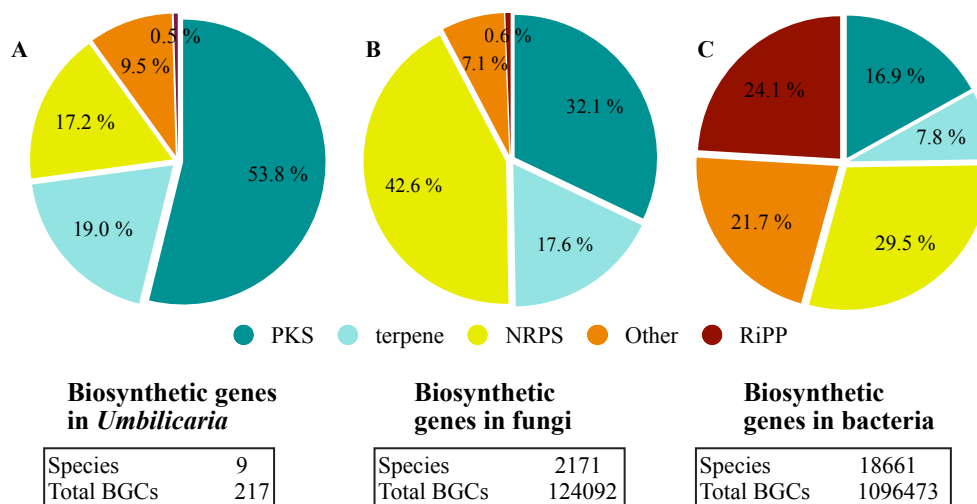
103 **Results**

104 **Overview of BGCs in the *Umbilicaria* genomes**

105 *Umbilicaria* genomes contain 20-33 BGCs, with the highest number of BGCs detected in *U.*
106 *deusta* and lowest in *U. phaea* (Fig. 1A). We did not observe a correlation between genome
107 size and number of BGCs (correlation coefficient = 0.10). *Umbilicaria* species contain an
108 average of 13 PKS clusters, and 4.2 NRPS clusters per species (Fig. 1B), making a PKS to
109 NRPS clusters proportion of 3.1). The most dominant class of BGC in *Umbilicaria* are the
110 ones with PKSs, amounting more than 50% of the total BGCs, followed by terpene clusters
111 (about 20%) and NRPS clusters (about 15%) respectively, (Fig. 2A). In contrast, NRPSs are
112 the most dominant class among fungal and bacterial BGCs (Fig. 2B, C), amounting to about
113 42% and 30% respectively.



114
 115 **Fig 1.** Genome quality metrics and diversity of biosynthetic genes in nine species of *Umbilicaria*. **A)**
 116 Genome metrics including the total number of biosynthetic gene clusters as predicted by antiSMASH,
 117 and number of genes and proteins estimated by InterProScan and SignalP as implemented in the
 118 funannotate pipeline. **B)** Diversity of biosynthetic gene clusters associated with major natural product
 119 categories, indicated as percentages (colored bars) and absolute numbers (numbers on bars).
 120



121
 122 **Fig 2.** Biosynthetic gene clusters in **A)** *Umbilicaria*, **B)** the full fungal BGC dataset and **C)** full
 123 bacterial BGC dataset. PKSs are the most dominant class of BGCs in *Umbilicaria* whereas in fungi
 124 and bacteria NRPSs are the predominant BGC class. Although the publicly available LFF genomes (>
 125 50) are much lower than the non-lichenized fungi (about 2100), all the LFF genomes analyzed for
 126 their BGCs have PKSs as the most common class of BGCs (see discussion for details), suggesting that
 127 the predominance of PKSs as observed here in *Umbilicaria* dataset is a common feature of LFF
 128 genomes.

129 **BGC clustering: BiG-FAM**

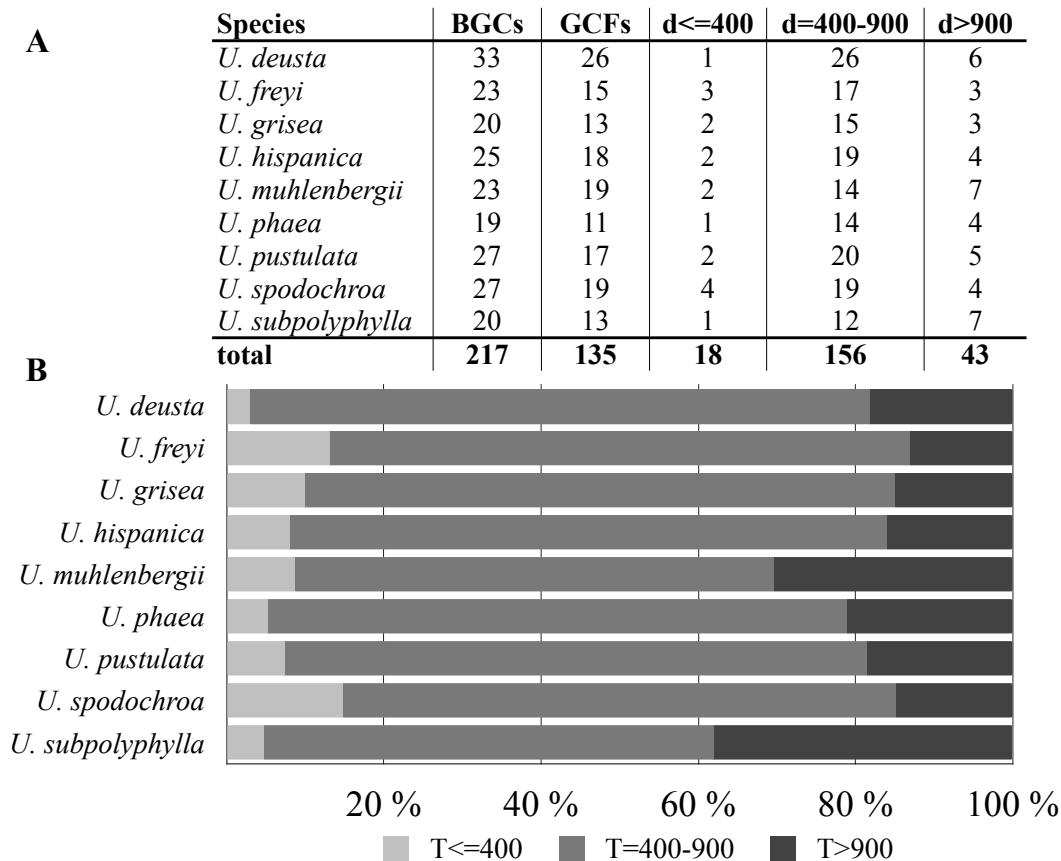
130 Of the total 217 BGCs found in 9 *Umbilicaria* species, 18 BGCs (8%) obtained a BGC-to-
131 GCFs (Gene Cluster Families) pairing distance lower than 400, indicating that they
132 potentially code for structurally very similar compounds known from the BGCs of their
133 respective GCFs (Fig. 3A, B); 156 (71%) had a pairing distance of 400-900, suggesting that
134 they share similar domain architectures with previously described BGCs in the BiG-FAM
135 database. We identify the clusters belonging to above two groups as “associated product
136 structurally known”. 43 BGCs (21%) had a pairing distance greater than 900 and are
137 potentially BGCs encoding novel natural products (Fig. 3 A). We identify these clusters as
138 “associated product putatively novel”. These BGCs belong to the class terpenes (1 BGCs),
139 NRPSs (12 BGCs) and PKSs (30 BGCs). The details of these BGCs and the sequence of the
140 core gene is provided in the Additional file 1.

141

142 **Within-genus comparison of BGCs: BiG-SLiCE**

143 We identified species-specific BGCs within *Umbilicaria* using BiG-SLiCE. Out of 217 total
144 BGCs, 159 (72%) grouped into 20 GCFs ($d=900$), suggesting they are similar clusters shared
145 by multiple species, while 58 (28%) had a $d > 900$, indicating that they were only distantly
146 related to other BGCs in *Umbilicaria*. Each *Umbilicaria* species contains four to ten (6.45 –
147 16.13%) unique, species-specific BGCs (Additional file 2A). In *U. deusta* we detected two
148 BGCs (both with PKSs) that were extremely divergent ($d > 1800$) within the genus
149 (Additional file 2B).

150 Out of these BGCs, 15 are unique within *Umbilicaria* as well divergent from the
151 BGCs to the known BGCs present in BiG-FAM database.



152

153 **Fig. 3 A)** Total BGCs in *Umbilicaria* and GCFs as identified by BiG-FAM and the number of BGCs
 154 clustering into a pre-characterized gene cluster families (GCFs) in BiG-FAM and their distance
 155 groups. $d \leq 400$ suggest that the cluster codes for a structurally and functionally similar NP, $d = 400-$
 156 900 indicates that the BGC codes for a related but structurally and functionally divergent NP, whereas
 157 $d > 900$ suggests that the BGC codes for a novel NP. **B)** Bar plots representing the percentage of BGCs
 158 in each *Umbilicaria* species with $d \leq 400$, $d = 400-900$ and $d > 900$. Only a small proportion of BGCs
 159 in each species could be grouped into a pre-characterized GCF in the BiG-FAM database (21,678
 160 species, 1,225,071 BGCs and 29,955 GCFs), whereas a large proportion of them is only distantly
 161 related to the pre-characterized BGCs. About 15-30% of BGCs could not be grouped into BiG-FAM
 162 gene cluster families and potentially code of structurally and functionally divergent NPs.
 163

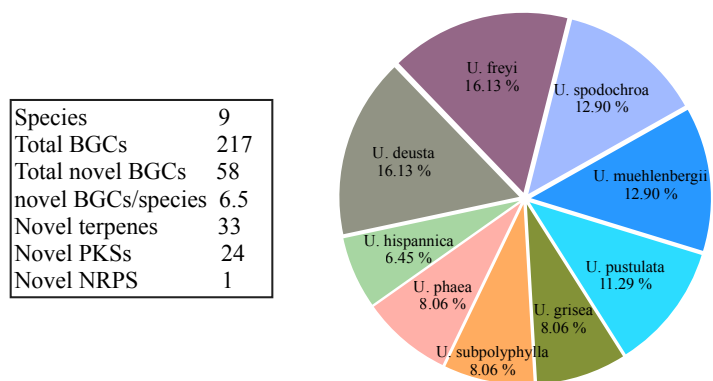


Fig. 4 Pie chart depicting the contribution of each species to the overall novel *Umbilicaria* BGCs (as identified by BiG-SLiCE, T>900) Each *Umbilicaria* species contains about 4-10 unique, species-specific BGCs. *U. freyi* and *U. deusta* contain the highest number of novel BGCs. The number of novel BGCs slightly positively correlated to the number of clusters (R=0.68). Out of 58 BGCs unique BGCs (T>900) 56.89% were terpene- and 41.37% were PKS clusters.

164

165

166 Discussion

167 Lichens produce a large number of natural products, and they have even more BGCs [27–29].

168 However, whether these BGCs encode hitherto unknown metabolic diversity/chemical

169 structures is not known. Here we quantify, for the first time, the proportion of BGCs linked to

170 putatively novel natural products in a group of closely related lichen-forming fungi. The

171 identification of 23 clusters encoding putatively novel chemical structures can be useful in the

172 search for new structures and drug leads.

173 In this study we mined the genomes of the *Umbilicaria* spp. to identify all the BGCs

174 (Fig. 1), followed by clustering the structurally and functionally similar BGCs into gene

175 cluster families (Fig. 3A, B) and identifying the gene clusters potentially coding for novel

176 NPs (Fig. 4, Additional File 1). Using *Umbilicaria* spp. as a study system, we show that LFF

177 biosynthetic landscape is diverse from that of non-lichenized fungi and bacteria, being

178 particularly rich in PKSs (Fig 2) and that a substantial portion for LFF BGCs (about 28% in

179 case of *Umbilicaria*) potentially codes for novel NPs (Fig. 3A, B). To the best of our

180 knowledge, this is the first investigation of this kind, implementing state of the art

181 computational tools to determines the proportion of metabolic diversity in LFF coding for

182 novel drugs and identifying candidate genes as a source of drug leads to prioritize them for
183 drug discovery efforts.

184

185 **Biosynthetic potential and BGC diversity of *Umbilicaria* spp.**

186 Although only PKSs-derived NPs are reported from *Umbilicaria* species (gyrophoric-,
187 umbilicatic-, and hiassic acid etc.) [30–32], we found that the *Umbilicaria* BGC landscape is
188 biosynthetically diverse and comprises three to five classes of NPs (Fig 1A, B). This is also
189 the case for most other LFF, for instance, PKS-derived NPs, are reported from *Bacidia* spp.,
190 *Cladonia* spp., *Endocarpon* spp., *Evernia prunastri*, *Umbilicaria pustulata*, *Pseudevernia*
191 *furfuracea*, but all of them contain several PKS, NRPS and terpene gene clusters [12,29,32–
192 34]. All these above-stated studies show that the biosynthetic potential of LFF vastly exceeds
193 their detectable chemical diversity. On average LFF may contain up to 30-40 BGCs but the
194 number of identified compounds per species is usually less than 10 [12,33,35]. This could be
195 because most of the clusters are silent and do not synthesize the NP or it could be simply
196 because of the failure to detect the NP. Bioinformatic characterization of entire BGC
197 landscape followed by identification of most distinct BGCs provides a way to estimate the
198 novelty of all the BGCs including the unexpressed and silent ones.

199

200 **BGC diversity of LFF as compared to bacteria and non-lichenized fungi**

201 We identified five classes of BGCs in the *Umbilicaria* genomes. PKSs were the most
202 dominant class, amounting to about 50%, followed by terpenes (19%), and NRPSs (14%)
203 (Fig. 1, Fig. 2 A). BGCs including PKSs typically make up the majority of BGCs in LFF:
204 *Evernia prunastri* (60%), *Pseudevernia furfuracea* (61%), *Cladonia* spp. (65%), *Endocarpon*
205 *pusillum* (58%), *Lobaria pulmonaria* (46%), and *Ramalina peruviana* (63%) (cite).

206 Although the number of publicly accessible, good quality LFF genomes are rather
207 scarce for LFF (<25) as compared to the bacteria and non-lichenized fungi, the data available
208 (9 *Umbilicaria* spp. genomes [36] plus 9 other publicly available lichen genomes) suggests
209 that the predominance of PKSs is a common feature of BGCs in LFF contributing more than
210 50% to the total BGCs. In contrast, in bacteria and non-lichenized fungi, NRPS are the most
211 prevalent BGC class, amounting to about 30% and 42% respectively, followed by the PKSs
212 (Fig. 2 B, C). This suggests that the biosynthetic potential of LFF is unique as compared to
213 the other organisms traditionally exploited for NPs, i.e., non-lichenized fungi and bacteria,
214 especially with respect to PKS diversity. In this regard, a recent study suggested that although
215 bacteria and fungi may share a few NPs, they do not have an overlapping chemical space and
216 instead have distinct biosynthetic potential [37]. LFF having a distinct BGC landscape
217 presents a complementary resource of NPs with promising medicinally-relevant biosynthetic
218 properties.

219

220 ***Umbilicaria* BGCs: Gene Cluster Families (GCFs) and novel NPs**

221 Gene cluster families (GCFs) are the groups of BGCs that encode the same or very similar
222 molecules. A total of 217 BGCs from nine *Umbilicaria* species were clustered into of 135
223 unique GCFs. (Fig 3 A) This suggests that *Umbilicaria* spp. are potentially capable of
224 synthesizing many structurally and functionally different natural products, although in nature
225 only one compound class is typically detected (depsides, linked to a BGC containing a PKS).

226 Only a small fraction of *Umbilicaria* BGCs, 8%, could be clustered with the pre-
227 characterized BGCs (Fig. 3A, B). About 71% of the BGCs were clustered to the BiG-FAM
228 GCFs with d= 400-900, indicating that they were only distantly related in structure and
229 function (Fig. 3 A, B). These BGCs are also interesting candidates to be investigated for their
230 biosynthetic properties as even a minor difference in the cluster and the chemistry of the final

231 metabolites could cause a crucial difference in bioactivity related to function and the
232 pharmacological potential of the product [38].

233 About 21% percent BGCs were highly divergent ($d > 900$) and are novel, potentially
234 coding for structurally and functionally unique NPs and could be an interesting target for NP-
235 based drug discovery (Fig. 3 B). The strikingly high number of novel BGCs in a fungal genus
236 adds to the mounting evidence that the non-model and understudied taxa are enormous,
237 untapped source of novel NPs.

238 Genome mining for large genomic regions, such as fungal BGCs, works best when the
239 genomes under study are highly complete and contiguous, as well as reliably annotated. Many
240 publicly available LFF genomes do not fulfill these criteria, preventing a taxonomically broad
241 study of biosynthetic novelty encoded in the genomes of LFF. We were surprised that even a
242 “chemically boring” lichen taxon, such as the genus *Umbilicaria*, harbored 43 BGCs
243 encoding putatively unknown natural product diversity. It lets us suspect that chemically more
244 diverse taxa, e.g. Lecanorales or Pertusariales, each including hundreds of species, are even
245 richer sources of BGCs with novel functions, and compounds with potential pharmaceutical
246 applications. Increased genome sequencing of taxonomically diverse LFF, combined with
247 higher genome qualities will facilitate BGC discovery.

248

249 **Unique BGCs within *Umbilicaria* spp.: BiG-SLiCE**

250 BGCs which are uniquely occurring in a species are candidates for interesting NPs [20,37,39].
251 On average each *Umbilicaria* species contains seven species-specific BGCs. Most of the
252 novel BGCs are present in *U. deusta* and *U. freyi* whereas *U. hispanica* has lowest number of
253 novel BGCs (Fig. 4). This suggests that even closely related species (species within a single
254 genus) contain diverse biosynthetic potential. Species or strain specific biosynthetic potential
255 has already been demonstrated for LFF, for example in *Umbilicaria pustulata* [32] and

256 *Pseudevernia furfuracea* [33] and it is a rather common occurrence in fungi [32,37,40]. For
257 instance, majority of the BGCs in *Streptomyces*, i.e., 57% have been shown to be strain-
258 specific [41]. The unique BGCs within *Umbilicaria* belong to the BGC classes PKSs,
259 terpenes, NRPS as well as to indoles (Supplementary information S2). Of these, mostly only
260 PKS derived NPs have been well studied in LFF and shown to have diverse pharmacological
261 properties [42–44].

262 Two PKS obtained a pairing distance greater than 1800. These were the most
263 divergent BGCs (Supplementary information S2) within *Umbilicaria* and were “orphan
264 BGCs”, i.e., for these clusters the corresponding metabolite cannot be predicted. Recently
265 several orphan clusters have been activated to synthesize a compound with useful
266 pharmacological properties, for example the antibiotic holomycin gene cluster from the
267 marine bacterium *Photobacterium galathea* was activated in culture [45–48]. The novel and
268 orphan clusters reported in this study are potentially interesting candidates for synthesizing
269 molecules with unique pharmacological properties and may serve as drug leads.

270 About 17% of fungal BGCs, 8% of bacterial BGCs and 19% of LFF BGCs comprise
271 terpenes (Fig. 2). Terpenes are pharmaceutically extremely versatile, with antimicrobial, anti-
272 inflammatory, neurodegenerative, and cytotoxic properties [49–54]. Some of the common
273 plant-derived terpenes and terpenoids are curcumin, Eucalyptus oil. Although several studies
274 reported pharmacological properties of fungal terpenes, such studies on LFF are missing
275 despite the slightly larger proportion of terpenes in LFF genomes. In this study we report
276 structurally and functionally unique terpenes as promising candidates, to be investigated for
277 their pharmaceutical potential.

278

279 **Conclusion**

280 In this study we identified the biosynthetic diversity of the lichen forming fungal genus
281 *Umbilicaria*, grouped the structurally and functionally related clusters into GCFs and
282 identified the most diverse, potentially novel clusters. Using *Umbilicaria* as model system we
283 show that LFF constitute a valuable source of novel NPs suggesting that there is tremendous
284 natural product diversity to be discovered in them. In particular they are rich source of novel
285 PKSs and terpenes. Combining this observation with other sequenced LFF we show that LFF
286 are indeed a source of untapped natural product diversity.

287

288 **Materials and methods**

289 **Dataset**

290 The genomes of the following *Umbilicaria* species were used for this study: *U. deusta*, *U.*
291 *freyi*, *U. grisea*, *U. subpolyphylla*, *U. hispanica*, *U. phaea*, *U. pustulata*, *U. muhlenbergii* and
292 *U. spodochoa*. Except *U. muhlenbergii* which belongs to the Bioproject PRJNA239196, all
293 the other genomes are a part of Bioproject PRJNA820300 (Table 1). The details of sample
294 and library preparation, as well as genome sequencing for *U. muhlenbergii* are available in
295 Park et al. [55] and for the other eight *Umbilicaria* spp in Singh et al. [36]. Briefly, all the
296 genomes except *U. muhlenbergii* were generated via PacBio SMRT sequencing on the Sequel
297 System II using the continuous long read (CLR) mode or the circular consensus sequencing
298 (CCS) mode. The continuous long reads (i.e. CLR reads) were then processed into highly
299 accurate consensus sequences (i.e. HiFi reads) and assembled into contigs using the assembler
300 metaFlye v2.7 [56]. The contigs were then scaffolded with LRScaf v1.1.12
301 (github.com/shingocat/lrscaf, [57]). We used only binned Ascomyocota reads for this study

302 (extracted using blastx in DIAMOND (--more-sensitive --frameshift 15 --range-culling) on a
303 custom database and following the MEGAN6 Community Edition pipeline [58]).

304

305 **BGC prediction and clustering: AntiSMASH**

306 BGCs were predicted using antiSMASH (antibiotics & SM Analysis Shell, v6.0) with scripts
307 implemented in the funannotate pipeline [18,59]. We tested, if a smaller genome size was
308 correlated with a lower number of BGCs. A correlation coefficient near 0 indicates no
309 correlation whereas a coefficient near 1 indicates a positive correlation.

310

Table 1. Voucher information of the genomes used in the study

Organism	Sample ID	Sequencing technology	BioProject	BioSample	Genome accession
<i>Umbilicaria deusta</i>	TBG_2334	PacBio sequal II	PRJNA820300	SAMN26992774	JALILR000000000
<i>Umbilicaria freyi</i>	TBG_2329	PacBio sequal II	PRJNA820300	SAMN26992773	JALILQ000000000
<i>Umbilicaria grisea</i>	TBG_2336	PacBio sequal II	PRJNA820300	SAMN26992780	JALILX000000000
<i>Umbilicaria hispanica</i>	TBG_2337	PacBio sequal II	PRJNA820300	SAMN26992775	JALILS000000000
<i>Umbilicaria muhlenbergii</i>	KoLRI No. LF000956	Illumina HiSeq	PRJNA239196	SAMN02650300	GCA_000611775.1
<i>Umbilicaria phaea</i>	TBG_1112	PacBio sequal II	PRJNA820300	SAMN26992776	JALILT000000000
<i>Umbilicaria pustulata</i>	TBG_2345	PacBio sequal II	PRJNA820300	SAMN26992777	JALILU000000000
<i>Umbilicaria spodochoa</i>	TBG_2434	PacBio sequal II	PRJNA820300	SAMN26992778	JALILV000000000
<i>Umbilicaria subpolyphylla</i>	TBG_2324	PacBio sequal II	PRJNA820300	SAMN26992779	JALILW000000000

311

312 **BGC clustering into BiG-FAM GCFs**

313 The homologous BGCs present in the *Umbilicaria* genomes were grouped into Gene Cluster
314 Families (GCFs) using BiG-FAM, which clusters structurally and functionally related BGCs
315 into GCFs and identifies the structurally most diverse BGCs by comparing the query BGCs to
316 the 1,225,071 BGCs of the BiG-FAM database. The 1,225,071 BGCs in BiG-FAM are

317 clustered into 29,955 GCFs based on similar domain architectures. A GCF comprises closely
318 related BGCs, potentially encoding the same or very similar compounds. By enabling such
319 clustering BiG-FAM establishes the degree of similarity of BGCs of a query taxon to
320 currently known (functionally pre-characterized) fungal and bacterial BGCs. The antiSMASH
321 job ID of each *Umbilicaria* species was used as input for BiG-FAM analysis.

322

323 **Quantification of BGC diversity and species specific BGCs in *Umbilicaria*: BiG-SLiCE**

324 We used BiG-SLiCE [20] to identify the most unique or species-specific BGCs within
325 *Umbilicaria*. While BiG-FAM identifies the most diverse BGCs compared to pre-
326 characterized BGCs from other taxa deposited in public repositories, BiG-SLiCE 1.1.0. is a
327 networking-based tool which assesses relations of BGCs of the dataset (i.e., *Umbilicaria*
328 BGCs in our study) and estimates their distance within the dataset to identify unique, species-
329 specific BGCs. The resulting distance indicates how closely a given BGC is related to other
330 BGCs. BiG-SLiCE was run on the *Umbilicaria* BGC dataset (i.e., 217 BGCs from nine
331 *Umbilicaria* spp.) using three different thresholds: 400, 900 and 1800.

332

333 **Declarations**

334 **Ethics approval and consent to participate:** Not applicable

335 **Consent for publication:** Not applicable

336 **Availability of data and materials:**

337 The dataset(s) supporting the conclusions of this article are available in the GenBank
338 repository, accession PRJNA820300, under the accession numbers JALILQ000000000 -
339 JALILY000000000. The lichen samples of the corresponding *Umbilicaria* Spp. are available
340 as Biosamples SAMN27294873 - SAMN27294881 and the mycobiont samples as

341 Biosamples SAMN26992773 - SAMN26992781. The antiSMASH files of *Umbilicaria* spp.
342 is available at figshare (doi: 10.6084/m9.figshare.19625997).

343 **Competing interests:** None

344 **Funding:** This research was funded by LOEWE-Centre TBG, funded by the Hessen
345 State Ministry of Higher Education, Research and the Arts (HMWK).

346 **Authors' contributions:**

347 GS analyzed and interpreted the data, generated the figures and tables and wrote the
348 manuscript.

349 FDG analyzed the data and assisted with the bioinformatic parts of the study.

350 IS interpreted the data, co-prepared the figures and co-wrote the manuscript.

351 All authors read and approved the final manuscript.

352 **Acknowledgements**

353 We thank Prof. Marnix Medema and Dr. Satria Kautsar for their support with BiG-SLiCE
354 program.

355

356 **Supporting Information**

357 **S1** Most divergent BGCs in *Umbilicaria* as identified by BiG-FAM, along with the cluster
358 information and sequence.

359 **S2** Most distantly related BGCs within *Umbilicaria* as identified by BiG-SLiCE along with
360 the cluster information.

361

362

363 **References**

- 364 1. Newman DJ, Cragg GM. Natural products as sources of new drugs over the 30 years from
365 1981 to 2010. *J Nat Prod.* 2012;75:311–35.
- 366 2. Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four
367 decades from 01/1981 to 09/2019. *J Nat Prod. American Chemical Society;* 2020;83:770–803.
- 368 3. Chakraborty K, Kizhakkekalam VK, Joy M, Chakraborty RD. A leap forward towards
369 unraveling newer anti-infective agents from an unconventional source: a draft genome
370 sequence illuminating the future promise of marine heterotrophic *Bacillus* sp. against drug-
371 resistant pathogens. *Mar Biotechnol.* 2021;23:790–808.
- 372 4. Demain AL. Importance of microbial natural products and the need to revitalize their
373 discovery. *J Ind Microbiol Biotechnol.* 2014;41:185–201.
- 374 5. Keller NP. Fungal secondary metabolism: regulation, function and drug discovery. *Nat Rev*
375 *Microbiol.* Nature Publishing Group; 2019;17:167–80.
- 376 6. Jensen PR. Natural products and the gene cluster revolution. *Trends Microbiol.* 2016. p.
377 968–77.
- 378 7. Calcott MJ, Ackerley DF, Knight A, Keyzers RA, Owen JG. Secondary metabolism in the
379 lichen symbiosis. *Chem Soc Rev.* 2018;47:1730–60. Available from:
380 <http://xlink.rsc.org/?DOI=C7CS00431A>
- 381 8. Rigali S, Anderssen S, Naômé A, van Wezel GP. Cracking the regulatory code of
382 biosynthetic gene clusters as a strategy for natural product discovery. *Biochem. Pharmacol.*
383 2018. p. 24–34.
- 384 9. Kim W, Liu R, Woo S, Kang K Bin, Park H, Yu YH, et al. Linking a gene cluster to
385 atranorin, a major cortical substance of lichens, through genetic dereplication and
386 heterologous expression. *MBio.* 2021;e0111121. Available from:

- 387 <https://pubmed.ncbi.nlm.nih.gov/34154413/>
- 388 10. Aigle B, Lautru S, Spiteller D, Dickschat JS, Challis GL, Leblond P, et al. Genome
389 mining of *Streptomyces ambofaciens*. J Ind Microbiol Biotechnol. 2014;41:251–63. Available
390 from: <https://pubmed.ncbi.nlm.nih.gov/24258629/>
- 391 11. Kum E, İnce E. Genome-guided investigation of secondary metabolites produced by a
392 potential new strain *Streptomyces* BA2 isolated from an endemic plant rhizosphere in Turkey.
393 Arch Microbiol. 2021;203:2431–8.
- 394 12. Calchera A, Dal Grande F, Bode HB, Schmitt I. Biosynthetic gene content of the
395 “perfume lichens” *Evernia prunastri* and *Pseudevernia furfuracea*. Molecules. 2019;24:203.
396 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/30626017>
- 397 13. Bertrand RL, Abdel-Hameed M, Sorensen JL. Lichen Biosynthetic Gene Clusters Part II:
398 Homology Mapping Suggests a Functional Diversity. J Nat Prod. 2018;81:732–48. Available
399 from: <https://pubs.acs.org/doi/10.1021/acs.jnatprod.7b00770>
- 400 14. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al.
401 antiSMASH: rapid identification, annotation and analysis of secondary metabolite
402 biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res.
403 2011;39:W339-46. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21672958>
- 404 15. Cragg GM, Newman DJ. Natural products: A continuing source of novel drug leads.
405 Biochim Biophys Acta - Gen Subj. 2013;1830:3670–95.
- 406 16. Yuan H, Ma Q, Ye L, Piao G. The traditional medicine and modern medicine from natural
407 products. Molecules. 2016;21:559. Available from:
408 <https://pubmed.ncbi.nlm.nih.gov/27136524/>
- 409 17. Newman DJ, Cragg GM, Snader KM. Natural products as sources of new drugs over the
410 period 1981-2002. J Nat Prod. 2003;66:1022–37.
- 411 18. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0:
- 19

- 412 updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.*
413 2019;47:W81–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/31032519>
- 414 19. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, Van Der Hooft JJJ, et al.
415 MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.*
416 2020;48:D454–8. Available from: <https://json-schema.org>
- 417 20. Kautsar SA, Van Der Hooft JJJ, De Ridder D, Medema MH. BiG-SLiCE: A highly
418 scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience.*
419 2021;10:1–17. Available from: <http://orcid.org/0000-0002-2191-2821>
- 420 21. Kautsar SA, Blin K, Shaw S, Weber T, Medema MH. BiG-FAM: The biosynthetic gene
421 cluster families database. *Nucleic Acids Res.* 2021;49:D490–7. Available from:
422 <https://palletsprojects.com/>
- 423 22. Shukla V, Joshi GP, Rawat MSM. Lichens as a potential natural source of bioactive
424 compounds: A review. *Phytochem. Rev.* Springer; 2010 [cited 2021 Feb 8]. p. 303–14.
425 Available from: <https://link.springer.com/article/10.1007/s11101-010-9189-6>
- 426 23. Shrestha G, St. Clair LL. Lichens: A promising source of antibiotic and anticancer drugs.
427 *Phytochem. Rev.* 2013. p. 229–44.
- 428 24. Boustie J, Grube M. Lichens—a promising source of bioactive secondary metabolites.
429 *Plant Genet Resour.* 2005;3:273–87. Available from:
430 https://www.cambridge.org/core/product/identifier/S1479262105000328/type/journal_article
- 431 25. Lumbsch HT. Chemical Fungal Taxonomy: An Overview. In: Frisvad JC, Bridge PD,
432 Arora DK, editors. *Chem Fungal Taxon.* 1st ed. CRC Press; 1998. p. 1–18. Available from:
433 [https://www.taylorfrancis.com/chapters/edit/10.1201/9781003064626-1/chemical-fungal-](https://www.taylorfrancis.com/chapters/edit/10.1201/9781003064626-1/chemical-fungal-taxonomy-overview-jens-frisvad-paul-bridge-dilip-arora)
434 [taxonomy-overview-jens-frisvad-paul-bridge-dilip-arora](https://www.taylorfrancis.com/chapters/edit/10.1201/9781003064626-1/chemical-fungal-taxonomy-overview-jens-frisvad-paul-bridge-dilip-arora)
- 435 26. Kealey JT, Craig JP, Barr PJ. Identification of a lichen depside polyketide synthase gene
436 by heterologous expression in *Saccharomyces cerevisiae*. *Metab Eng Commun.*

- 437 2021;13:e00172. Available from:
438 <https://linkinghub.elsevier.com/retrieve/pii/S2214030121000122>
- 439 27. Meiser A, Otte J, Schmitt I, Grande FD. Sequencing genomes from mixed DNA samples -
440 Evaluating the metagenome skimming approach in lichenized fungi. *Sci Rep.* 2017;7:1–13.
441 Available from: www.nature.com/scientificreports/
- 442 28. Bertrand RL, Sorensen JL. A comprehensive catalogue of polyketide synthase gene
443 clusters in lichenizing fungi. *J. Ind. Microbiol. Biotechnol.* Springer Verlag; 2018. p. 1067–
444 81.
- 445 29. Gerasimova J V, Beck A, Werth S, Resl P. High diversity of type I polyketide genes in
446 *Bacidia rubella* as revealed by the comparative analysis of 23 lichen genomes. *J Fungi.*
447 2022;8:449. Available from: <https://doi.org/10.3390/jof8050449>
- 448 30. Davydov EA, Peršoh D, Rambold G. Umbilicariaceae (lichenized Ascomycota) – Trait
449 evolution and a new generic concept. *Taxon.* 2017;66:1282–303. Available from:
450 <https://onlinelibrary.wiley.com/doi/abs/10.12705/666.2>
- 451 31. Posner B, Feige GB, Huneck S. Studies on the chemistry of the lichen genus *Umbilicaria*
452 hoffm. *Zeitschrift fur Naturforsch - Sect C J Biosci.* 1992;47:1–9. Available from:
453 <https://www.degruyter.com/view/journals/znc/47/1-2/article-p1.xml>
- 454 32. Singh G, Calchera A, Schulz M, Drechsler M, Bode HB, Schmitt I, et al. Climate-specific
455 biosynthetic gene clusters in populations of a lichen-forming fungus. *Environ Microbiol.*
456 2021;00:1462-2920.15605. Available from:
457 <https://onlinelibrary.wiley.com/doi/10.1111/1462-2920.15605>
- 458 33. Singh G, Armaleo D, Dal Grande F, Schmitt I. Depside and depsidone synthesis in
459 lichenized fungi comes into focus through a genome-wide comparison of the olivetoric acid
460 and physodic acid chemotypes of *Pseudevernia furfuracea*. *Biomolecules.* 2021;11:1445.
461 Available from: <https://www.mdpi.com/2218-273X/11/10/1445>

- 462 34. Wang J, Nielsen J, Liu Z. Synthetic biology advanced natural product discovery.
463 Metabolites. 2021;11:785. Available from: <https://pubmed.ncbi.nlm.nih.gov/34822443/>
- 464 35. Pizarro D, Divakar PK, Grewe F, Crespo A, Dal Grande F, Lumbsch HT. Genome-wide
465 analysis of biosynthetic gene cluster reveals correlated gene loss with absence of usnic acid in
466 lichen-forming fungi. Genome Biol Evol. 2020;12:1858–68. Available from:
467 <https://academic.oup.com/gbe/article/12/10/1858/5903737>
- 468 36. Singh G, Calchera A, Merges D, Otte J, Schmitt I, Grande FD. A candidate gene cluster
469 for the bioactive natural product gyrophoric acid in lichen-forming fungi. bioRxiv.
470 2022;2022.01.14.475839. Available from:
471 <https://www.biorxiv.org/content/10.1101/2022.01.14.475839v1>
- 472 37. Robey MT, Caesar LK, Drott MT, Keller NP, Kelleher NL. An interpreted atlas of
473 biosynthetic gene clusters from 1,000 fungal genomes. Proc Natl Acad Sci U S A. 2021;118.
474 Available from: <https://doi.org/10.1073/pnas.2020230118>
- 475 38. Lautié E, Russo O, Ducrot P, Boutin JA. Unraveling plant natural chemical diversity for
476 drug discovery purposes. Front. Pharmacol. Frontiers Media S.A.; 2020. p. 397.
- 477 39. Navarro-Muñoz JC, Selem-Mojica N, Mallowney MW, Kautsar SA, Tryon JH, Parkinson
478 EI, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem
479 Biol. 2020;16:60–8. Available from: <https://pubmed.ncbi.nlm.nih.gov/31768033/>
- 480 40. Alam K, Islam MM, Li C, Sultana S, Zhong L, Shen Q, et al. Genome mining of
481 *Pseudomonas* species: diversity and evolution of metabolic and biosynthetic potential.
482 Molecules. 2021 [cited 2022 Mar 8];26:7524. Available from: [https://www.mdpi.com/1420-](https://www.mdpi.com/1420-3049/26/24/7524)
483 [3049/26/24/7524](https://www.mdpi.com/1420-3049/26/24/7524)
- 484 41. Choudoir MJ, Pepe-Ranne C, Buckley DH. Diversification of secondary metabolite
485 biosynthetic gene clusters coincides with lineage divergence in *Streptomyces*. Antibiotics.
486 2018;7:1–15. Available from: <https://pubmed.ncbi.nlm.nih.gov/29438308/>

- 487 42. Ingelfinger R, Henke M, Roser L, Ulshöfer T, Calchera A, Singh G, et al. Unraveling the
488 pharmacological potential of lichen extracts in the context of cancer and inflammation with a
489 broad screening approach. *Front Pharmacol.* 2020;11:1322. Available from:
490 <https://www.frontiersin.org/article/10.3389/fphar.2020.01322/full>
- 491 43. Manojlović N, Ranković B, Kosanić M, Vasiljević P, Stanojković T. Chemical
492 composition of three *Parmelia* lichens and antioxidant, antimicrobial and cytotoxic activities
493 of some their major metabolites. *Phytomedicine.* 2012;19:1166–72.
- 494 44. Cardile V, Graziano ACE, Avola R, Piovano M, Russo A. Potential anticancer activity of
495 lichen secondary metabolite physodic acid. *Chem Biol Interact.* 2017;263:36–45.
- 496 45. Shi J, Zeng YJ, Zhang B, Shao FL, Chen YC, Xu X, et al. Comparative genome mining
497 and heterologous expression of an orphan NRPS gene cluster direct the production of
498 ashimides. *Chem Sci.* 2019;10:3042–8. Available from:
499 <https://pubmed.ncbi.nlm.nih.gov/30996885/>
- 500 46. Mattern DJ, Schoeler H, Weber J, Novohradská S, Kraibooj K, Dahse HM, et al.
501 Identification of the antiphagocytic tryptacidin gene cluster in the human-pathogenic fungus
502 *Aspergillus fumigatus*. *Appl Microbiol Biotechnol.* 2015;99:10151–61. Available from:
503 <https://pubmed.ncbi.nlm.nih.gov/26278536/>
- 504 47. Buijs Y, Isbrandt T, Zhang S-D, Larsen TO, Gram L. The antibiotic andrimid produced by
505 *Vibrio coralliilyticus* increases expression of biosynthetic gene clusters and antibiotic
506 production in *Photobacterium Galathea*e. *Front Microbiol.* 2020;11:3276. Available from:
507 <https://www.frontiersin.org/articles/10.3389/fmicb.2020.622055/full>
- 508 48. Ziko L, Saqr AHA, Ouf A, Gimpel M, Aziz RK, Neubauer P, et al. Antibacterial and
509 anticancer activities of orphan biosynthetic gene clusters from Atlantis II Red Sea brine pool.
510 *Microb Cell Fact.* 2019;18:56. Available from:
511 <https://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-019-1103-3>

- 512 49. Cox-Georgian D, Ramadoss N, Dona C, Basu C. Therapeutic and medicinal uses of
513 terpenes. *Med Plants From Farm to Pharm.* 2019. p. 333–59. Available from:
514 </pmc/articles/PMC7120914/>
- 515 50. Guimarães AC, Meireles LM, Lemos MF, Guimarães MCC, Endringer DC, Fronza M, et
516 al. Antibacterial activity of terpenes and terpenoids present in essential oils. *Molecules.*
517 2019;24:2471. Available from: </pmc/articles/PMC6651100/>
- 518 51. Jiang M, Wu Z, Guo H, Liu L, Chen S. A review of terpenes from marine-derived fungi:
519 2015-2019. *Mar Drugs.* 2020;18:321. Available from: www.mdpi.com/journal/marinedrugs
- 520 52. Del Prado-Audelo ML, Cortés H, Caballero-Florán IH, González-Torres M, Escutia-
521 Guadarrama L, Bernal-Chávez SA, et al. Therapeutic applications of terpenes on
522 inflammatory diseases. *Front Pharmacol.* 2021;12:2114. Available from:
523 <https://www.frontiersin.org/articles/10.3389/fphar.2021.704197/full>
- 524 53. Jaeger R, Cuny E. Terpenoids with special pharmacological significance: A review. *Nat*
525 *Prod Commun.* 2016;11:1373–90.
- 526 54. Yang W, Chen X, Li Y, Guo S, Wang Z, Yu X. Advances in pharmacological activities of
527 terpenoids. *Nat. Prod. Commun.* 2020. p. 1–13. Available from:
528 <https://journals.sagepub.com/doi/full/10.1177/1934578X20903555>
- 529 55. Park SY, Choi J, Lee GW, Jeong MH, Kim JA, Oh SO, et al. Draft genome sequence of
530 *Umbilicaria muelenbergii* KoLRILF000956, a lichen-forming fungus amenable to genetic
531 manipulation. *Genome Announc.* 2014;2:e00357. Available from:
532 <https://pubmed.ncbi.nlm.nih.gov/24762942/>
- 533 56. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using
534 repeat graphs. *Nat Biotechnol.* 2019;37:540–6. Available from:
535 <https://www.nature.com/articles/s41587-019-0072-8>
- 536 57. Qin M, Wu S, Li A, Zhao F, Feng H, Ding L, et al. LRScf: Improving draft genomes

537 using long noisy reads. BMC Genomics. 2019;20:955. Available from:
538 <https://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-019-6337-2>
539 58. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community
540 Edition - Interactive exploration and analysis of large-scale microbiome sequencing data.
541 PLOS Comput Biol. 2016;12:e1004957. Available from:
542 <https://dx.plos.org/10.1371/journal.pcbi.1004957>
543 59. Palmer J, Stajich J. Funannotate v1.7.4. Zenodo. 2019;
544