1 **Stone-dwelling actinobacteria Blastococcus saxobsidens, Modestobacter marinus &**
2 **Geodermatophilus *obscurus* proteogenomes**

3 Haïtham Sghaier[1], Karima Hezbri[2], Faten Ghodhbane-Gtari[2], Petar Pujic[3], Arnab Sen[4], Daniele
4 Daffonchio[5], Abdellatif Boudabous[2], Louis S Tisa[6], Hans-Peter Klenk[7], Jean Armengaud[8], Philippe
5 Normand[3*], Maher Gtari[2]
6
7 [1] National Center for Nuclear Sciences and Technology, Sidi Thabet Technopark, 2020 Ariana, Tunisia.
8 [2] Laboratoire Microorganismes et Biomolécules Actives, Université de Tunis Elmanar (FST) & Université de Carthage
9 (INSAT), Tunis, 2092, Tunisia.
10 [3] Université de Lyon, Université Lyon 1, Lyon, France; CNRS, UMR 5557, Ecologie Microbienne, 69622
11 Villeurbanne, Cedex, France.
12 [4] NBU Bioinformatics Facility, Department of Botany, University of North Bengal, Siliguri, 734013, India.
13 [5] King Abdullah University of Science and Technology (KAUST), BESE, Biological and Environmental Sciences and
14 Engineering Division, Thuwal, 23955-6900, Kingdom of Saudi Arabia & Department of Food, Environmental and
15 Nutritional Sciences (DeFENS), University of Milan, via Celoria 2, 20133 Milan, Italy.
16 [6] Department of Molecular, Cellular & Biomedical Sciences, University of New Hampshire, 46 College Road, Durham,
17 NH 03824-2617, USA.
18 [7] School of Biology, Newcastle University, Ridley Building, Newcastle upon Tyne, NE1 7RU, United Kingdom.
19 [8] CEA, DSV, IBITEC-S, Lab Innovative technologies for Detection and Diagnostic, Bagnols-sur-Cèze, F-30207,
20 France.
21
22 [*] To whom correspondence should be addressed: Philippe Normand, Université Lyon 1, Lyon, France; 69622
23 Villeurbanne, France. Tel: + 33 4 72431676, e-mail: philippe.normand@univ-lyon1.fr
24
25 **Running title:** Proteogenomes of actinobacterial pioneers.
26
27 **Subject Category:** Integrated genomics and post-genomics approaches in microbial ecology.
28

29 **Abstract**

30 The Geodermatophilaceae are unique model systems to study the ability to thrive on or within
31 stones and their proteogenomes (referring to the whole protein arsenal encoded by the genome)
32 could provide important insight into their adaptation mechanisms. Here, we report the detailed
33 comparative genome analysis of *Blastococcus saxobsidens* (Bs), *Modestobacter marinus* (Mm) and
34 *Geodermatophilus obscurus* (Go) isolated respectively from the interior and the surface of
35 calcarenite stones and from desert sandy soils. The genome-scale analysis of Bs, Mm and Go
36 illustrates how adaptation to these niches can be achieved through various strategies including
37 "molecular tinkering/opportunism" as shown by the high proportion of lost, duplicated or
38 horizontally transfered genes and ORFans. Using high-throughput discovery proteomics, the three
39 proteomes under unstressed conditions were analyzed, highlighting the most abundant biomarkers
40 and the main protein players. Proteomic data corroborated previously demonstrated stone-related
41 ecological distribution. For instance, these data showed starvation-inducible, biofilm-related and
42 DNA-protection proteins as signatures of the microbes associated with the interior, surface and
43 outside of stones, respectively.

44

45 **Keywords:** *Blastococcus* / *Geodermatophilus* / *Modestobacter* / proteogenome / soil / stone.

46

Introduction

Geodermatophilaceae are an actinobacterial family (Normand, 2006) in the order Geodermatophiliales (Sen *et al*., 2014) that comprises three genera: *Geodermatophilus*, *Blastococcus* and *Modestobacter* initially isolated from desert soils (Luedemann, 1968), sea water (Ahrens and Moll, 1970) and Antarctic regolith (Mevs *et al*., 2000), respectively. These three genera have a complex life cycle, and produce remarkably resistant enzymes such as esterases (Essoussi *et al*., 2010; Jaouani *et al*., 2012; Normand *et al*., 2014). They also have the ability to resist adverse environmental conditions such as ultraviolet (UV) light, ionizing radiation (IR), desiccation and heavy metals (Gtari *et al*., 2012; Montero-Calasanz *et al*., 2014; Montero-Calasanz *et al*., 2015; Rainey *et al*., 2005). This resistance to environmental hazards represents a trait of Terrabacteria, a well-supported phylogenetic group composed of Actinobacteria and four other major lineages of eubacteria (Firmicutes, Cyanobacteria, Chloroflexi and *Deinococcus*-*Thermus*) that colonized land 3.05–2.78 Ga (Battistuzzi *et al*., 2004; Battistuzzi and Hedges, 2009; Tunnacliffe and Lapinski, 2003).

Surprisingly, Geodermatophilaceae are present in a variety of biotopes including prominently rocks (Eppard *et al*., 1996) and desert sandy soils (Liu *et al*., 2014; Montero-Calasanz *et al*., 2013; Montero-Calasanz *et al*., 2013; Montero-Calasanz *et al*., 2013; Montero-Calasanz *et al*., 2013; Montero-Calasanz *et al*., 2013; Montero-Calasanz *et al*., 2013; Montero-Calasanz *et al*., 2013). Although considered endemic to soils, evolution of Geodermatophilaceae has continued in specialized land biotopes. Indeed, soil and stone niches have yielded a wealth of knowledge regarding the extant distribution of Geodermatophilaceae (Gtari *et al*., 2012; Normand *et al*., 2014), raising questions about their evolution and the mechanisms of adaptation to harsh environments.

After their uplift by storms, Geodermatophilaceae have the potential to travel thousands of kilometers in the atmosphere (Chuvochina *et al*., 2011). Consequently, stone surfaces can be colonized by these wind-borne microbes (Essoussi *et al*., 2012). These surfaces are often covered

72 with growth (called “patinas”, “varnish-like” or “tintenstriches”), that are comprised of complex
73 communities of eukaryotes and prokaryotes, recurrent among which are Actinobacteria (Eppard *et*
74 *al*., 1996; Kuhlman *et al*., 2006) including Geodermatophilaceae (Urzi *et al*., 2001). With regard to
75 biopitting, one hypothesis is that acid secretion and high carbon dioxide (CO_2) emissions from
76 combustion engines result in alternating episodes of calcareous solubilization and precipitation. The
77 microbial communities located in biopits have been analyzed by microbiological and molecular
78 methods and found to be complex, with a recurrence of Geodermatophilaceae. To deduce strategies
79 for the dispersal of established biofilms and propose restoration approaches, identification of the
80 components of the matrix of these biopolymers is required (Nijland *et al*., 2010).

81 To understand how Geodermatophilaceae adapt to stones and soil, here we present a
82 proteogenome analysis and detailed comparison of Bs (Chouaia *et al*., 2012), Mm (Normand *et al*.,
83 2012) and Go (Ivanova *et al*., 2010). To get a close-up view of their physiology, we analyzed the
84 proteome content of stationary-phase Bs, Mm and Go cells by a liquid chromatography (LC)-
85 tandem mass spectrometry (MS/MS) shotgun approach and semi-quantification by spectral
86 counting. We focussed on the identification of interesting proteic biomarkers of potential
87 physiological value. Bs, Mm and Go are the first Geodermatophilaceae whose genomes and
88 proteomes have been analysed jointly. Now, they represent new models of choice for studies of
89 niche adaptation amongst Terrabacteria.

90

91 **Materials and methods**

92 ***Bioinformatics approaches***

93 Genes were classed into Clusters of Orthologous Groups (COG) (Tatusov *et al*., 2001) and retrieved
94 from the Mage platform (Vallenet *et al*., 2006). Metabolic pathways were analysed using BioCyc
95 (Caspi *et al*., 2010). Identification of duplicated, lost or horizontally transferred genes was done
96 using the Mage platform (phyloprofile) as was the identification of the core and extended genome.

97 Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) were identified with the
98 CRISPI database http://crispi.genouest.org/ (Rousseau *et al*., 2009). Genomic islands were
99 identified with IslandViewer available at http://www.pathogenomics.sfu.ca/islandviewer/query.php
100 (Langille and Brinkman, 2009). Phylogenetic analysis was done using MEGA6 (Tamura *et al*.,
101 2013); and the inferred topology was drawn and integrated with the genomic context at the
102 Microbial Genomic context Viewer (MGcV) accessible at http://mgcv.cmbi.ru.nl/ (Overmars *et al*.,
103 2013).
104 Correspondence analysis was done as previously described (Benzécri, 1973) using the R
105 software (R Development Core Team, 2007) on COG numbers (obtained from the Mage
106 platform/Genomic Tools), and on numbers of Coding DNA Sequences (CDS) containing
107 Transcription and Signaling domains (obtained from the Mage platform/Search Interpro) as
108 previously described (Santos *et al*., 2009).
109
110 ***Bacterial growth and proteome sample preparation***
111 Cells of *B. saxobsidens* DD2 (Bs), *M. marinus* BC501 (Mm) and *G. obscurus* DSM 43160 (Go)
112 were plated onto Luedemann medium (yeast extract, malt extract, glucose, soluble starch and
113 calcium carbonate ($CaCO_3$)) and incubated for 72h at 28°C as described previously (Luedemann,
114 1968). Bacterial cultures (15 mg wet weight) were resuspended into 90 µl of lithium dodecyl sulfate
115 β-mercaptoethanol protein gel sample buffer (Invitrogen), and incubated at 99ºC for 5 min as
116 indicated previously (Hartmann *et al*., 2014). Prior to SDS-PAGE analysis on 10% Bis-Tris
117 NuPAGE gels (Invitrogen), the samples were briefly centrifuged to remove large aggregates. A
118 volume of 20 µl of the proteome of Bs, Mm and Go (corresponding to 160 µg of total proteins) was
119 loaded per well. Three independent biological replicates were analyzed per microorganism. SDS-
120 PAGE was carried out in 1X 3-(N-morpholino) propanesulfonic acid solution (Invitrogen) on a
121 XCell SureLock Mini-cell (Invitrogen) under a constant voltage of 200 V for 5 min. Gels were

122 stained with SimplyBlue SafeStain, a ready-to-use Coomassie G-250 stain (Invitrogen). SeeBlue
123 Plus2 (Invitrogen) was used as a molecular weight marker. Polyacrylamide gel bands (equivalent in
124 volume to 50 µl) comprising the entire proteomes—one band per entire proteome—were cut and
125 processed for in-gel proteolysis with Trypsin Sequencing Grade (Roche) followed by the
126 ProteaseMax protocol (Promega) as described previously (Clair *et al*., 2010).

127

128 ***NanoLC-MS/MS Analysis***

129 Peptide digests were resolved on an UltiMate 3000 LC system (Dionex-LC Packings) prior to
130 MS/MS measurements done with a LTQ-Orbitrap XL (Thermo-Fisher) as described previously
131 (Dedieu *et al*., 2011). MS/MS spectra were processed and interpreted with the MASCOT 2.3.02
132 search engine (Matrix Science) with standard parameters as indicated previously (Hartmann and
133 Armengaud, 2014) against databases corresponding to a complete list of annotated CDS from either
134 Bs (NCBI RefSeq: NC_016943.1), Mm (BioProject: PRJEA167487, PRJEA82845) or Go
135 (BioProject: PRJNA43725, PRJNA29547). Peptide matches with a score above their peptidic
136 identity threshold were filtered at $p<0.05$. A protein was only validated when at least two peptides
137 had been assigned. Using a previously described approach (Liu *et al*., 2004; Zivanovic *et al*., 2009),
138 protein abundance was evaluated by shotgun analysis using MS/MS spectral counts. Normalized
139 Spectral count Abundance Factors (NSAF) were calculated (Paoletti *et al*., 2006). The sum of all
140 NSAF—100%—was calculated for each bacterium: Bs (142.63), Mm (46.38) and Go (162.87).
141 Accordingly, all values in this manuscript were separated from the locus tags with a comma or
142 between parentheses and represent NSAF percentages, unless otherwise stipulated. The mass
143 spectrometry proteomics data have been deposited at the open access library of ProteomeXchange
144 Consortium (http://www.proteomexchange.org/) (Vizcaino *et al*., 2014) with the data set identifiers
145 PXD001519, PXD001518 and PXD001520 for Bs, Mm and Go, respectively.

146

147 **Results and Discussion**

148 ***Characteristics of proteogenomes***

149 Life in biotopes with low trophic resources has driven the three Geodermatophilaceae
150 members toward medium-sized genomes (Chouaia *et al*., 2012; Ivanova *et al*., 2010; Normand *et*
151 *al*., 2012) from 4.87 to 5.32 mega base pairs (Mb) (Figure 1; Supplementary Table 1). The three
152 plasmid-less genomes had very high G+C% (72.95–74.1%). Under unstressed conditions, the three
153 proteomes were analysed by a high-throughput shotgun procedure (Christie-Oleza and Armengaud,
154 2010). For Bs (PXD001519, 39889 MS/MS spectra, Additional data 1), Mm (PXD001518, 14729
155 MS/MS spectra, Additional data 2) and Go (PXD001520, 14829 MS/MS spectra, Additional data
156 3), 5506, 1940 and 6884 spectra could be assigned to 553, 100 and 370 proteins, respectively. These
157 three datasets represent the first proteogenome references for Geodermatophilaceae. Figure 1
158 depicts the proteins detected in this study.

159 We have predicted 3,277 genes (30% amino acid identity) shared by the three genomes,
160 which were sorted into seven possible and mutually exclusive Venn groups (Supplementary Figure
161 1). Significant similarity to previously reported genes of known function allowed us to assign a
162 putative function to 3,231, 3,643 and 3,351 protein-coding genes in Bs, Mm and Go, respectively.
163 The remaining genes were designated "proteins with unknown function". The analysed proteomes
164 of Bs, Mm and Go indicate the expression of four, two and three such "proteins of unknown
165 function" (Supplementary Table 2) among the most highly expressed proteins—55, nine and 42 as
166 summarised in Supplementary Figure 2—that account for half of the total number of assigned
167 spectra, respectively. As will become apparent below, some of these "proteins of unknown
168 function" seem to play a primordial role in niche adaptation of host bacteria.

169 The three genomes have 70% of their CDS that could be ascribed to the COG category
170 (Supplementary Table 3) (Tatusov *et al*., 2001; Vallenet *et al*., 2006). Correspondance analysis
171 showed that the three Geodermatophilaceae genomes are close to one another. All three,

172 particularly Bs and Mm, have a high proportion of [T] (signal) and [P] (inorganic ion transport and
173 metabolism) categories, which is evocative of a lifestyle in a mineral-rich biotope (Supplementary
174 Figure 3A). The overall distribution of the COG profile as well as the abundance of transcription
175 factors and signaling molecules (Supplementary Figures 3B and 3C), constitute a signature that may
176 be associated to lifestyle (Santos *et al*., 2009). Compared to other Actinobacteria, the three genomes
177 contained the highest absolute numbers of CDS containing PAS and EAL domains (Supplementary
178 Figure 3C).

179 Our proteomic results showed that the most represented COG categories among the most
180 highly expressed proteins are: [J] translation, ribosomal structure and biogenesis (~33%),
181 [C] energy production and conversion (~22%) and [I] lipid transport and metabolism (~9%) in Bs;
182 [G] carbohydrate transport and metabolism (~33%), [E] amino acid transport and metabolism
183 (~22%) and [P] (~22%) in Mm and [J] (~33%), [C] (~26%) and [P] (~7%) in Go. The category
184 [P] is represented by more than 5% of the most highly expressed proteins of Bs. These proteomic
185 results are a further support for the presence of the monophyletic group composed of Bs and Go,
186 that has been confirmed *in silico* by bioinformatics (Sen *et al*., 2014) as well as *in vitro* by
187 microbiological and biochemical markers (Normand *et al*., 2014).

188

189 ***"Molecular tinkering/opportunism"strategies***

190 Bs, Mm and Go genomes exhibit at least three strategies related to "molecular
191 tinkering/opportunism" (Gogarten *et al*., 2002; Jacob, 1977; Laubichler, 2007) (Supplementary
192 Table 4): (i) domain duplication, (ii) horizontal gene transfer (HGT)—genes absent in two of the
193 analyzed genomes but present in one of a group of more distant Actinobacteria—and (iii) rapid
194 evolution—to create ORFans (Daubin and Ochman, 2004; Fukuchi and Nishikawa, 2004). Mm has
195 429 duplicated genes (7.9% of the genome) while in Bs and Go the number of these genes is
196 slightly lower (representing 4.8 and 6.5%, respectively). Removal of a complex nutrient induced a

197 motile state in these bacteria—motile budding rods called R-forms (Ishiguro and Wolfe, 1970;
198 Ishiguro and Wolfe, 1974). The most highly expressed protein in Bs and Mm was flagellin
199 synthesis.

200 lin (BLASA_0851, 4.96%; MODMU_1040, 11.51%) encoded by a duplicated gene that has
201 a paralog encoding a flagellar hook-associated protein FigL (BLASA_0855; MODMU_1044),
202 respectively. The same duplication event was observed in Go between similar paralogs, a flagellin
203 (Gobs_0985, 0.02%) and a FigL (Gobs_0990). Rates of HGT amount to 6.9, 8.9 and 6.8% in Bs,
204 Mm and Go, respectively, which is consistant with hostile rock environments where antibiosis is
205 often an unaffordable luxury (Friedmann and Ocampo-Friedmann, 1984). For example, our
206 proteomic analyses indicated the presence of a transposase in Bs (BLASA_4384, 0.01%). The
207 number of ORFans with 7.4–10.2% of the three genomes is much higher than the number found for
208 *E. coli* (3.5%) (Daubin and Ochman, 2004), a difference that could be linked to an unexpectedly
209 higher rate of HGT in, on the surface and outside stones than in the promiscuous gastrointestinal
210 tract. Our proteomic approach allowed the identification of five, two and six ORFans that may have
211 important functional roles in Bs, Mm and Go, respectively (Supplementary Table 2). A computation
212 of lost genes—genes present in two genomes but absent in the third—shows that Bs, by far, had the
213 highest number of lost genes (515 CDS) (Supplementary Table 4).
214

215 ***First line defense strategies***

216 Genome analyses indicate that Bs, Mm and Go possess several genes putatively involved in
217 carotenoid biosynthesis (Supplementary Table 5). The Bs orange pigment absorbs at 230–270 nm
218 and at 450–500 nm; whereas Mm and Go pigments are quite comparable and absorb almost
219 continuously between 200 and 750 nm (Supplementary Figure 4). Bs has a putative operon
220 (BLASA_0209–0214, *crtB2*, *hopC*, *ispA*, *shc*, *hpnH*, *ilvC*) that is absent in Mm and Go; and this

221 could explain its intense orange pigment. Moreover, the expression of an uncharacterized enzyme
222 involved in pigment biosynthesis (BLASA_3306, 0.05%) was detected in Bs.

223 The three genomes also possess impressive arrays of genes involved in stress relief, reactive
224 oxygen species (ROS—superoxide anions ($O2^{\bullet -}$), hydrogen peroxide (H_2O_2) and hydroxyl radicals
225 ($HO^{\bullet}$)) detoxification and DNA protection and repair (Supplementary Table 6). The Bs, Mm and Go
226 proteomes express a nickel-containing superoxide dismutase (SOD) (BLASA_3991, 0.83%;
227 MODMU_4573, 0.57%; Gobs_4176, 0.29%) and two catalases. Catalase (KatE) was one of the
228 most highly expressed proteins in Mm (MODMU_2078, 4.64%). Its ortholog, Gobs_2125 (0.74%),
229 belongs to the first 42 proteins that accounted for half of the total number of assigned spectra to the
230 proteome of Go. Bs also highly expressed KatE (BLASA_3094, 0.54%). In addition, Bs expressed a
231 manganese-containing catalase (KatA: BLASA_0196, 0.04%), but to a lesser extent than Go KatE
232 (only 2 versus 75 spectral counts, when cumulating data from the triplicated samples). The
233 transformation of the non-essential amino acid sarcosine, a source of carbon and energy derived
234 from the osmoprotectant betaine, into the essential amino acid glycine generates H_2O_2. The
235 sarcosine oxidase subunits (*soxB*, *soxD* and *soxA*) and carbon monoxide (CO) dehydrogenase
236 subunit G (*coxG*) genes form an operon in ROS-resistant Mm and Go (MODMU_3072–
237 MODMU_3075 and Gobs_2883–Gobs_2880, respectively), and were not detected in the ROS-
238 sensitive Bs (Gtari *et al*., 2012). Supplementary Table 7 lists other selected physiological features
239 present in Bs, Mm and Go. For instance, in accordance with our previously published experimental
240 data (Gtari *et al*., 2012), orthologs of metal tolerance determinants (Janssen *et al*., 2010) were
241 detected in the three studied genomes.

242 Linear density of genomic DNA double-strand breaks (DSBs) inflicted per gray (Gy) per
243 mega base pairs (Mbp) (0.002–0.004) is similar for diverse bacteria ((Daly, 2009; Daly, 2011) and
244 references therein). Acute doses of 0.9, 6 and 9 kGy (Gtari *et al*., 2012) are predicted to inflict ~18,
245 ~128 and ~192 DSBs in Bs (~4.88 Mbp), Mm (~5.33 Mbp) and Go (~5.33 Mbp), respectively.

246 Although absent in almost all actinobacterial species, all three genomes contained a non-
247 homologous end joining (NHEJ) operon, BLASA_3099–BLASA_3097 in Bs, MODMU_2074–
248 MODMU_2076 in Mm and Gobs_2119–Gobs_2121 in Go, (Supplementary Figure 5). Bs possesses
249 a supplementary putative Ku protein (BLASA_1744) that forms another operon with "proteins of
250 unknown function". These findings suggest that NHEJ may be a major pathway of DSBs repair in
251 Geodermatophilaceae.
252
253 ***Most abundant proteomic biomarkers and niche signatures***
254 Bs was isolated from deep (2cm) from stones found around the Mediterranean using a chisel
255 and hammer to eliminate the surface layers (Urzi *et al*., 2004); and it is predominant in the deeper
256 fraction—about 2 cm below the stone surface (Gtari *et al*., 2012). The proteomic analysis results
257 (Supplementary Figure 2) suggest that unstressed Bs has evolved a survival strategy inside stones
258 based on: (i) heavy investment in protein synthesis and in preventing their aggregation (ribosomal
259 proteins and GroEL); (ii) detection and response to the changes of the environmental external
260 nutrients (UDP-glucose, phosphate, *etc*.) concentrations (LysM, UshA, PhoU) (Buist *et al*., 2008;
261 Marzan and Shimizu, 2011); (iii) scavenging ROS (SOD) and (iv) transport of oxygen
262 (hemerythrin). The presence of enzymes using anaerobic terminal electron acceptors in Bs, Mm and
263 Go (Supplementary Table 7) indicates that formate (HCO_2^-) and nitrate (NO_3^-) anaerobic
264 respiration may be possible.
265 We also discovered four highly expressed biomarkers (MODMU_0153, 5.83%;
266 MODMU_0507, 5.41%; MODMU_1130, 3.92%; MODMU_3547, 3.17%) of Mm, isolated from a
267 white marble surface (Carrara, Tuscany, Italy) (Urzi *et al*., 2001)—and predominant in the upper
268 fraction (about 2 mm of stone surface) (Gtari *et al*., 2012), (Supplementary Figure 2). Orthologs of
269 these four highly expressed biomarkers are proteins implicated in the development of biofilms.

270 The outside of stone has also been investigated (Berdoulay and Salvado, 2009; Macedo *et*
271 *al*., 2009; Urzi *et al*., 2001). Growth on stone surfaces means either reliance on photosynthesis or on
272 nutrients carried by rain, air or through the stone itself. Concerning operons encoding genes for
273 photosynthesis reactions, Go, isolated from soil of the Amargosa Desert (Nevada, USA)
274 (Luedemann, 1968), contains three (Gobs_1696–Gobs_1703, Gobs_4550–Gobs_4544, Gobs_4558–
275 Gobs_4551) and Bs contains two (BLASA_0681, BLASA_2555–BLASA_2552). Surprisingly,
276 Mm contains only a single NADPH-ferredoxin reductase (*fprA*) gene (MODMU_0890). Two genes
277 encoding ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) were identified only in Go
278 (Gobs_1448, 0.03%; Gobs_2026) suggesting that the strain obtains both carbon and energy via
279 carboxydotrophy. Also, Go is characterized by the presence of DNA-related biomarkers including a
280 highly expressed DNA-binding histone-like protein (Gobs_0298, 1.13%) (Supplementary Figure 2).
281 In contrast to Mm, both Go and Bs have many similar highly expressed proteomic biomarkers like a
282 Dps-like iron-chelating protein (Gobs_3661 and BLASA_1121) that may limit, through the
283 confinement of free iron, the Fenton derived production of $HO^{\bullet}$ (Confalonieri and Sommer, 2011;
284 Williams *et al*., 2007). Yet, Bs has more highly expressed biomarkers associated with the
285 production of ROS (Supplementary Figure 2)—cytochromes, flavoproteins, *etc*.—representing a
286 cellular benchmark for the proclivity of cells to resist to stress like IR ((Ghosal *et al*., 2005) and
287 references therein). Given this current state of affairs, it comes as no surprise that Bs is ROS- and
288 IR-sensitive (Gtari *et al*., 2012).

289 Rain is known to carry nitrogen compounds (Singh and Agrawal, 2008) such as nitric acid
290 as well as traces of sulfur (Raybould *et al*., 1977). Besides glutamine synthetase and glutamate
291 synthase for ammonium assimilation, the three genomes contain a conserved operon coding for an
292 ammonium transporter (AmtB) and a nitrogen regulatory protein (GlnB) (Supplementary Table 7).
293 Air carries numerous volatile organic compounds, prominent among which is CO (Austin *et al*.,
294 2001). Study of the genomes revealed that Bs, Mm and Go have several copies of the *coxLMS*

295 operon (Supplementary Table 7), which would help them oxidize CO. Such multiple copies (Wu *et*
296 *al*., 2005) are always an indication of a strong selective pressure (Lee *et al*., 2009; Oda *et al*., 2005).
297 Expression of the proteins—CO dehydrogenase subunits and acetyl-coenzyme A synthetase
298 (AcsA)—(Supplementary Table 2) corresponding to some of the identified orthologous genes
299 (Supplementary Table 7) of the Wood–Ljungdahl pathway has been detected under standard growth
300 conditions of Bs and Go, suggesting metabolic utility of this pathway. Contrarily to CoxM and
301 CoxL, AcsA protein was not detected in Mm (Supplementary Table 2). Thus, the three strains
302 inhabit exacting biotopes, which necessitate a rich array of transport systems, storage components, a
303 motility machinery and energy generating pathways (Figure 3). These biomarkers have shed new
304 light on the microniche signature for each rock dweeling terrabacteria.
305

306 ***Conclusion and perspectives***

307 Here, the complete genome sequences of three Geodermatophilaceae members, Bs, Mm and
308 Go, with contrasted physiologies and ecological microniches (Normand *et al*., 2014), together with
309 the analysis of their proteomes under unstressed conditions, should help provide a solid foundation
310 for investigating the varied strategies to adapt to their lifestyles. Particularly, comparison of the
311 three genomes provided an opportunity to analyze how Bs, Mm and Go can respond to stresses such
312 as ROS mainly via pigmentation and catalase production and DSB through the NHEJ pathway.
313 Moreover, highly expressed proteomic biomarkers of Bs, Mm and Go were depicted. The
314 identification of thse biomarkers have shed new light on the physiological and biochemical traits
315 that are unique to each species and its ecological microniche. Particularly, the Mm exoproteome
316 was almost as dominant as the cellular proteome, which hinders a deeper proteomic view
317 (Armengaud *et al*., 2012). Undoubtedly, much of the future progress in studying Bs, Mm and Go
318 rests squarely on the shoulders of research performed with their stressed proteogenomes.
319

320 **Acknowledgements**

321 Thanks are expressed to Danis Abrouk (University of Lyon) for help with statistics. The support of
322 the EU project BIODESERT (European Community's Seventh Framework Programme CSA-SA
323 REGPOT-2008-2 under grant agreement no. 245746) is acknowledged.

324

325 **Conflict of interest**

326 The authors declare no conflict of interest.

327

328 Supplementary Information accompanies this paper on The ISME Journal website.

329

330 **References**

331 Ahrens R, Moll G (1970). Ein neues knospendes Bakterium aus der Ostsee. [A new budding
332 bacterium from the Baltic Sea]. *Arch Mikrobiol* **70**, 243–265.
333 Armengaud J, Christie-Oleza JA, Clair G, Malard V, Duport C (2012). Exoproteomics: exploring
334 the world around biological systems. *Expert Rev Proteomics* **9**, 561–575.
335 Austin CC, Wang D, Ecobichon DJ, Dussault G (2001). Characterization of volatile organic
336 compounds in smoke at experimental fires. *J Toxicol Environ Health A* **63**, 191–206.
337 Battistuzzi FU, Feijao A, Hedges SB (2004). A genomic timescale of prokaryote evolution: insights
338 into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol*
339 **4**, 44.
340 Battistuzzi FU, Hedges SB (2009). A major clade of prokaryotes with ancient adaptations to life on
341 land. *Mol Biol Evol* **26**, 335–343.
342 Beech IB, Gaylarde CC (1999). Recent advances in the study of biocorrosion - an overview. *Revista*
343 *de Microbiologia* **30**, 177–190.
344 Benzécri J-P (1973). L'Analyse des Données. Volume II. L'Analyse des Correspondances. Paris,
345 France Dunod.
346 Berdoulay M, Salvado JC (2009). Genetic characterization of microbial communities living at the
347 surface of building stones. *Lett Appl Microbiol* **49**, 311–316.
348 Buist G, Steen A, Kok J, Kuipers OP (2008). LysM, a widely distributed protein motif for binding
349 to (peptido)glycans. *Mol Microbiol* **68**, 838–847.
350 Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F *et al*. (2010). The MetaCyc database
351 of metabolic pathways and enzymes and the BioCyc collection of pathway/genome
352 databases. *Nucleic Acids Res* **38**, D473–479.
353 Chouaia B, Crotti E, Brusetti L, Daffonchio D, Essoussi I, Nouioui I *et al*. (2012). Genome
354 sequence of *Blastococcus saxobsidens* DD2, a stone-inhabiting bacterium. *J Bacteriol* **194**,
355 2752–2753.
356 Christie-Oleza JA, Armengaud J (2010). In-depth analysis of exoproteomes from marine bacteria
357 by shotgun liquid chromatography-tandem mass spectrometry: the *Ruegeria pomeroyi* DSS-
358 3 case-study. *Mar Drugs* **8**, 2223–2239.
359 Chuvochina MS, Marie D, Chevaillier S, Petit JR, Normand P, Alekhina IA *et al*. (2011).
360 Community variability of bacteria in alpine snow (Mont Blanc) containing Saharan dust
361 deposition and their snow colonisation potential. *Microbes Environ* **26**, 237–247.
362 Clair G, Roussi S, Armengaud J, Duport C (2010). Expanding the known repertoire of virulence
363 factors produced by *Bacillus cereus* through early secretome profiling in three redox
364 conditions. *Mol Cell Proteomics* **9**, 1486–1498.
365 Confalonieri F, Sommer S (2011). Bacterial and archaeal resistance to ionizing radiation. *J Phys:*
366 *Conf Ser* **261**, doi:10.1088/1742-6596/1261/1081/012005.
367 Costerton JW, Stewart PS, Greenberg EP (1999). Bacterial biofilms: a common cause of persistent
368 infections. *Science* **284**, 1318–1322.
369 Daly MJ (2009). A new perspective on radiation resistance based on *Deinococcus radiodurans*. *Nat*
370 *Rev Microbiol* **7**, 237–245.
371 Daly MJ (2011). Death by protein damage in irradiated cells. *DNA Repair (Amst)* **11**, 12–21.
372 Daubin V, Ochman H (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in
373 *E. coli*. *Genome Res* **14**, 1036–1042.
374 de la Torre JR, Goebel BM, Friedmann EI, Pace NR (2003). Microbial diversity of cryptoendolithic
375 communities from the McMurdo Dry Valleys, Antarctica. *Appl Environ Microbiol* **69**,
376 3858–3867.
377 Dedieu A, Gaillard JC, Pourcher T, Darrouzet E, Armengaud J (2011). Revisiting iodination sites in
378 thyroglobulin with an organ-oriented shotgun strategy. *J Biol Chem* **286**, 259–269.

379 Eppard M, Krumbein WE, Koch C, Rhiel E, Staley JT, Stackebrandt E (1996). Morphological,
380 physiological, and molecular characterization of actinomycetes isolated from dry soil, rocks,
381 and monument surfaces. *Arch Microbiol* **166**, 12-22.
382 Essoussi I, Boujmil R, Nouioui I, Abbassi-Ghozzi I, Hamza A, Boudabous A *et al*. (2012). Genetic
383 diversity and esterase-profiling of Actinobacteria isolated from Sahara desert stones and
384 monuments. *Geomicrobiology Journal* **29**, 23–28.
385 Essoussi I, Ghodhbane-Gtari F, Amairi H, Sghaier H, Jaouani A, Brusetti L *et al*. (2010). Esterase
386 as an enzymatic signature of *Geodermatophilaceae* adaptability to Sahara desert stones and
387 monuments. *J Appl Microbiol* **108**, 1723–1732.
388 Fletcher M (1994). Bacterial biofilms and biofouling. *Curr Opin Biotechnol* **5**, 302–306.
389 Friedmann EI, Ocampo-Friedmann RO (1984). Endolithic microorganisms in extreme dry
390 environments: analysis of a lithobiontic habitat. In: Klug MJ and Reddy CA (eds.), Current
391 Perspectives in Microbial Ecology. Proc. Third Internat. Symp. Microb. Ecol., Am. Soc.
392 Microbiol., Washington, D.C., pp. 177–185.
393 Fukuchi S, Nishikawa K (2004). Estimation of the number of authentic orphan genes in bacterial
394 genomes. *DNA Res* **11**, 219–231, 311–313.
395 Ghosal D, Omelchenko MV, Gaidamakova EK, Matrosova VY, Vasilenko A, Venkateswaran A *et*
396 *al*. (2005). How radiation kills cells: survival of *Deinococcus radiodurans* and *Shewanella*
397 *oneidensis* under oxidative stress. *FEMS Microbiol Rev* **29**, 361–375.
398 Gogarten JP, Doolittle WF, Lawrence JG (2002). Prokaryotic evolution in light of gene transfer.
399 *Mol Biol Evol* **19**, 2226–2238.
400 Gtari M, Essoussi I, Maaoui R, Sghaier H, Boujmil R, Gury J *et al*. (2012). Contrasted resistance of
401 stone-dwelling *Geodermatophilaceae* species to stresses known to give rise to reactive
402 oxygen species. *FEMS Microbiol Ecol* **80**, 566–577.
403 Hartmann EM, Allain F, Gaillard JC, Pible O, Armengaud J (2014). Taking the shortcut for high-
404 throughput shotgun proteomic analysis of bacteria. *Methods Mol Biol* **1197**, 275–285.
405 Hartmann EM, Armengaud J (2014). Shotgun proteomics suggests involvement of additional
406 enzymes in dioxin degradation by *Sphingomonas wittichii* RW1. *Environ Microbiol* **16**,
407 162–176.
408 Horath T, Bachofen R (2009). Molecular characterization of an endolithic microbial community in
409 dolomite rock in the central Alps (Switzerland). *Microb Ecol* **58**, 290–306.
410 Ishiguro EE, Wolfe RS (1970). Control of morphogenesis in *Geodermatophilus*: ultrastructural
411 studies. *J Bacteriol* **104**, 566–580.
412 Ishiguro EE, Wolfe RS (1974). Induction of morphogenesis in *Geodermatophilus* by inorganic
413 cations and by organic nitrogenous cations. *J Bacteriol* **117**, 189–195.
414 Ito A, Taniuchi A, May T, Kawata K, Okabe S (2009). Increased antibiotic resistance of
415 *Escherichia coli* in mature biofilms. *Appl Environ Microbiol* **75**, 4093–4100.
416 Ivanova N, Sikorski J, Jando M, Munk C, Lapidus A, Glavina Del Rio T *et al*. (2010). Complete
417 genome sequence of *Geodermatophilus obscurus* type strain (G-20T). *Stand Genomic Sci* **2**,
418 158–167.
419 Jacob F (1977). Evolution and tinkering. *Science* **196**, 1161–1166.
420 Janssen PJ, Van Houdt R, Moors H, Monsieurs P, Morin N, Michaux A *et al*. (2010). The complete
421 genome sequence of *Cupriavidus metallidurans* strain CH34, a master survivalist in harsh
422 and anthropogenic environments. *PLoS One* **5**, e10433.
423 Jaouani A, Neifar M, Hamza A, Chaabouni S, Martinez MJ, Gtari M (2012). Purification and
424 characterization of a highly thermostable esterase from the actinobacterium
425 *Geodermatophilus obscurus* strain G20. *J Basic Microbiol* **52**, 653–660.
426 Kuhlman KR, Fusco WG, La Duc MT, Allenbach LB, Ball CL, Kuhlman GM *et al*. (2006).
427 Diversity of microorganisms within rock varnish in the Whipple Mountains, California.
428 *Appl Environ Microbiol* **72**, 1708-1715.

429 Langille MG, Brinkman FS (2009). IslandViewer: an integrated interface for computational
430 identification and visualization of genomic islands. *Bioinformatics* **25**, 664–665.
431 Laubichler MD (2007). Tinkering: a conceptual and historical evaluation. *Novartis Found Symp*
432 **284**, 20-29; discussion 29-34, 110–115.
433 Lee ZM, Bussema C, 3rd, Schmidt TM (2009). rrnDB: documenting the number of rRNA and
434 tRNA genes in bacteria and archaea. *Nucleic Acids Res* **37**, D489–493.
435 Liu H, Sadygov RG, Yates JR, 3rd (2004). A model for random sampling and estimation of relative
436 protein abundance in shotgun proteomics. *Anal Chem* **76**, 4193–4201.
437 Liu R, Li K, Zhang H, Zhu J, Joshi D (2014). Spatial distribution of microbial communities
438 associated with dune landform in the Gurbantunggut Desert, China. *J Microbiol* **52**, 898–
439 907.
440 Luedemann GM (1968). *Geodermatophilus*, a new genus of the Dermatophilaceae
441 (Actinomycetales). *J Bacteriol* **96**, 1848–1858.
442 Macedo MF, Miller AZ, Dionisio A, Saiz-Jimenez C (2009). Biodiversity of cyanobacteria and
443 green algae on monuments in the Mediterranean Basin: an overview. *Microbiology* **155**,
444 3476–3490.
445 Marzan LW, Shimizu K (2011). Metabolic regulation of *Escherichia coli* and its *phoB* and *phoR*
446 genes knockout mutants under phosphate and nitrogen limitations as well as at acidic
447 condition. *Microb Cell Fact* **10**, 39.
448 Mevs U, Stackebrandt E, Schumann P, Gallikowski CA, Hirsch P (2000). *Modestobacter*
449 *multiseptatus* gen. nov., sp. nov., a budding actinomycete from soils of the Asgard Range
450 (Transantarctic Mountains). *Int J Syst Evol Microbiol* **50 Pt 1**, 337–346.
451 Montero-Calasanz MC, Goker M, Broughton WJ, Cattaneo A, Favet J, Potter G *et al*. (2013).
452 *Geodermatophilus tzadiensis* sp. nov., a UV radiation-resistant bacterium isolated from sand
453 of the Saharan desert. *Syst Appl Microbiol* **36**, 177–182.
454 Montero-Calasanz MC, Goker M, Potter G, Rohde M, Sproer C, Schumann P *et al*. (2013).
455 *Geodermatophilus africanus* sp. nov., a halotolerant actinomycete isolated from Saharan
456 desert sand. *Antonie Van Leeuwenhoek* **104**, 207–216.
457 Montero-Calasanz MC, Goker M, Potter G, Rohde M, Sproer C, Schumann P *et al*. (2013).
458 *Geodermatophilus arenarius* sp. nov., a xerophilic actinomycete isolated from Saharan
459 desert sand in Chad. *Extremophiles* **16**, 903–909.
460 Montero-Calasanz MC, Goker M, Potter G, Rohde M, Sproer C, Schumann P *et al*. (2013).
461 *Geodermatophilus normandii* sp. nov., isolated from Saharan desert sand. *Int J Syst Evol*
462 *Microbiol* **63**, 3437–3443.
463 Montero-Calasanz MC, Goker M, Potter G, Rohde M, Sproer C, Schumann P *et al*. (2013).
464 *Geodermatophilus saharensis* sp. nov., isolated from sand of the Saharan desert in Chad.
465 *Arch Microbiol* **195**, 153–159.
466 Montero-Calasanz MC, Goker M, Potter G, Rohde M, Sproer C, Schumann P *et al*. (2013).
467 *Geodermatophilus telluris* sp. nov., an actinomycete isolated from Saharan desert sand. *Int J*
468 *Syst Evol Microbiol* **63**, 2254–2259.
469 Montero-Calasanz MC, Goker M, Rohde M, Schumann P, Potter G, Sproer C *et al*. (2013).
470 *Geodermatophilus siccatus* sp. nov., isolated from arid sand of the Saharan desert in Chad.
471 *Antonie Van Leeuwenhoek* **103**, 449–456.
472 Montero-Calasanz MC, Hofner B, Goker M, Rohde M, Sproer C, Hezbri K *et al*. (2014).
473 *Geodermatophilus poikilotrophi* sp. nov.: a multitolerant actinomycete isolated from
474 dolomitic marble. *Biomed Res Int* **2014**, 11.
475 Montero-Calasanz MD, Hezbri K, Goker M, Sghaier H, Rohde M, Sproer C *et al*. (2015).
476 Description of gamma radiation-resistant *Geodermatophilus dictyosporus* sp. nov. to
477 accommodate the not validly named *Geodermatophilus obscurus* subsp. *dictyosporus*
478 (Luedemann, 1968). *Extremophiles* **19**, 77–85.

479 Nijland R, Hall MJ, Burgess JG (2010). Dispersal of biofilms by secreted, matrix degrading,
480 bacterial DNase. *PLoS One* **5**, e15668.
481 Normand P (2006). *Geodermatophilaceae* fam. nov., a formal description. *Int J Syst Evol Microbiol*
482 **56**, 2277–2278.
483 Normand P, Daffonchio D, Gtari M. (2014). Family *Geodermatophilaceae*, p. 19. *In* Rosenberg E
484 (ed.), The prokaryotes – Actinobacteria. Springer-Verlag, Berlin, Heidelberg.
485 Normand P, Gury J, Pujic P, Chouaia B, Crotti E, Brusetti L *et al*. (2012). Genome sequence of
486 radiation-resistant *Modestobacter marinus* strain BC501, a representative actinobacterium
487 that thrives on calcareous stone surfaces. *J Bacteriol* **194**, 4773–4774.
488 Oda Y, Samanta SK, Rey FE, Wu L, Liu X, Yan T *et al*. (2005). Functional genomic analysis of
489 three nitrogenase isozymes in the photosynthetic bacterium *Rhodopseudomonas palustris*. *J*
490 *Bacteriol* **187**, 7784–7794.
491 Overmars L, Kerkhoven R, Siezen RJ, Francke C (2013). MGcV: the microbial genomic context
492 viewer for comparative genome analysis. *BMC Genomics* **14**, 209.
493 Paoletti AC, Parmely TJ, Tomomori-Sato C, Sato S, Zhu D, Conaway RC *et al*. (2006).
494 Quantitative proteomic analysis of distinct mammalian Mediator complexes using
495 normalized spectral abundance factors. *Proc Natl Acad Sci U S A* **103**, 18928–18933.
496 R Development Core Team (2007). R: a language and environment for statistical computing. *R*
497 *Foundation for Statistical Computing, Vienna, Austria,* http://www.R-project.org.
498 Rainey FA, Ray K, Ferreira M, Gatz BZ, Nobre MF, Bagaley D *et al*. (2005). Extensive diversity of
499 ionizing-radiation-resistant bacteria recovered from Sonoran Desert soil and description of
500 nine new species of the genus *Deinococcus* obtained from a single soil sample. *Appl*
501 *Environ Microbiol* **71**, 5225–5235.
502 Raybould CC, Unsworth MH, Gregory PJ (1977). Sources of sulphur in rain collected below a
503 wheat canopy. *Nature* **267**, 146–147.
504 Rousseau C, Gonnet M, Le Romancer M, Nicolas J (2009). CRISPI: a CRISPR interactive
505 database. *Bioinformatics* **25**, 3317–3318.
506 Santos CL, Tavares F, Thioulouse J, Normand P (2009). A phylogenomic analysis of bacterial
507 helix-turn-helix transcription factors. *FEMS Microbiol Rev* **33**, 411–429.
508 Sen A, Daubin V, Abrouk D, Gifford I, Berry AM, Normand P (2014). The phylogeny of
509 actinobacteria revisited in the light of complete genomes, the orders Frankiales and
510 Micrococcales should be split into coherent entities. Proposal of *Frankiales* ord. nov.,
511 *Geodermatophilales* ord. nov., *Acidothermales* ord. nov. and *Nakamurellales* ord. nov. *Int J*
512 *Syst Evol Microbiol* **64**, 3821–3832.
513 Singh A, Agrawal M (2008). Acid rain and its ecological consequences. *J Environ Biol* **29**, 15–24.
514 Singh AK, Singh R, Tomar D, Pandya CD (2012). The leucine aminopeptidase of *Staphylococcus*
515 *aureus* is secreted and contributes to biofilm formation. *Int J Infect Dis* **16**, e375–381.
516 Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: Molecular Evolutionary
517 Genetics Analysis version 6.0. *Mol Biol Evol* **30**, 2725–2729.
518 Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS *et al*. (2001). The
519 COG database: new developments in phylogenetic classification of proteins from complete
520 genomes. *Nucleic Acids Res* **29**, 22–28.
521 Tunnacliffe A, Lapinski J (2003). Resurrecting Van Leeuwenhoek's rotifers: a reappraisal of the
522 role of disaccharides in anhydrobiosis. *Philos Trans R Soc Lond B Biol Sci* **358**, 1755–1771.
523 Urzi C, Brusetti L, Salamone P, Sorlini C, Stackebrandt E, Daffonchio D (2001). Biodiversity of
524 *Geodermatophilaceae* isolated from altered stones and monuments in the Mediterranean
525 basin. *Environ Microbiol* **3**, 471–479.
526 Urzi C, Salamone P, Schumann P, Rohde M, Stackebrandt E (2004). *Blastococcus saxobsidens* sp.
527 nov., and emended descriptions of the genus *Blastococcus* Ahrens and Moll 1970 and
528 *Blastococcus aggregatus* Ahrens and Moll 1970. *Int J Syst Evol Microbiol* **54**, 253–259.

529 Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S *et al*. (2006). MaGe: a microbial
530 genome annotation system supported by synteny results. *Nucleic Acids Res* **34**, 53–65.
531 Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D *et al*. (2014).
532 ProteomeXchange provides globally coordinated proteomics data submission and
533 dissemination. *Nat Biotechnol* **32**, 223–226.
534 Walker JJ, Pace NR (2007). Phylogenetic composition of Rocky Mountain endolithic microbial
535 ecosystems. *Appl Environ Microbiol* **73**, 3497–3504.
536 White-Ziegler CA, Um S, Perez NM, Berns AL, Malhowski AJ, Young S (2008). Low temperature
537 (23 degrees C) increases expression of biofilm-, cold-shock- and RpoS-dependent genes in
538 *Escherichia coli* K-12. *Microbiology* **154**, 148–166.
539 Williams E, Lowe TM, Savas J, DiRuggiero J (2007). Microarray analysis of the hyperthermophilic
540 archaeon *Pyrococcus furiosus* exposed to gamma irradiation. *Extremophiles* **11**, 19–29.
541 Wu M, Ren Q, Durkin AS, Daugherty SC, Brinkac LM, Dodson RJ *et al*. (2005). Life in hot carbon
542 monoxide: the complete genome sequence of Carboxydothermus hydrogenoformans Z-
543 2901. *PLoS Genet* **1**, e65.
544 Zhou J, Rudd KE (2013). EcoGene 3.0. *Nucleic Acids Res* **41**, D613–624.
545 Zivanovic Y, Armengaud J, Lagorce A, Leplat C, Guerin P, Dutertre M *et al*. (2009). Genome
546 analysis and genome-wide proteomics of *Thermococcus gammatolerans*, the most
547 radioresistant organism known amongst the Archaea. *Genome Biol* **10**, R70.
548

549 **Titles and legends to figures**

550

551 **Figure 1.** Circular representation of the three genomes with detected proteins. From the outside in
552 are 1-the coordinates, 2-the G+C% (ranging from 70 to 80%), 3-the HGT predicted by the software
553 RGP run on Mage (in grey), 4-the core genome or the genes shared by the three genomes (threshold
554 of 50% identity over 80% of the length of the shorter sequence, present in a synton), 5-the genes
555 specific to each genome (absent from the other two genomes, minus the "genes of unknown
556 function" and 6-the proteins detected in this study (in red).

557

558 **Figure 2.** Schematic view of the Geodermatophilaceae physiological determinants. Transport
559 systems, storage components, the motility machinery and the main energy generating pathways are
560 represented.
561

562 **Supplementary information files**

563 The following additional data are available with the online version of this paper.

564

565 **Supplementary Figure 1.** Venn diagram showing the core genome and shared genes between the
566 three Geodermatophilaceae, *Blastococcus saxobsidens* DD2 (Bs), *Modestobacter marinus* BC501
567 (Mm) and *Geodermatophilus obscurus* DSM 43160 (Go). At a threshold level of 30% over 80% of
568 the length of the shortest protein sequence, the three taxa have a common genome of 3277 CDS,
569 while the extended genome would comprise 7739 CDS. The homology constraints are 30% over
570 80% of the length of the shortest protein sequence. Cap is capsule synthesis protein, Cox is carbon
571 monoxide dehydrogenase, Crt is 9-cis-epoxycarotenoid dioxygenase (1st step of abscisic-acid
572 biosynthesis from carotenoids), Fli is flagellum synthesis, Hup is hydrogenase uptake, Shc is
573 squalene hopene cyclase, SodC is copper/zinc superoxide dismutase, SOS is "Save our soul"
574 response, Uid is beta-glucuronidase and Vao is vanillyl-alcohol oxidase (Format: JPEG).

575

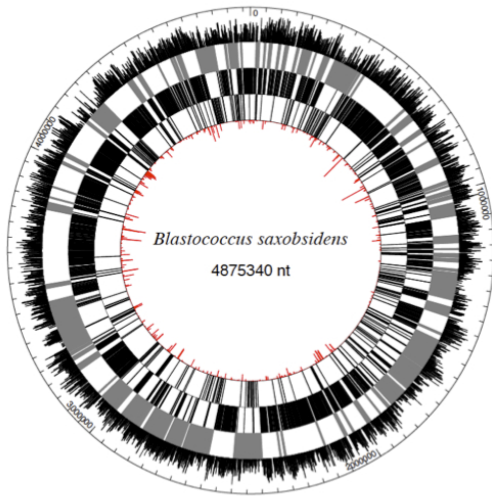
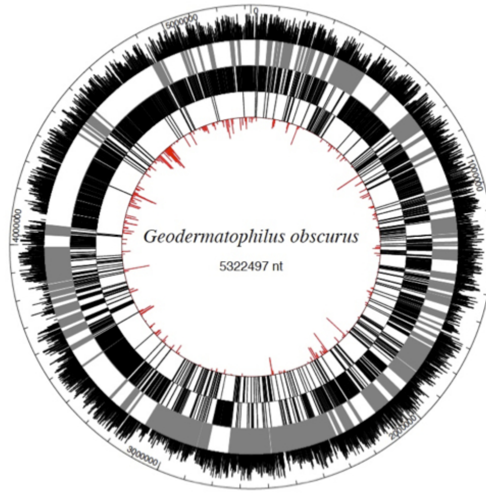
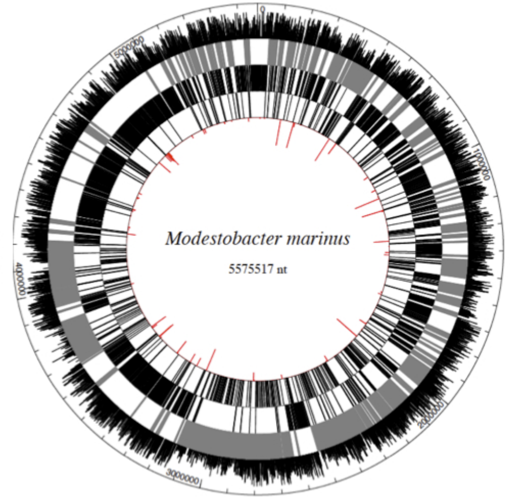
576 **Supplementary Figure 2.** Summary of the most abundant proteins of *Blastococcus saxobsidens*
577 DD2 (Bs), *Modestobacter marinus* BC501 (Mm) and *Geodermatophilus obscurus* DSM 43160
578 (Go), detected under our standard experimental conditions, and ordered for each bacterium based on
579 their Normalized Spectral count Abundance Factors (NSAF) percentages. Red-to-green (low-to-
580 high) lines were drawn under Open Reading Frame (ORF) boxes based on the input scores for each
581 expressed protein. Subcellular localizations of expressed and neighboring proteins are indicated by
582 the colors of ORF boxes—cytoplasmic, green; cytoplasmic membrane, yellow; outer membrane,
583 dark blue; extracellular, light blue; cell wall, orange; periplasmic, magenta and unknown, grey
584 (Format: PDF).

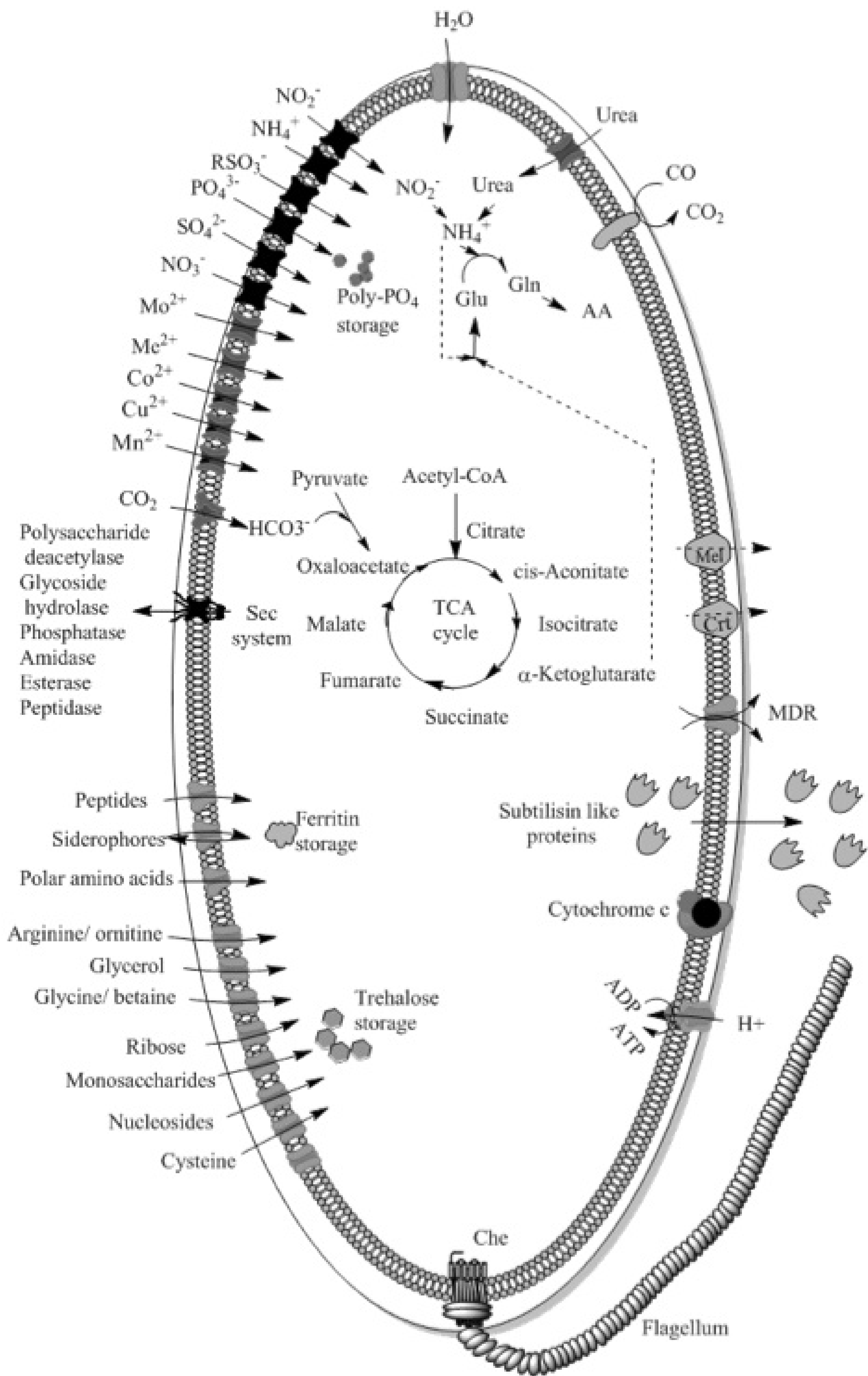
585

586 **Supplementary Figure 3.** Correspondence analyses for the three Geodermatophilaceae genomes
587 (Bs, *Blastococcus saxobsidens* DD2; Mm, *Modestobacter marinus* BC501 and
588 Go, *Geodermatophilus obscurus* DSM 43160).
589 **(A)** Correspondence analysis made with COG categories (red dots) extracted from the Mage
590 platform for the three Geodermatophilaceae genomes (Bs, Mm and Go) (black diamonds) and
591 related actinobacteria (blue asterisks: Ac, *Acidothermus cellulolyticus*; Bl, *Bifidobacterium longum*;
592 Fa, *Frankia alni*; Fc, *Frankia* CcI3; Fe, *Frankia* EaN1pec; Ms, *Mycobacterium smegmatis* and Mt,
593 *Mycobacterium tuberculosis*). The first dimension (Dim1) explains 43.18% of the variance while
594 the second dimension (Dim2) explains 25.12% (Format: JPEG).
595 **(B)** Correspondence analysis made with regulators numbers extracted from the Mage platform for
596 the three Geodermatophilaceae genomes—Bs, Mm and Go—(black diamonds) and related
597 actinobacteria (blue asterisks: Ac; Bl; Fa; Fc; Fe; Ms; Mt; Sc, *Streptomyces coelicolor* and Tw,
598 *Tropheryma whipplei*). Bl and Tw have been removed from this analysis as too far away from the
599 rest) (Format: JPEG). The first dimension (Dim1) explains 12.63% of the variance while the second
600 dimension (Dim2) explains 21.83%.
601 **(C)** Correspondence analysis made with signaling molecules numbers extracted from the Mage
602 platform for the three Geodermatophilaceae genomes—Bs, Mm and Go—(black diamonds) and
603 related actinobacteria (blue asterisks: Ac, Fa, Fc, Fe, Ms, Mt and Sc) (Format: JPEG). The first
604 dimension (Dim1) explains 42.1% of the variance while the second dimension (Dim2) explains
605 28.8%.
606
607 **Supplementary Figure 4.** Absorbance curves of whole cell extracts of the three
608 Geodermatophilaceae genomes (*Blastococcus saxobsidens* DD2 (Bs), red; *Modestobacter marinus*
609 BC501 (Mm), cyan; *Geodermatophilus obscurus* DSM 43160 (Go), blue; control, black). The Bs

610 orange pigment absorbs at 230–270 nm and at 450–500 nm; whereas pigments of Mm and Go
611 absorb almost continuously between 200 and 750 nm (Format: JPEG).
612
613 **Supplementary Figure 5.** Phylogenetic tree of the Ku-proteins coding gene sequences in
614 *Blastococcus saxobsidens* DD2, *Modestobacter marinus* BC501, *Geodermatophilus obscurus* DSM
615 43160 and other Actinobacteria. The neighbor-joining tree was inferred from DNA sequences using
616 MEGA6 then integrated with the genomic context at the MGcV (see Methods). Dark green ORF
617 boxes indicate the species G+C percentages. Less darkened colors indicate a decrease in G+C
618 content. The scale bar corresponds to estimated sequence divergence and attached numbers
619 represent the rate of substitutions per nucleotide site (Format: JPEG).
620
621 **Supplementary Table 1.** General characteristics of *Blastococcus saxobsidens* DD2 (Bs),
622 *Modestobacter marinus* BC501 (Mm) and *Geodermatophilus obscurus* DSM 43160 (Go) genomes
623 (Format: DOC).
624
625 **Supplementary Table 2.** Selected important enzymes and detected ORFans and "proteins of
626 unknown function" in the proteogenomes of *Blastococcus saxobsidens* DD2 (Bs) *Modestobacter*
627 *marinus* BC501 (Mm) and *Geodermatophilus obscurus* DSM 43160 (Go).
628
629 **Supplementary Table 3.** Clusters of Orthologous Groups (COG) classification of *Blastococcus*
630 *saxobsidens* DD2 (Bs), *Modestobacter marinus* BC501 (Mm) and *Geodermatophilus obscurus*
631 DSM 43160 (Go) genomes (Format: DOC).
632

633 **Supplementary Table 4.** Number of lost, horizontally transferred, duplicated or created genes
634 (ORFans) and genomic islands in *Blastococcus saxobsidens* DD2 (Bs), *Modestobacter marinus*
635 BC501 (Mm) and *Geodermatophilus obscurus* DSM 43160 (Go) (Format: DOC).
636
637 **Supplementary Table 5.** List of genes putatively involved in pigments synthesis in *Blastococcus*
638 *saxobsidens* DD2 (Bs), *Modestobacter marinus* BC501 (Mm) and *Geodermatophilus obscurus*
639 DSM 43160 (Go) (Format: DOC).
640
641 **Supplementary Table 6.** List of stress-related and DNA repair genes in *Blastococcus saxobsidens*
642 DD2 (Bs), *Modestobacter marinus* BC501 (Mm) and *Geodermatophilus obscurus* DSM 43160
643 (Go) (Format: DOC).
644
645 **Supplementary Table 7.** Selection of pathways and enzymes of physiological pertinence in
646 *Blastococcus saxobsidens* DD2 (Bs), *Modestobacter marinus* BC501 (Mm) and *Geodermatophilus*
647 *obscurus* DSM 43160 (Go) (Format: DOC).
648
649 **Additional Data.** The following additional data are available at http://www.ebi.ac.uk/pride.
650 **(1)** Additional data 1 is related to *Blastococcus saxobsidens* DD2 proteomics project (accession:
651 PXD001519, DOI: 10.6019/PXD001519).
652 **(2)** Additional data 2 is related to *Modestobacter marinus* proteomics project (accession:
653 PXD001518, DOI: 10.6019/PXD001518).
654 **(3)** Additional data 3 is related to *Geodermatophilus obscurus* DSM43160 proteomics project
655 (accession: PXD001520, DOI: 10.6019/PXD001520).


Blastococcus saxobsidens
4875340 nt

Geodermatophilus obscurus
5322497 nt

Modestobacter marinus
5575517 nt



H_2O
NO_2^-
NH_4^+
RSO_3^-
PO_4^{3-}
SO_4^{2-}
NO_3^-
Mo^{2+}
Me^{2+}
Co^{2+}
Cu^{2+}
Mn^{2+}
CO_2
Urea
CO
CO_2
NO_2^-
Urea
NH_4^+
Gln
Glu
AA
Poly-PO_4
storage
Pyruvate
Acetyl-CoA
$HCO3^-$
Citrate
Oxaloacetate
cis-Aconitate
TCA
cycle
Malate
Isocitrate
Fumarate
α-Ketoglutarate
Succinate
Polysaccharide
deacetylase
Glycoside
hydrolase
Phosphatase
Amidase
Esterase
Peptidase
Sec
system
Mel
Crt
MDR
Peptides
Siderophores
Ferritin
storage
Polar amino acids
Subtilisin like
proteins
Cytochrome c
Arginine/ ornitine
Glycerol
Glycine/ betaine
Trehalose
storage
Ribose
Monosaccharides
Nucleosides
Cysteine
ADP
ATP
H+
Che
Flagellum