Robustness analysis of metabolic predictions in algal microbial communities based on different annotation pipelines

Elham Karimi [Corresp., 1] , **Enora Geslain** [1, 2] , **Arnaud Belcour** [3] , **Clémence Frioux** [4] , **Méziane Aïte** [3] , **Anne Siegel** [3] , **Erwan Corre** [2] , **Simon M. Dittami** [Corresp. 1]

[1] UMR8227, Integrative Biology of Marine Models, Sorbonne Université/CNRS, Station Biologique de Roscoff,, Roscoff, France

[2] FR2424, Sorbonne Université/CNRS, Station Biologique de Roscoff, Roscoff, France

[3] Equipe Dyliss, Univ Rennes, Inria, CNRS, IRISA, Rennes, France

[4] Inria, Bordeaux Sud-Ouest Université, Talence, France

Corresponding Authors: Elham Karimi, Simon M. Dittami
Email address: ekarimi@sb-roscoff.fr, simon.dittami@sb-roscoff.fr

Animals, plants, and algae rely on symbiotic microorganisms for their development and functioning. Genome sequencing and genomic analyses of these microorganisms provide opportunities to construct metabolic networks and to analyze the metabolism of the symbiotic communities they constitute. Genome-scale metabolic network reconstructions rest on information gained from genome annotation. As there are multiple annotation pipelines available, the question arises to what extent differences in annotation pipelines impact outcomes of these analyses. Here, we compare five commonly used pipelines (Prokka, MaGe, IMG, DFAST, RAST) from predicted annotation features (coding sequences, Enzyme Commission numbers, hypothetical proteins) to the metabolic network-based analysis of symbiotic communities (biochemical reactions, producible compounds, and selection of minimal complementary bacterial communities). While Prokka and IMG produced the most extensive networks, RAST and DFAST networks produced the fewest false positives and the most connected networks with the fewest dead-end metabolites. Our results underline differences between the outputs of the tested pipelines at all examined levels, with small differences in the draft metabolic networks resulting in the selection of different microbial consortia to expand the metabolic capabilities of the algal host. However, the consortia generated yielded similar predicted producible compounds and could therefore be considered functionally interchangeable. This contrast between selected communities and community functions depending on the annotation pipeline needs to be taken into consideration when interpreting the results of metabolic complementarity analyses. In the future, experimental validation of bioinformatic predictions will likely be crucial to both evaluate and refine the pipelines and needs to be coupled with increased efforts to expand and improve annotations in reference databases.

1 **Robustness analysis of metabolic predictions in algal microbial communities**
2 **based on different annotation pipelines**

4 **Elham Karimi[1*], Enora Geslain[1,2], Arnaud Belcour[3], Clémence Frioux[4], Méziane Aïte[3],**
5 **Anne Siegel[3], Erwan Corre[2], Simon M. Dittami[1] ***
6
7 [1] Sorbonne Université/CNRS, Station Biologique de Roscoff, UMR8227, Integrative Biology of
8 Marine Models, CS 90074, Roscoff, France.
9 [2] Sorbonne Université/CNRS, Station Biologique de Roscoff, FR2424, CS 90074, Roscoff,
10 France Roscoff, France.
11 [3]Univ Rennes, Inria, CNRS, IRISA, Equipe Dyliss, Rennes, France.
12 [4] Inria, Bordeaux Sud-Ouest Université, 33405 Talence, France.
13
14 Corresponding Authors:
15 Elham Karimi[1] & Simon M. Dittami[1]
16 Place Georges Teissier, 29680 Roscoff, France
17 Email address: elham.karimi@sb-roscoff.fr. & simon.dittami@sb-roscoff.fr.

19 ## Abstract

20 Animals, plants, and algae rely on symbiotic microorganisms for their development and
21 functioning. Genome sequencing and genomic analyses of these microorganisms provide
22 opportunities to construct metabolic networks and to analyze the metabolism of the symbiotic
23 communities they constitute. Genome-scale metabolic network reconstructions rest on
24 information gained from genome annotation. As there are multiple annotation pipelines
25 available, the question arises to what extent differences in annotation pipelines impact outcomes
26 of these analyses. Here, we compare five commonly used pipelines (Prokka, MaGe, IMG,
27 DFAST, RAST) from predicted annotation features (coding sequences, Enzyme Commission
28 numbers, hypothetical proteins) to the metabolic network-based analysis of symbiotic
29 communities (biochemical reactions, producible compounds, and selection of minimal
30 complementary bacterial communities). While Prokka and IMG produced the most extensive
31 networks, RAST and DFAST networks produced the fewest false positives and the most
32 connected networks with the fewest dead-end metabolites. Our results underline differences
33 between the outputs of the tested pipelines at all examined levels, with small differences in the
34 draft metabolic networks resulting in the selection of different microbial consortia to expand the
35 metabolic capabilities of the algal host. However, the consortia generated yielded similar
36 predicted producible compounds and could therefore be considered functionally interchangeable.
37 This contrast between selected communities and community functions depending on the
38 annotation pipeline needs to be taken into consideration when interpreting the results of
39 metabolic complementarity analyses. In the future, experimental validation of bioinformatic
40 predictions will likely be crucial to both evaluate and refine the pipelines and needs to be
41 coupled with increased efforts to expand and improve annotations in reference databases.
42 **Keywords:** Gene prediction; functional annotation; genome-scale metabolic networks;
43 metabolic complementary analyses; metabolic exchanges; holobionts
44

Introduction

Plants, animals, and algae are hosts to a large diversity of microorganisms. The importance of these symbiotic microbes for the development and functioning of their hosts is widely accepted (Amin et al. 2015; Fraune & Bosch 2010; McFall-Ngai et al. 2013; Philippot et al. 2013). This is also true for brown algal surfaces, which provide an attractive substrate for different bacterial phyla, most importantly the *Proteobacteria*, *Bacteroidetes*, *Firmicutes*, and *Actinobacteria*. These bacteria are important for the brown algal host e.g. by promoting its growth and by providing defenses against pathogens (KleinJan et al. 2017; Singh & Reddy 2014). The mutualistic relationships between algae and their microbiota are so tightly interwoven that both parts together are frequently considered a functional entity, the so-called holobiont (Egan et al. 2013).

Recent advances in sequencing technology have generated new opportunities for genome sequencing and genomic analyses of microbial populations (Dittami et al. 2021). A key step in interpreting this wealth of genomic data is their structural and functional annotation, and multiple prokaryotic annotation pipelines have been developed to accomplish this task (Siezen & van Hijum 2010). Based on functional annotations, genome-scale metabolic networks can then be generated, providing a formal representation of an organisms' metabolism (Haggart et al. 2011). Finally, metabolic complementarity can highlight potentially beneficial metabolic interactions between a host and its symbionts (Dittami et al. 2014; Muller et al. 2018), and the degree of metabolic complementarity between two organisms has been suggested to be a direct indicator of beneficial vs. competitive interaction in a system (Burgunter-Delamare et al. 2020; Levy et al. 2015). It can even be used as a criterion for the selection of minimal consortia where the host attains maximal metabolic potential while minimizing exchanges (Frioux et al. 2018; Levy et al. 2015). Although these tools and pipelines provide a promising framework for a comprehensive understanding of the metabolism of organisms, it is currently unknown to what extent outcomes of such analyses are sensitive to gene-associations present in the metabolic networks they rely on, and thus the precision of the prokaryotic annotation pipelines used. For metabarcoding analyses, for instance, the choice of the bioinformatic pipeline has been shown to have a significant impact on some of the biological conclusions that could be drawn (Siegwald et al. 2019).

Here we compared a selection of widespread genome annotation pipelines, examined their impact on the reconstructed genome-scale metabolic networks, and assessed whether the same conclusion about metabolic complementarity could be reached with the different pipelines. Our work was carried out using the filamentous brown alga *Ectocarpus subulatus* as a model system. Different strains of *Ectocarpus* have been established as a genomic and genetic model (Cock et al. 2010) and an increasing amount of information is available on its microbiome (Burgunter-Delamare et al. 2020; Dittami et al. 2015; Tapia et al. 2016), including a collection of 72 genomes (Karimi et al. 2019) of bacteria isolated from *Ectocarpus* as a host (Dittami et al. 2020b).

Materials & Methods

Genome Annotation

90
91 Eighty-one publicly available genomes were used for this study (Table S1), all of them
92 corresponding to strains that have been isolated from the brown alga *Ectocarpus subulatus*. The
93 genomes were structurally and functionally annotated using five popular annotation pipelines:
94 the Magnifying Genomes (MaGe) platform (Vallenet et al. 2019), the Rapid Annotations using
95 Subsystems Technology (RAST) toolkit (Aziz et al. 2008), the Integrated Microbial Genomes
96 (IMG) server (Markowitz et al. 2009), the Prokka tool v1.13 (Seemann 2014) and the DDBJ Fast
97 Annotation and Submission Tool (DFAST) (Tanizawa et al. 2017). Each pipeline uses one or
98 more unique databases to predict and search for genes and gene functions. MaGe searches for
99 functional features using UniProtKB/Swiss-Prot, Interpro, FIGFAM, COG, ENZYME, and
100 Diamond as a search tool (Vallenet et al. 2019); RAST predicts gene functions using the SEED
101 database (Aziz et al. 2008); IMG predicts features for genes based on COGs, Pfams, TIGRfams,
102 as well as the KEGG and MetaCyc; Prokka is a command-line tool using UniProt, Pfam, and
103 TIGRFAMs; DFAST uses ortholog searches with reciprocal BLAST searches, HMM searches
104 against TIGRFAM, and COG assignments for functional annotations (Tanizawa et al. 2017). All
105 of the pipelines were run with default parameters (Table S2). For Prokka, we also performed
106 tests with more stringent e-value settings (1e-15 vs 1e-6(default)); as these results were similar to
107 those obtained with default parameters they are included only in the supplementary tables.
108 supplementary tables.
109
110 *Draft metabolic network reconstruction*
111
112 The draft metabolic network of the host alga *Ectocarpus subulatus* was already available
113 (Dittami et al. 2020a) and had been reconstructed using Pathway Tools version 23 (Karp et al.
114 2019) and the AuReMe pipeline (Aite et al. 2018). To be able to use this metabolic network of
115 the host for complementarity analyses, the same version of Pathway Tools was also used to
116 reconstruct the bacterial draft metabolic networks. For each annotated bacterial genome (i.e. 5
117 pipelines x 81 strains), a draft metabolic network was automatically reconstructed using Mpwt,
118 the PathoLogic multiprocessing wrapper Python package (Belcour et al. 2020) taking into
119 account the protein description, GO terms, and complete EC numbers. The output files created
120 by Pathway Tools were then converted to the SBML format with the Padmet software. During
121 this process the reactions without gene association were removed. Then, the list of reactions as
122 well as network statistics were extracted using the report_network.py script from the Padmet
123 software (Aite et al. 2018). Consequently, for each of the five annotation pipelines, a set of 81
124 draft metabolic networks was obtained and used for metabolic complementarity analyses.
125 Furthermore, one additional set was produced for each strain by merging the metabolic networks
126 obtained for each annotation pipeline.
127
128 *Predicting Metabolic Complementarity*
129
130 The SBML files were utilized to predict potential metabolic complementarity between the algal
131 host and the bacteria using the MiSCoTo tool (Frioux et al. 2018). Based on a list of compounds
132 present in the culture medium, i.e. Provasoli-enriched seawater as defined by Prigent et al.
133 (2014), MiSCoTo predicts the production of compounds (i.e. the scope) in combined metabolic
134 networks based on a discrete abstraction, which assumes recursively that a compound is
135 producible either if it belongs to the culture medium if it is the product of a reaction in the

136 combined metabolic networks whose substrates are themselves producible, as stated in the
137 network expansion algorithm (Ebenhöh et al. 2004). It thus does not consider the kinetics, or
138 stoichiometry of the reactions, nor the relative abundance of the organism corresponding to the
139 metabolic networks. MisCoTo furthermore allows for the selection of minimal symbiont
140 communities that enable the production of a given set of compounds. In our analyses of the data
141 obtained with the five different pipelines individually plus one analysis merging all genomes
142 from all pipelines, the set of host metabolites that became producible through symbiotic
143 cooperation was computed with miscoto_scopes. This set was then used as a target for
144 community selection with miscoto_mincom using "--soup" as an option, *i.e.* ignoring the
145 putative cost of exchanges associated with ensuring the producibility of these novel metabolites
146 by the host. In each experiment, the minimal community size ensuring the producibility of these
147 algal compounds was computed, together with the set of symbionts that appears in at least one of
148 the minimal communities ("--union" option). Both the list of producible metabolites and the
149 selected microbial communities were compared for all tested annotation pipelines. The detailed
150 description of the code and data used for the above steps is available on Github
151 https://github.com/ElhamKarimi/Metabolic-predictions_different-Pipelines.
152
153 *Indirect measures of network quality and manual examination of pipeline-specific reactions*
154 *reconstructed metabolic networks*
155
156 To gain insight into the quality of metabolic networks generated based on the different pipelines,
157 dead-end metabolites were computed for each pipeline using Memote (Lieven et al. 2020).
158 Furthermore, sizes of the largest strongly connected components were counted in each of the
159 networks to obtain a measure of its overall connectivity. First, directed metabolite graphs were
160 constructed from the SBML files using the Networkx package (Hagberg et al. 2008). Then, 23
161 ubiquitous metabolites were removed (water, proton, ATP, ADP, 3-5-ADP, AMP, NADP,
162 NADPH, NAD, NADH, FAD, UDP, GTP, GDP, oxygen molecule, inorganic phosphate,,
163 diphosphate, carbon dioxide, ammonia, hydrogen peroxide, coenzyme A, H_2 acceptor, H_2 donor-
164 H2). Finally, for each graph, strongly connected components were found with the Networkx
165 strongly_connected_components function. Then, for each metabolic network, the ratio between
166 the size of the largest strongly connected component and the total number of metabolites was
167 computed.
168 In addition to these indirect measures, we randomly chose 100 reactions that were exclusively
169 found in one pipeline (20 per pipeline) for manual curation. Protein sequences of the bacterial
170 genes that led to the prediction of these reactions were aligned with the Swissprot/Uniprot
171 database using blastP and were then classified as high confidence (>30% ID; > 70% coverage
172 with a protein with the same metabolic function, functional domains present), low confidence
173 (lower similarity with a protein with the same function, or similarity with a protein with a
174 distinct but similar function), or false (>30% ID; >70% coverage with proteins known to carry
175 out a different function).
176
177
178 *Statistical Analyses*
179
180 All statistical analyses were carried out using RStudio (v.1.1.463). Differences in the number of
181 annotation features per pipeline were determined using an ANOVA followed by a Tukey's HSD
182 (honestly significant difference) test, considering each genome as a replicate. Nonmetric

183   multidimensional scaling (NMDS) was performed on a presence/absence matrix comprising all
184   reactions for each genome/annotation pipeline. The Bray-Curtis index was used as a measure of
185   quantification of compositional dissimilarity between genomes annotated by different pipelines.
186   To verify if the annotation pipeline had an impact on the reaction, an ANOSIM test (number of
187   permutations = 9999) was performed using the Vegan package (Oksanen et al. 2020). UpSet
188   diagrams were generated using the UpSetR package (Conway et al. 2017). Comparisons of the
189   overall numbers of predicted EC-numbers per category were tested using a Binomial test and a
190   Bonferroni correction for multiple testing.

191

Results

193   81 bacterial genomes were annotated using the Prokka, RAST, IMG, MaGe, and DFAST
194   pipelines, and the number of predicted features (Coding Sequences (CDSs), Enzyme
195   Commission (EC) numbers, hypothetical proteins) are shown for each genome and pipeline in
196   Table S1. The number of predicted CDSs across all genomes and pipelines ranged from 2180 to
197   6755; the number of predicted EC-numbers ranged from 443 to 2106. The number of
198   hypothetical proteins ranged from 56 to3566. Finally, the average number of predicted metabolic
199   reactions, after automatic metabolic network reconstruction with Pathway Tools, ranged from
200   1034 to 2052. Figure 1 shows the impact of the tested annotation pipelines on these metrics. At
201   the level of CDS prediction, we observed divergence with MaGe, for instance, predicting more
202   CDSs per genome than all other pipelines. Also, MaGe annotation contained fewer unknown
203   hypothetical proteins.

204

205   Prokka predicted more EC numbers than the other pipelines, regardless of the e-value cutoff
206   used. The precision of the predictions was similar in Prokka, IMG, MaGe, and DFAST with 85,
207   88, 84, and 85% of all EC numbers being complete. The only outlier was RAST, which provided
208   only complete EC numbers (Table S3). We did, however, observe significant differences
209   between the pipelines regarding certain categories of EC numbers. For instance, DFAST
210   predicted most oxidoreductases acting on NADH/NADPH, IMG predicted most Ligases forming
211   phosphoric ester bonds, MAGE was the only pipeline to predict translocases (E.C. 7.-.-.-),
212   PROKKA predicted most hydrolases acting on phosphorus-nitrogen bonds, and RAST predicted
213   most hydrolases acting on ether bonds. A detailed list of overrepresented EC numbers per
214   annotation pipeline is provided in Table 1.

215

216   In terms of predicted metabolic reactions, IMG predicted the highest numbers (4.2% more than
217   average) closely followed by Prokka with default settings (3.1%), while DFAST annotations
218   resulted in the fewest reactions (7.7% fewer than average). Similar patterns were also obtained
219   when running the analyses separately for the different bacterial phyla (Fig. S1), except that EC-
220   number predictions for *Actinobacteria* by DFAST were lower compared to other pipelines, and
221   CDS prediction for *Firmicutes* by IMG were slightly higher. Finally, we examined three indirect
222   measures of quality for the network reconstructions: the percentage of dead-ends, i.e. metabolites
223   which can be produced but are not consumed; orphan metabolites, i.e. metabolites that are
224   consumed but not produced in the metabolic networks; and finally, the ratio between the size of
225   the largest strongly connected components and the total number of metabolites, a direct estimate
226   of the "connectedness" of the network. The percentage of dead-end metabolites was similar
227   across all pipelines: 30.35% for Prokka, 29.66% for IMG, 29.17 for DFAST, 29.02% for MaGe,
228   and 28.4% for RAST. The percentage of orphan metabolites was also similar: 30.63% for

229   Prokka, 30.49% for IMG, 29.81% for DFAST, 30.37% for MaGe, and 29.48% for RAST. Lastly,
230   this was also true for the ratio between the size of the largest of strongly connected components
231   and the total number of metabolites: 26% for Prokka, 28% for IMG, 29% for DFAST, 27% for
232   MaGe, and 29% for RAST.
233   **To determine how similar networks based on different annotation pipelines were,** the Bray-
234   Curtis dissimilarity, a unidimensional distance measure between two matrices or networks, was
235   calculated for all network comparisons. Non-Metric Multidimensional Scaling (NMDS) was then
236   used to display the resulting distance matrix in the two-dimensional space. In the resulting graph
237   (Fig. 2 and Table S4) the distance between two points represents the dissimilarity of the
238   underlying networks. It shows that the similarity between networks is determined by the bacterial
239   phyla, and less so by the annotation pipeline, but within each phylum, there was neither a clear
240   clustering according to strain nor according to annotation pipeline. ANOSIM analyses were then
241   used to determine if these patterns were statistically significant. They show that both phylum and
242   annotation pipeline had a significant effect on the similarity of the networks ($p<0.0001$ in both
243   cases), but confirm that the effect was stronger for phylum (R=0.93) than for annotation pipeline
244   (R=0.41).
245   To further explore the **differences in the metabolic network reconstructions** and to highlight
246   the specificities of each annotation pipeline we aimed to determine the conserved and unique
247   content between the genome annotations from these different pipelines. This was done using the
248   UpSetR package. Besides generating lists of reactions specific to different subsets of pipelines,
249   this package also generates an UpSet diagram (Fig. 3), which displays the size of these sets. For
250   instance, 2535 reactions were predicted in at least one genome with each of the 5 pipelines.
251   Prokka and MaGe annotation resulted in the most pipeline-specific reactions: 390 (325 with
252   more stringent e-value) and 256 respectively. These two tools also shared a high number of
253   pairwise exclusive metabolic reactions, i.e. reactions not found by any other pipeline (307 vs 15-
254   65 for all other pairwise comparisons). The lists of reactions exclusively found in each pipeline
255   as well as lists of the metabolic pathways they belong to are provided in Tables S5 and S6.
256
257   **To get an overview of the dominant metabolic functions specific to the networks** based on
258   the different pipelines, we examined pipeline-specific metabolic reactions at the level of the
259   pathways they participate in. In total, 180 pathways comprised the reactions specific to Prokka,
260   and among them, 28 were constituted of >50% Prokka-specific reactions, including nitrobenzene
261   degradation (PWY-5637) I, taurine degradation I (PWY-1263), albaflavenone biosynthesis
262   (PWY-5887), and novobiocin biosynthesis (PWY-7287). Similarly, for MaGe the specific
263   reactions could be associated with 127 pathways, and 11 were constituted of >50% specific
264   reactions, including nitroethane degradation (PWY-5355), sulfite oxidation I (PWY-527), and
265   sulfite oxidation IV (PWY-5326). For RAST 99 pathways including 6 pathways >50% complete
266   presented the specific reactions, e.g. acetaldehyde biosynthesis II (PWY-6330) and
267   trypanothione biosynthesis (TRYPANOSYN-PWY). IMG-specific reactions represented 52
268   pathways with only 4 >50% complete, including tRNA splicing II (PWY-7803) and aldoxime
269   degradation (P345-PWY). Finally, reactions predicted specifically by DFAST corresponded to
270   42 pathways with only Nitrogen fixation II (PWY-7576) and Fatty acid biosynthesis (PWY-
271   5970) being >50% complete (Table S6).
272
273   **To evaluate the accuracy of the annotation pipelines, manual curation of 20 randomly**
274   **selected reactions that were unique to each pipeline** was carried out. DFAST has the highest

275   number of high confidence assignations (10) followed by Prokka and MAGE (6), IMG(4), and
276   RAST(1). RAST, on the other hand, did not produce any annotations that could be clearly
277   identified as false, followed by Prokka (1), DFAST (2), IMG (3), and MaGe (3). In Prokka,
278   increasing the stringency of the e-value cutoff did not affect the proportions of high confidence,
279   low confidence or false reactions. In DFAST, although the correct reaction was associated with
280   the gene, we observed discrepancies between the product name of genes and the EC number
281   associated in 3 cases. In MAGE and IMG we found 1 and 3 predicted proteins without any
282   similarity in the database and classified them as low confidence (Table S7).
283
284   To assess how these differences in the draft metabolic networks impacted the function of the
285   predicted metabolism of algal-bacterial holobionts, we next examined the list of **metabolites**
286   **that could be produced** by the algal metabolic network when combined with the 81 draft
287   bacterial networks for each annotation pipeline (i.e. the predicted producible compounds). As
288   shown in Table 2, based on Prokka, RAST, IMG, MaGe, and DFAST 506, 492, 484, 549, and
289   504 producible compounds were identified, respectively (see Table S8 for a list of compounds).
290   448 (i.e. between 81.6% and 92.6% of the producible compounds) were shared among all tools,
291   but MaGe, RAST, and Prokka also had 34, 11, and 10 exclusive compounds (Fig. 3B).
292   Furthermore, 12 additional compounds became producible by merging the networks of all strains
293   and annotation pipelines.
294
295   Next, we computed **how many strains were needed to enable the production of the same**
296   **metabolic compounds as the entire bacterial communities** for each pipeline (Table 2). As
297   expected, the annotation pipeline affected the number of strains involved in a minimal
298   community, which increased with the number of metabolites that could be produced by the entire
299   bacterial community. The number of selected strains ranged from seven strains needed to
300   produce the 484 metabolites producible by the metabolic networks reconstructed with the IMG
301   pipeline to 13 strains needed to produce the 549 metabolites producible by the MaGe metabolic
302   networks. Fourteen strains were required to produce the full 590 compounds producible by the
303   combination of all networks from all pipelines. We also examined which bacteria were involved
304   in these minimal communities, taking into consideration the union of all minimal communities
305   proposed by MisCoTo. Here, the annotation pipeline had a strong effect on these communities,
306   with strains being selected from sub-sets of bacteria ranging from 15 (Prokka) to 61 (RAST)
307   strains (Table 2). The number of bacteria in each of the sub-sets was not related to the number of
308   producible compounds nor the size of the minimal communities. Fifteen strains were selected in
309   only one of the five pipelines, 14 of which exclusively in the RAST-based community, and one,
310   *Pseudomonas* 9AG, was included in minimal communities only when they were calculated from
311   merged data from all annotation pipelines. Most strains were jointly selected by two or more
312   pipelines, but even Prokka and IMG, for which minimal communities were based on the smallest
313   subset of bacteria (15 and 17 selected strains, respectively) had only six strains in common.
314   However, we also found commonalities between the selected communities: all communities
315   comprised at least one representative of the *Actinomycetales* (*Actinobacteria*), *Bacillales*
316   (*Firmicutes*), *Rhizobiales* (*Alphaproteobacteria*), *Sphingomonadales* (*Alphaproteobacteria*), and
317   *Pseudomonadales* (*Gammaproteobacteria*). Lastly, one strain, *Maribacter* strain 151
318   (*Flavobacteriaceae*, *Bacteroidetes*), was consistently selected in all data sets regardless of the
319   annotation tool (Table 2).
320

321 Finally, given the high variability between the selected communities, we tested how specific
322 these communities were to the dataset used to select them. To this means, we selected the strains
323 comprised in the minimal **microbial communities predicted based on the Prokka** annotations,
324 i.e. 15 bacteria, and examined the metabolites predicted to be produced by this community
325 together with the host when applied to the datasets based on the other pipelines. Briefly, the
326 DFAST, IMG, MaGe, and RAST metabolic networks with the Prokka community predicted
327 similar scopes as the larger communities specifically selected for the dataset: 477 compounds vs
328 492 for RAST, 469 vs 484 for IMG, 488 vs 549 for MaGe, and 448 vs 504 producible
329 compounds for DFAST, respectively (Table S9). This suggests that differences in the community
330 size and composition between the pipelines can be explained mainly by the need to complement
331 each community with bacteria enabling the production of less than 10% of the metabolites.
332
333 **Discussion & Conclusions**
334 In this study, we sought to assess the impact of selected annotation pipelines on draft metabolic
335 network reconstructions and downstream analyses of metabolic complementarity. Our results
336 highlight significant differences between the output of the five tested standard pipelines at all
337 examined levels, from the prediction of coding sequences to the selection of microbial
338 communities, especially for EC annotations and hypothetical proteins, with levels of variability
339 similar to those previously reported by Griesemer et al. (2018). Overall, the number of reactions
340 predicted in the final network mirrored the number of EC numbers predicted, with Prokka and
341 IMG yielding both the highest number of EC annotations and reactions. This underlines the
342 importance of this type of annotation for metabolic reconstructions with Pathway Tools - a link
343 that is not surprising as EC numbers are directly referenced in MetaCyc (Caspi et al. 2017), the
344 database used by Pathway Tools for draft metabolic network reconstruction. Each complete EC
345 number can therefore be translated directly into one or several corresponding metabolic
346 reactions.
347
348 Our analyses also highlight a number of reactions that were predicted specifically for some
349 pipelines, but our analysis of the corresponding pathways did not clearly show any pipelines to
350 favor the annotation of specific biological processes over others. Furthermore, the fact that we
351 observed little differences in the relative performance of the pipelines when comparing them
352 separately for different bacterial phyla, suggests that there is no strong phylum-specificity of the
353 tested pipelines. Our manual examination of a subset of pipeline-specific reactions showed that
354 the reliability of these pipeline-specific predictions is generally low, with a tendency for those
355 pipelines that predict the fewest pipeline-specific reactions (DFAST, RAST) to produce the
356 highest proportion of high-confidence prediction, as well as the most connected networks in
357 terms of dead-end metabolites. The overall high proportion of dead-end metabolites can be
358 explained by the facts that i) during the reconstruction process we discarded reactions without
359 genetic support (gene association), and that ii) metabolic networks underwent no or very little
360 curation. Furthermore, the genomes analyzed here belong to non-model organisms, and hence
361 their metabolism is more difficult to reconstruct than in well-established models.
362
363 Given these differences, a key question is how reliable functional analyses based on these draft
364 metabolic networks are, and how much their results change according to the annotation pipeline
365 employed. Here our results may seem contradictory at first: on one hand, variability at the level
366 of producible metabolites was low; on the other hand, during the selection of microbial

367 communities based on metabolic complementarities, even these small differences in the draft
368 metabolic networks resulted in largely different consortia. These small differences, whether they
369 are erroneous or missing annotations, accumulate across the 81 annotated genomes and are
370 particularly likely to impact the overall set of producible metabolites and hence community
371 selection, as any possible cooperation with an added value to the host is selected. Indeed,
372 although small, differences in the producible metabolites require complementing the bacterial
373 community with strains that specifically enable the production of these metabolites. Overall,
374 however, the low number of pipeline-specific metabolites indicated that all of the selected
375 consortia were functionally similar. Furthermore, we have shown that a given minimal bacterial
376 community (in our case one generated with Prokka data), also yields similar producible
377 metabolites regardless of the annotation pipeline used for the generation of the metabolic
378 networks. This means that the different selected bacterial consortia are likely able to fulfill the
379 same or similar metabolic roles in our metabolic model. This phenomenon has previously also
380 been described in natural alga-associated bacterial consortia (Burke et al. 2011) and was
381 explained by the competitive lottery theory (Sale 1979): several, but not all species/strains can
382 occupy a given niche, and among them, random processes govern which species prevails (lottery
383 model). In our case, such "random" processes could be generated by the 'noise' in the genome
384 annotations and the specificities of each pipeline. Consequently, the niche would correspond to
385 the provision of specific metabolic functions. In line with this analogy, the composition of the
386 bacterial communities in our analyses was not entirely random: although only one strain was
387 present in all communities, each community also contained at least one member of several major
388 taxonomic groups. These groups likely possess specific metabolic capacities absent from other
389 groups making their presence indispensable in all selected consortia. These requirements were
390 detected and met by MiSCoTo regardless of the annotation pipeline.
391
392 Regarding the use of metabolic complementarity as a criterion to select microbial communities,
393 this implies that, as also confirmed experimentally by Burgunter-Delamare et al. (2020), metabolic
394 complementarity can be used to select microbial functions important for a symbiotic community.
395 However, we need to expect high variability in the composition of the selected communities if
396 several strains contain similar metabolic capacities. Just like in nature, there may simply not be
397 one ideal solution that clearly excludes all others. In this sense our data does not provide any
398 evidence that one or another pipeline is more suitable for metabolic network reconstruction – we
399 can only state that for our dataset Prokka and IMG, on average, produced more EC numbers and
400 larger draft metabolic networks with Pathway Tools, while DFAST and RAST are likely to
401 produce fewer false-positive reactions and dead-end metabolites, on average. Furthermore, MaGe
402 produced the largest global scope for the community. One approach to avoid biases introduced by
403 annotation pipelines is to merge results from different pipelines thus maximizing the number of
404 annotations (Kalkatawi et al. 2015). Our data obtained for the merged metabolic networks of all
405 pipelines suggest that this approach would also further extend our metabolic networks and the
406 global scope of the community. However, this approach also comes with a risk, as every additional
407 annotation pipeline may introduce additional errors (Poptsova & Gogarten 2010), especially if the
408 pipeline is not regularly updated (Salzberg 2019). Based on our manual curation of pipeline-
409 specific reactions we may even consider the opposite i.e. basing metabolic complementarity
410 analysis exclusively on reactions that have been predicted independently by two or more pipelines.
411 This will reduce the scope of the networks but likely also result in a reduction of noise and false-
412 positive associations. An additional option may be to increase the stringency of the pipeline

413 settings. Here moderate modification of the e-value in Prokka had little effect on the final results,
414 but further adjustments are possible and their efficiency would likely need to be adapted for each
415 dataset.
416
417 Another application of metabolic complementarity using different annotation pipeline could also
418 be in the context of metagenome analysis, on Metagenome Assembled Genomes (MAGs) of
419 uncultured bacteria. Belcour et al. (2020) have shown that reconstructed metabolic networks from
420 MAGs are similar to the reference genomes, although the average number of reactions was higher
421 in MAGs. Also, the stability of the producible metabolites and selected minimal symbionts was
422 demonstrated for degraded genomes with 2% or more of genes randomly removed. It, therefore,
423 seems probable that the overall stability observed in our study in terms of the predicted metabolic
424 contributions of the symbionts to the algal metabolism regardless of the annotation pipeline will
425 also hold true for metagenomic communities.
426
427 In the long run, continued and extensive experimental validation of bioinformatic predictions in
428 both a culture and a metagenomic context will be key to evaluate and refine the pipelines
429 (Poptsova & Gogarten 2010) and needs to be coupled with increased efforts to expand and
430 improve annotations in reference databases (Carr & Borenstein 2014). In the meantime, one
431 approach to overcome the variability brought by the annotation pipelines is to carefully curate
432 the resulting metabolic networks. This is an indispensable step in obtaining high-quality
433 metabolic networks as stated by Thiele & Palsson (2010). However, this step is costly as it
434 usually requires human expertise and thorough literature exploration. As more and more
435 genomes and metagenomes are available, there is a need for curation-free and reliable metabolic
436 networks to surmount this bottleneck. Non-curated metabolic networks can be informative but
437 are likely to contain specific possibly false-positive functions brought by the annotation, these
438 functions deserve to be examined closely when used for selecting communities of interest.
439
440

Acknowledgments

442 We would like to thank Alireza Asvadi for his help writing a python script for merging tables of
443 all strains across pipelines, and the reviewers of this paper for their constructive remarks. We are
444 grateful to the Institut Français de Bioinformatique (ANR-11-INBS-0013) and BioGenouest
445 Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr) for providing computing
446 resources.
447
448
449

References

451 Aite M, Chevallier M, Frioux C, Trottier C, Got J, Cortés MP, Mendoza SN, Carrier G, Dameron
452   O, Guillaudeux N, Latorre M, Loira N, Markov GV, Maass A, and Siegel A. 2018.
453   Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of
454   genome-scale metabolic models. *PLoS Computational Biology* 14:e1006146.
455   10.1371/journal.pcbi.1006146

456 Amin SA, Hmelo LR, van Tol HM, Durham BP, Carlson LT, Heal KR, Morales RL, Berthiaume
457   CT, Parker MS, Djunaedi B, Ingalls AE, Parsek MR, Moran MA, and Armbrust EV.
458   2015. Interaction and signalling between a cosmopolitan phytoplankton and associated
459   bacteria. *Nature* 522:98-101. 10.1038/nature14488
460 Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass
461   EM, and Kubal M. 2008. The RAST server: rapid annotations using subsystems
462   technology. *BMC Genomics* 9:1. 10.1186/1471-2164-9-75
463 Belcour A, Frioux C, Aite M, Bretaudeau A, Hildebrand F, and Siegel A. 2020. Metage2Metabo,
464   microbiota-scale metabolic complementarity for the identification of key species. *eLife*
465   9:e61968. 10.7554/eLife.61968.
466 Burgunter-Delamare B, KleinJan H, Frioux C, Fremy E, Wagner M, Corre E, Le Salver A,
467   Leroux C, Leblanc C, Boyen C, Siegel A, and Dittami SM. 2020. Metabolic
468   Complementarity Between a Brown Alga and Associated Cultivable Bacteria Provide
469   Indications of Beneficial Interactions. *Frontiers in Marine Science* 7:85.
470   10.3389/fmars.2020.00085.
471 Burke C, Steinberg P, Rusch D, Kjelleberg S, and Thomas T. 2011. Bacterial community
472   assembly based on functional genes rather than species. *Proceedings of the National*
473   *Academy of Sciences* 108:14288-14293. 10.1073/pnas.1101591108.
474 Carr R, and Borenstein E. 2014. Comparative Analysis of Functional Metagenomic Annotation
475   and the Mappability of Short Reads. *PLoS ONE* 9:e105776.
476   10.1371/journal.pone.0105776.
477 Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M,
478   Midford PE, Ong Q, Ong WK, Paley S, Subhraveti P, and Karp PD. 2017. The MetaCyc
479   database of metabolic pathways and enzymes. *Nucleic Acids Research* 46:D633-D639.
480   10.1093/nar/gkx935.
481 Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F,
482   Aury J-M, Badger JH, Beszteri B, Billiau K, Bonnet E, Bothwell JH, Bowler C, Boyen C,
483   Brownlee C, Carrano CJ, Charrier B, Cho GY, Coelho SM, Collén J, Corre E, Da Silva
484   C, Delage L, Delaroque N, Dittami SM, Doulbeau S, Elias M, Farnham G, Gachon
485   CMM, Gschloessl B, Heesch S, Jabbari K, Jubin C, Kawai H, Kimura K, Kloareg B,
486   Küpper FC, Lang D, Le Bail A, Leblanc C, Lerouge P, Lohr M, Lopez PJ, Martens C,
487   Maumus F, Michel G, Miranda-Saavedra D, Morales J, Moreau H, Motomura T,
488   Nagasato C, Napoli CA, Nelson DR, Nyvall-Collén P, Peters AF, Pommier C, Potin P,
489   Poulain J, Quesneville H, Read B, Rensing SA, Ritter A, Rousvoal S, Samanta M,
490   Samson G, Schroeder DC, Ségurens B, Strittmatter M, Tonon T, Tregear JW, Valentin K,
491   von Dassow P, Yamagishi T, Van de Peer Y, and Wincker P. 2010. The *Ectocarpus*
492   genome and the independent evolution of multicellularity in brown algae. *Nature*
493   465:617. 10.1038/nature09016.
494 Conway JR, Lex A, and Gehlenborg N. 2017. UpSetR: an R package for the visualization of
495   intersecting sets and their properties. *Bioinformatics* 33:2938-2940.
496   10.1093/bioinformatics/btx364.
497 Dittami SM, Arboleda E, Auguet J-C, Bigalke A, Briand E, Cárdenas P, Cardini U, Decelle J,
498   Engelen A, Eveillard D, Gachon CMM, Griffiths S, Harder T, Kayal E, Kazamia E,
499   Lallier FH, Media M, Marzinelli EM, Morganti T, Núñez Pons L, Prado S, Pintado J,
500   Saha M, Selosse M-A, Skillings D, Stock W, Sunagawa S, Toulza E, Vorobev A,
501   Leblanc C, and Not F. 2021. A community perspective on the concept of marine

502 holobionts: state-of-the-art, challenges, and future directions. *PeerJ*
503 9:e10911.10.7717/peerj.1091.
504 Dittami SM, Corre E, Brillet-Guéguen L, Lipinska AP, Pontoizeau N, Aite M, Avia K, Caron C,
505 Cho CH, Collén J, Cormier A, Delage L, Doubleau S, Frioux C, Gobet A, González-
506 Navarrete I, Groisillier A, Hervé C, Jollivet D, KleinJan H, Leblanc C, Liu X, Marie D,
507 Markov GV, Minoche AE, Monsoor M, Pericard P, Perrineau M-M, Peters AF, Siegel A,
508 Siméon A, Trottier C, Yoon HS, Himmelbauer H, Boyen C, and Tonon T. 2020a. The
509 genome of Ectocarpus subulatus – A highly stress-tolerant brown alga. *Marine Genomics*
510 52:100740. 10.1016/j.margen.2020.100740.
511 Dittami SM, Duboscq-Bidot L, Perennou M, Gobet A, Corre E, Boyen C, and Tonon T. 2015.
512 Host–microbe interactions as a driver of acclimation to salinity gradients in brown algal
513 cultures. *The ISME journal* 10:51. 10.1038/ismej.2015.104.
514 Dittami SM, Eveillard D, and Tonon T. 2014. A metabolic approach to study algal–bacterial
515 interactions in changing environments. *Molecular Ecology* 23:1656-1660.
516 10.1111/mec.12670.
517 Dittami SM, Peters AF, West JA, Cariou T, KleinJan H, Burgunter-Delamare B, Prechoux A,
518 Egan S, and Boyen C. 2020b. Revisiting Australian Ectocarpus subulatus (Phaeophyceae)
519 From the Hopkins River: Distribution, Abiotic Environment, and Associated Microbiota.
520 *Journal of Phycology* 56:719-729. 10.1111/jpy.12970.
521 Ebenhöh O, Handorf T, and Heinrich R. 2004. Structural analysis of expanding metabolic
522 networks. *Genome Inform* 15:35-45.
523 Egan S, Harder T, Burke C, Steinberg P, Kjelleberg S, and Thomas T. 2013. The seaweed
524 holobiont: understanding seaweed–bacteria interactions. *FEMS Microbiology Reviews*
525 37:462-476. 10.1111/1574-6976.12011.
526 Fraune S, and Bosch TCG. 2010. Why bacteria matter in animal development and evolution.
527 *Bioessays* 32:571-580. 10.1002/bies.200900192.
528 Frioux C, Fremy E, Trottier C, and Siegel A. 2018. Scalable and exhaustive screening of
529 metabolic functions carried out by microbial consortia. *Bioinformatics* 34:i934-i943.
530 10.1093/bioinformatics/bty588.
531 Griesemer M, Kimbrel JA, Zhou CE, Navid A, and D'haeseleer P. 2018. Combining multiple
532 functional annotation tools increases coverage of metabolic annotation. *BMC Genomics*
533 19:948. 10.1186/s12864-018-5221-9.
534 Hagberg A, Swart P, and S Chult D. 2008. Exploring network structure, dynamics, and function
535 using networkx. *Proceedings of the 7th Python in Science Conference (SciPy2008),*:11-
536 18.
537 Haggart CR, Bartell JA, Saucerman JJ, and Papin JA. 2011. Chapter twenty-one - Whole-
538 Genome Metabolic Network Reconstruction and Constraint-Based Modeling. In:
539 Jameson D, Verma M, and Westerhoff HV, eds. *Methods in Enzymology*: Academic
540 Press, 411-433.
541 Kalkatawi M, Alam I, and Bajic VB. 2015. BEACON: automated tool for Bacterial GEnome
542 Annotation ComparisON. *BMC Genomics* 16:616. 10.1186/s12864-015-1826-4.
543 Karimi E, Geslain E, KleinJan H, Tanguy G, Legeay E, Corre E, and Dittami SM. 2019. Genome
544 Sequences of 72 Bacterial Strains Isolated from Ectocarpus subulatus: A Resource for
545 Algal Microbiology. *Genome Biology and Evolution* 12:3647-3655. 10.1093/gbe/evz278.
546 Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, Ong WK,
547 Subhraveti P, Caspi R, Fulcher C, Keseler IM, and Paley SM. 2019. Pathway Tools

548   version 23.0 update: software for pathway/genome informatics and systems biology.
549   *Briefings in Bioinformatics* 8:bbz104. 10.1093/bib/bbz104.
550 KleinJan H, Jeanthon C, Boyen C, and Dittami SM. 2017. Exploring the cultivable Ectocarpus
551   microbiome. *Frontiers in Microbiology* 8:2456. 10.3389/fmicb.2017.02456.
552 Levy R, Carr R, Kreimer A, Freilich S, and Borenstein E. 2015. NetCooperate: a network-based
553   tool for inferring host-microbe and microbe-microbe cooperation. *BMC Bioinformatics*
554   16:164-164. 10.1186/s12859-015-0588-y.
555 Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, Bartell JA, Blank LM,
556   Chauhan S, Correia K, Diener C, Dräger A, Ebert BE, Edirisinghe JN, Faria JP, Feist
557   AM, Fengos G, Fleming RMT, García-Jiménez B, Hatzimanikatis V, van Helvoirt W,
558   Henry CS, Hermjakob H, Herrgård MJ, Kaafarani A, Kim HU, King Z, Klamt S, Klipp
559   E, Koehorst JJ, König M, Lakshmanan M, Lee D-Y, Lee SY, Lee S, Lewis NE, Liu F,
560   Ma H, Machado D, Mahadevan R, Maia P, Mardinoglu A, Medlock GL, Monk JM,
561   Nielsen J, Nielsen LK, Nogales J, Nookaew I, Palsson BO, Papin JA, Patil KR, Poolman
562   M, Price ND, Resendis-Antonio O, Richelle A, Rocha I, Sánchez BJ, Schaap PJ, Malik
563   Sheriff RS, Shoaie S, Sonnenschein N, Teusink B, Vilaça P, Vik JO, Wodke JAH, Xavier
564   JC, Yuan Q, Zakhartsev M, and Zhang C. 2020. MEMOTE for standardized genome-
565   scale metabolic model testing. *Nature Biotechnology* 38:272-276. 10.1038/s41587-020-
566   0446-y.
567 Markowitz VM, Mavromatis K, Ivanova NN, Chen IA, Chu K, and Kyrpides NC. 2009. IMG
568   ER: a system for microbial genome annotation expert review and curation.
569   *Bioinformatics* 25:2271-2278. 10.1093/bioinformatics/btp393.
570 McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, Dubilier
571   N, Eberl G, Fukami T, Gilbert SF, Hentschel U, King N, Kjelleberg S, Knoll AH, Kremer
572   N, Mazmanian SK, Metcalf JL, Nealson K, Pierce NE, Rawls JF, Reid A, Ruby EG,
573   Rumpho M, Sanders JG, Tautz D, and Wernegreen JJ. 2013. Animals in a bacterial
574   world, a new imperative for the life sciences. *Proceedings of the National Academy of*
575   *Sciences* 110:3229-3236. 10.1073/pnas.1218525110.
576 Muller EEL, Faust K, Widder S, Herold M, Martínez Arbas S, and Wilmes P. 2018. Using
577   metabolic networks to resolve ecological properties of microbiomes. *Current Opinion in*
578   *Systems Biology* 8:73-80. 10.1016/j.coisb.2017.12.004.
579 Oksanen J, Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P, O'Hara R,
580   Simpson G, and Solymos P. 2020. Vegan: Community Ecology Package. R package
581   version 2.5-5. 2019. https://CRAN.R-project.org/package=vegan.
582 Philippot L, Raaijmakers JM, Lemanceau P, and van der Putten WH. 2013. Going back to the
583   roots: the microbial ecology of the rhizosphere. *Nature Reviews Microbiology* 11:789-
584   799. 10.1038/nrmicro3109.
585 Poptsova MS, and Gogarten JP. 2010. Using comparative genome analysis to identify problems
586   in annotated microbial genomes. *Microbiology* 156:1909-1917. 10.1099/mic.0.033811-0.
587 Prigent S, Collet G, Dittami SM, Delage L, Ethis de Corny F, Dameron O, Eveillard D, Thiele S,
588   Cambefort J, Boyen C, Siegel A, and Tonon T. 2014. The genome-scale metabolic
589   network of Ectocarpus siliculosus (EctoGEM): a resource to study brown algal
590   physiology and beyond. *The Plant Journal* 80:367-381. 10.1111/tpj.12627.
591 Sale PF. 1979. Recruitment, loss and coexistence in a guild of territorial coral reef fishes.
592   *Oecologia* 42:159-177. 10.1007/BF00344855.

593   Salzberg SL. 2019. Next-generation genome annotation: we still struggle to get it right. *Genome*
594   *Biology* 20:92. 10.1186/s13059-019-1715-2.
595   Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068-2069.
596   10.1093/bioinformatics/btu153.
597   Siegwald L, Caboche S, Even G, Viscogliosi E, Audebert C, and Chabé M. 2019. The Impact of
598   Bioinformatics Pipelines on Microbiota Studies: Does the Analytical "Microscope"
599   Affect the Biological Interpretation? *Microorganisms* 7:393.
600   Siezen RJ, and van Hijum SAFT. 2010. Genome (re-)annotation and open-source annotation
601   pipelines. *Microbial biotechnology* 3:362-369. 10.1111/j.1751-7915.2010.00191.x.
602   Singh RP, and Reddy CRK. 2014. Seaweed–microbial interactions: key functions of seaweed-
603   associated bacteria. *FEMS Microbiology Ecology* 88:213-230. 10.1111/1574-
604   6941.12297.
605   Tanizawa Y, Fujisawa T, and Nakamura Y. 2017. DFAST: a flexible prokaryotic genome
606   annotation pipeline for faster genome publication. *Bioinformatics* 34:1037-1039.
607   10.1093/bioinformatics/btx713.
608   Tapia JE, González B, Goulitquer S, Potin P, and Correa JA. 2016. Microbiota influences
609   morphology and reproduction of the brown alga *Ectocarpus* sp. *Frontiers in*
610   *Microbiology* 7:197. 10.3389/fmicb.2016.00197.
611   Thiele I, and Palsson BØ. 2010. A protocol for generating a high-quality genome-scale metabolic
612   reconstruction. *Nature Protocols* 5:93-121. 10.1038/nprot.2009.203.
613   Vallenet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, Burlot L, Bussell X, Fouteau
614   S, Gautreau G, Lajus A, Langlois J, Planel R, Roche D, Rollin J, Rouy Z, Sabatet V, and
615   Médigue C. 2019. MicroScope: an integrated platform for the annotation and exploration
616   of microbial gene functions through genomic, pangenomic and metabolic comparative
617   analysis. *Nucleic Acids Research* 48:D579-D589. 10.1093/nar/gkz926.
618

Figure 1

Boxplot of the number of A) CDSs, B) EC-numbers (considers all EC-number), C) Hypothetical proteins, and D) Metabolic reactions of all annotated bacterial species through the five pipelines.

The graphs show the deviation of each genome annotated with every one pipeline from the mean across all pipelines in %. The thick horizontal line represents the median of the distribution across all genomes for that pipeline, and circles represent outlier data. Letters above the box-plots indicate statistically significant differences (Tukey's HSD test). Pipelines share the same letter, the differences between them are not significant ($p >= 0.05$). (See Table S10 for details)

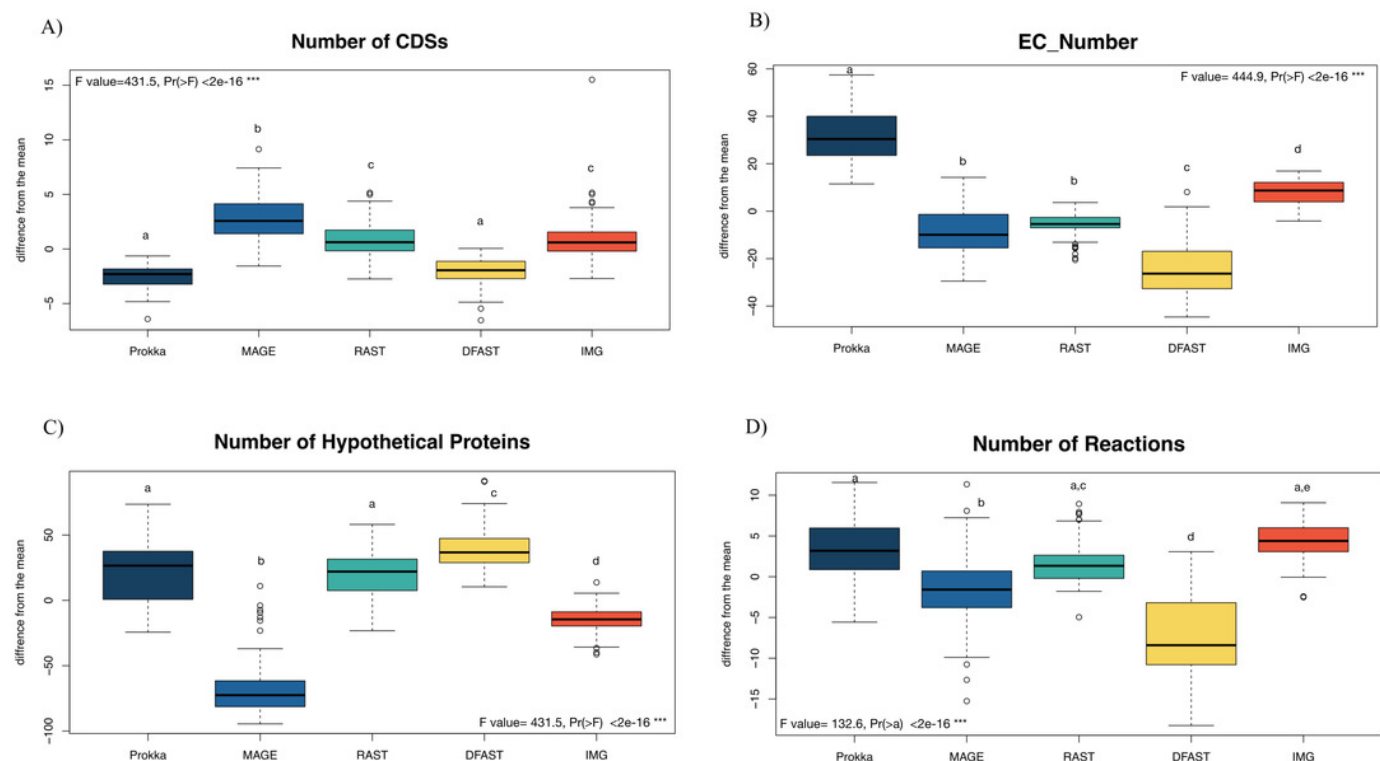



Figure 2

Non-metric Multi-Dimensional Scaling (NMDS) plot of the metabolic networks of the different bacterial strains and pipelines.

The Bray-Curtis index was used as a dissimilarity measure; ellipses show clusters of strains based on bacterial phyla; grey polygons connect annotations of the same genome performed with different pipelines.

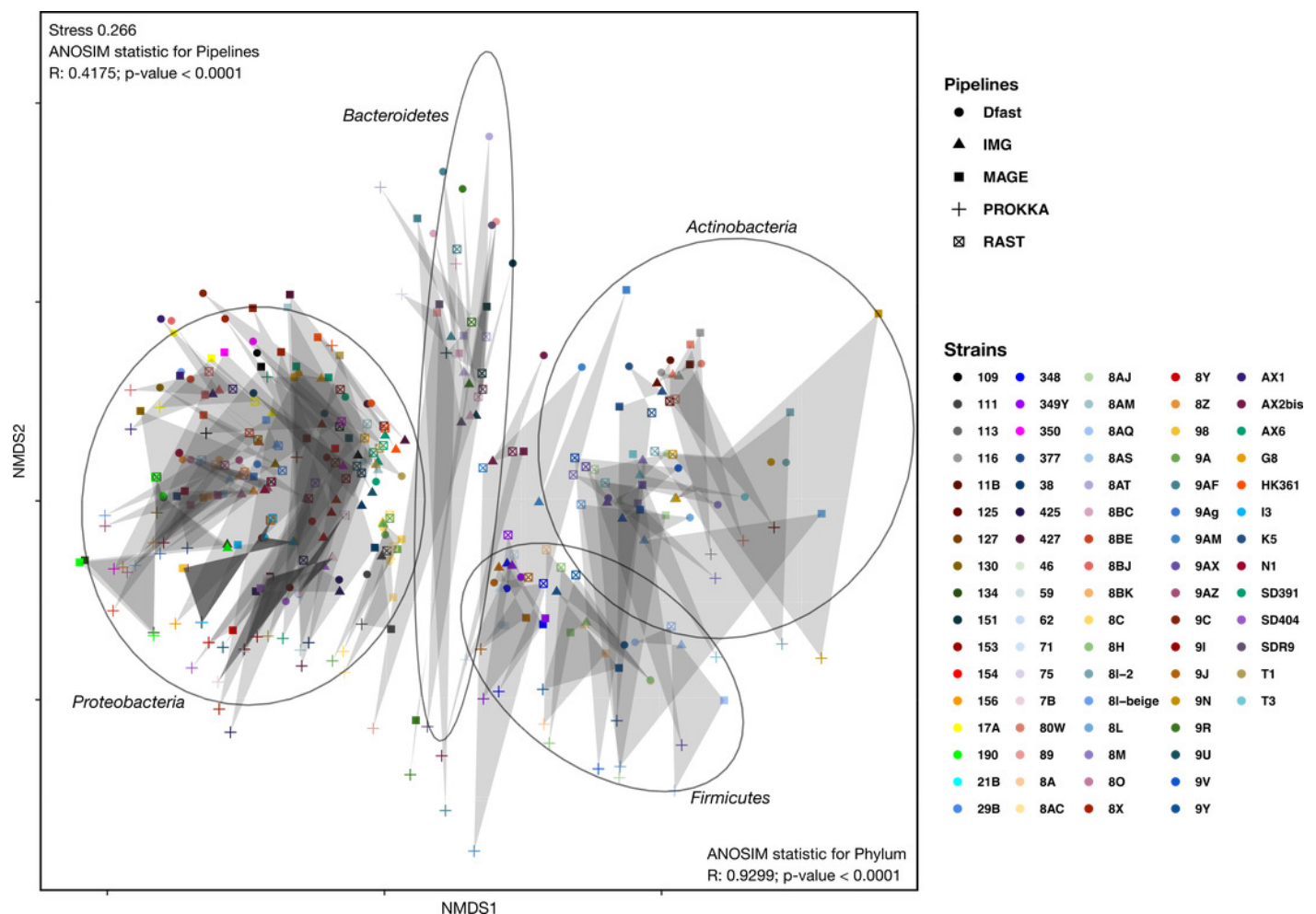
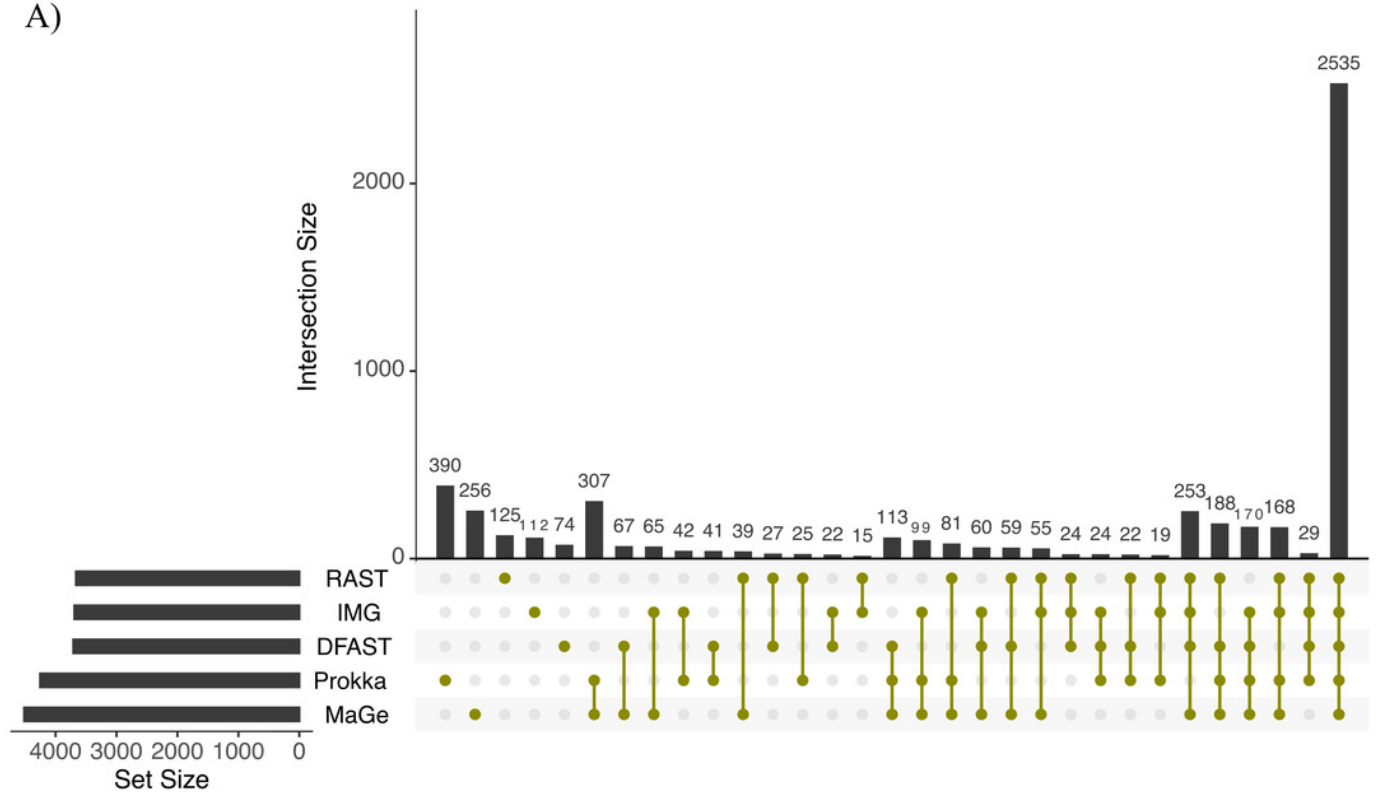
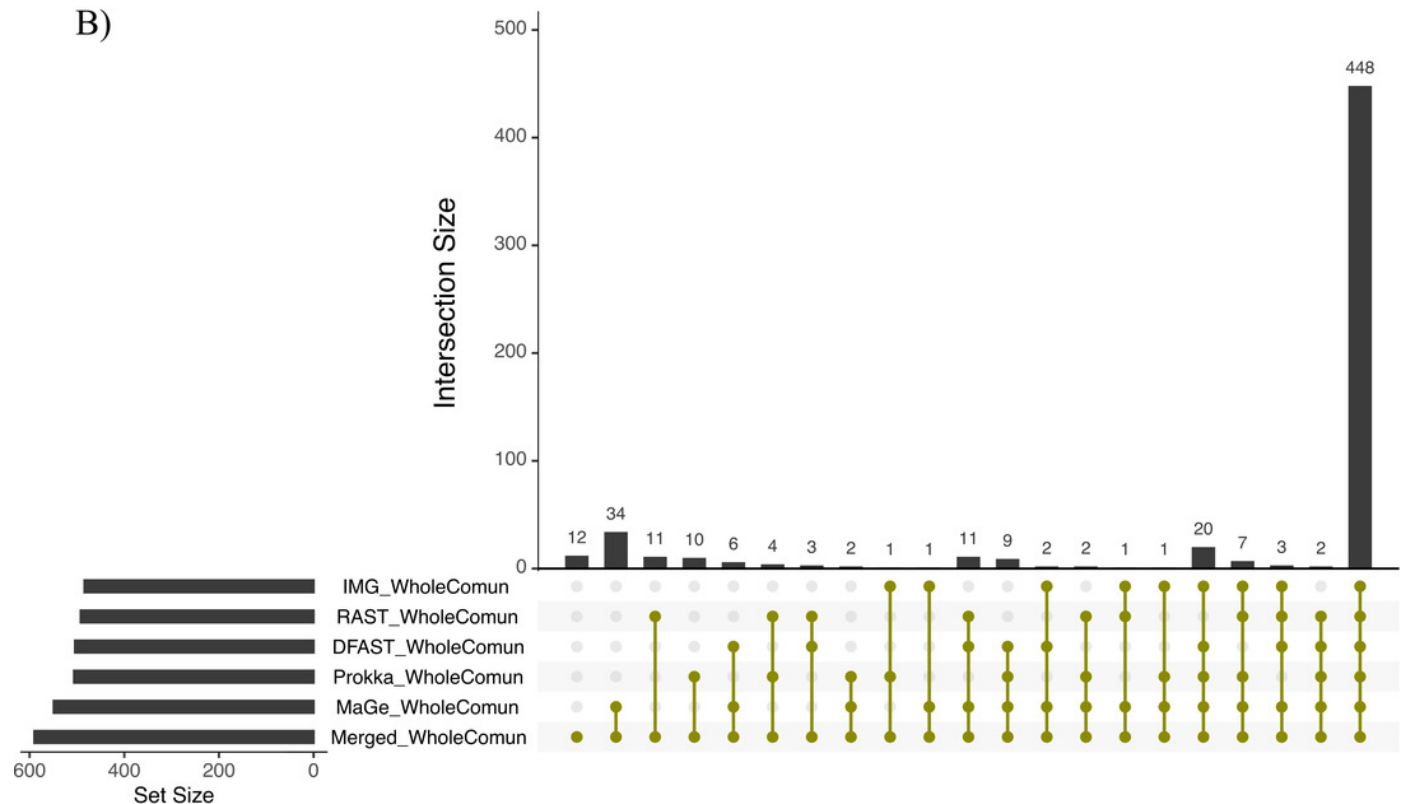



Figure 3

Upset chart showing the overlap in predicted biochemical reactions (A) and producible compounds by the algal host and the complete set of 81 bacteria (scope) (B) annotated with each pipeline.

The horizontal bars on the left show the total number of predicted reactions (A) or producible compounds (B) in each pipeline. The vertical histogram on the right shows the number of overlapping reactions (A) or compounds (B). The total number of reactions and producible compounds for each tool is indicated on the left as 'set size per tool'.

A)

B)

Overrepresented EC numbers in the tested pipelines compared to the average of all pipelines.

The "%over" indicates how many more EC numbers were found in a specific pipeline compared to all pipelines; "∞" indicates that the EC category was predicted only in this pipeline. p-values correspond to a binomial test after Bonferroni correction.

	Description	%over	p
DFAST	1.6 Oxidoreductases acting on NADH or NADPH	52%	<0.001
	2.1 Transferase transferring one-carbon groups	12%	<0.001
	2.7 Transferase transferring phosphorus-containing groups	8%	<0.001
	3.6 Hydrolases acting on acid anhydrides	22%	<0.001
	6.1 Ligases forming carbon-oxygen bonds	17%	<0.001
	6.3 Ligases forming carbon-nitrogen bonds	11%	<0.001
IMG	2.1 Transferase transferring one-carbon groups	7%	0.019
	3.1 Hydrolases acting on ester bonds	9%	<0.001
	3.4 Hydrolases acting on peptide bonds	14%	<0.001
	3.6 Hydrolases acting on acid anhydrides	15%	<0.001
	6.5 Ligases forming phosphoric ester bonds	68%	<0.001
MaGe	1.19 Oxidoreductases acting on reduced flavodoxin as donor	5658%	<0.001
	4.3 Carbon–nitrogen lyases	17%	0.036
	5.6 Isomerases altering macromolecular conformation	∞	<0.001
	7.1 Translocases catalysing the translocation of hydrons	∞	<0.001
	7.2 Translocases catalysing the translocation of inorganic cations and chelates	∞	<0.001
	7.3 Translocases catalysing the translocation of inorganic anions	∞	<0.001
	7.5 Translocases catalysing the translocation of carbohydrates and derivatives	∞	<0.001
	7.6 Translocases catalysing the translocation of other compounds	∞	<0.001
Prokka	1.1 Oxidoreductases acting on CH-OH group of donors	10%	<0.001
	1.10 Oxidoreductases acting on diphenols and related substances as donors	56%	<0.001
	1.12 Oxidoreductases acting on hydrogen as donors	196%	<0.001
	1.13 Oxygenases	37%	<0.001
	1.14 Oxidoreductases acting on paired donors, with oxygen	78%	<0.001
	1.16 Oxidoreductases oxidizing metal ions	56%	<0.001
	2.4 Glycosyltransferases	21%	<0.001
	2.8 Transferase transferring sulfur-containing groups	16%	0.004
	3.1 Hydrolases acting on ester bonds	15%	<0.001
	3.2 Glycosylases	15%	0.004
	3.4 Hydrolases acting on peptide bonds	9%	0.025
	3.6 Hydrolases acting on acid anhydrides	12%	<0.001
	3.8 Hydrolases acting on halide bonds	86%	<0.001
	3.9 Hydrolases acting on phosphorus-nitrogen bonds	856%	<0.001
	4.5 Carbon–halide lyases	∞	<0.001
RAST	1.2 Oxidoreductases acting on the aldehyde or oxo group of donors	18%	<0.001
	2.3 Acyltransferases	11%	<0.001
	2.7 Transferase transferring phosphorus-containing groups	11%	<0.001
	3.3 Hydrolases acting on ether bonds	72%	<0.001
	4.1 Lyases Carbon–carbon lyases	15%	<0.001
	4.2 Carbon–oxygen lyases	12%	<0.001
	5.1 Isomerases - racemases, epimerases	17%	0.008
	5.3 Isomerases - intramolecular oxidoreductases	16%	0.004
	5.4 Isomerases - intramolecular transferases	28%	<0.001
	6.1 Ligases forming carbon-oxygen bonds	12%	0.010
	6.3 Ligases forming carbon-nitrogen bonds	10%	0.002
	6.4 Ligases forming carbon-carbon bonds	41%	<0.001

Selection of minimal bacterial communities based on annotation pipelines.

Strains	Phylogeny	Prokka	IMG	Ma-Ge	DF-AST	RA-ST	Mer-ged
Curtobacterium sp. 8I-2	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Microbacteriaceae(1); Curtobacterium(1);					✓	✓
Microbacterium sp. 8M	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Microbacteriaceae(1); Microbacterium(1);					✓	✓
Plantibacter sp. T3	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Microbacteriaceae(1); Plantibacter(1);					✓	✓
Pseudoclavibacter sp. 8L	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Microbacteriaceae(1); Pseudoclavibacter(1);	✓			✓	✓	✓
Arthrobacter sp. 8AJ	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Micrococcaceae(1); Arthrobacter(1);					✓	✓
Arthrobacter sp. 9V	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Micrococcaceae(1); Arthrobacter(1);			✓			✓
Citricoccus sp. K5	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Micrococcaceae(1); Citricoccus(1);				✓	✓	✓
Micrococcus sp. 116	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Micrococcaceae(1); Micrococcus(1);	✓				✓	✓
Micrococcus sp. 11B	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Micrococcaceae(1); Micrococcus(1);					✓	✓
Micrococcus sp. 80W	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Micrococcaceae(1); Micrococcus(1);					✓	✓
Aeromicrobium sp. 9AM	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Nocardioidaceae(1); Aeromicrobium(1);	✓				✓	✓
Nocardioides sp. AX2bis	Actinobacteria(1); Actinobacteria(1); Actinomycetales(1); Nocardioidaceae(1); Nocardioides(1);			✓		✓	✓
Imperialibacter sp. 89	Bacteroidetes(1); Cytophagia(1); Cytophagales(1); Flammeovirgaceae(1); Imperialibacter(1);			✓	✓		✓
Imperialibacter sp. EC-SDR9	Bacteroidetes(1); Cytophagia(1); Cytophagales(1); Flammeovirgaceae(1); Imperialibacter(1);			✓	✓		
Imperialibacter sp. SDR9	Bacteroidetes(1); Cytophagia(1); Cytophagales(1); Flammeovirgaceae(1); Imperialibacter(1);			✓	✓		✓
Flavobacterium sp. 9R	Bacteroidetes(1); Flavobacteriia(1); Flavobacteriales(1); Flavobacteriaceae(1); Flavobacterium(1);	✓	✓			✓	✓
Frigoribacterium sp. 9N	Bacteroidetes(1); Flavobacteriia(1); Flavobacteriales(1); Flavobacteriaceae(1); Flavobacterium(1);	✓				✓	✓
Maribacter sp. 151	Bacteroidetes(1); Flavobacteriia(1); Flavobacteriales(1); Flavobacteriaceae(1); Maribacter(1);	✓	✓	✓	✓	✓	✓
Sphingobacterium sp. 8BC	Bacteroidetes(1); Sphingobacteriia(1); Sphingobacteriales(1); Sphingobacteriaceae(1); Sphingobacterium(1);					✓	✓
Bacillus sp. 348	Firmicutes(1); Bacilli(1); Bacillales(1); Bacillaceae 1(1); Bacillus(1);					✓	
Bacillus sp. 349Y	Firmicutes(1); Bacilli(1); Bacillales(1); Bacillaceae 1(1); Bacillus(1);				✓	✓	
Bacillus sp. 71	Firmicutes(1); Bacilli(1); Bacillales(1); Bacillaceae 1(1); Bacillus(1);					✓	
Bacillus sp. 9J	Firmicutes(1); Bacilli(1); Bacillales(1); Bacillaceae 1(1); Bacillus(1);	✓				✓	
Exiguobacterium sp. 8A	Firmicutes(1); Bacilli(1); Bacillales(1); Bacillales_Incertae Sedis XII(1); Exiguobacterium(1);					✓	✓
Exiguobacterium sp. 8H	Firmicutes(1); Bacilli(1); Bacillales(1); Bacillales_Incertae Sedis XII(1); Exiguobacterium(1);					✓	✓
Exiguobacterium sp. 9Y	Firmicutes(1); Bacilli(1); Bacillales(1); Bacillales_Incertae Sedis XII(1); Exiguobacterium(1);					✓	✓
Flavobacterium sp. 9AF	Firmicutes(1); Bacilli(1); Bacillales(1); Bacillales_Incertae Sedis XII(1); Exiguobacterium(1);	✓		✓	✓	✓	✓
Staphylococcus sp. 8AQ	Firmicutes(1); Bacilli(1); Bacillales(1); Staphylococcaceae(1); Staphylococcus(1);			✓	✓	✓	✓
Brevundimonas sp. G8	Proteobacteria(1); Alphaproteobacteria(1); Caulobacterales(1); Caulobacteraceae(1); Brevundimonas(1);						✓
Bosea sp. 125	Proteobacteria(1); Alphaproteobacteria(1); Rhizobiales(1); Bradyrhizobiaceae(1); Bosea(1);				✓	✓	
Bosea sp. 127	Proteobacteria(1); Alphaproteobacteria(1); Rhizobiales(1); Bradyrhizobiaceae(1); Bosea(1);				✓	✓	
Bosea sp. 21B	Proteobacteria(1); Alphaproteobacteria(1); Rhizobiales(1); Bradyrhizobiaceae(1); Bosea(1);				✓	✓	
Bosea sp. 29B	Proteobacteria(1); Alphaproteobacteria(1); Rhizobiales(1); Bradyrhizobiaceae(1); Bosea(1);				✓	✓	
Bosea sp. 46	Proteobacteria(1); Alphaproteobacteria(1); Rhizobiales(1); Bradyrhizobiaceae(1); Bosea(1);				✓	✓	
Bosea sp. 62	Proteobacteria(1); Alphaproteobacteria(1); Rhizobiales(1); Bradyrhizobiaceae(1); Bosea(1);				✓	✓	
Bosea sp. 7B	Proteobacteria(1); Alphaproteobacteria(1); Rhizobiales(1); Bradyrhizobiaceae(1); Bosea(1);				✓	✓	
Hoeflea sp. HK425	Proteobacteria(1); Alphaproteobacteria(1); Rhizobiales(1); Phyllobacteriaceae(1); Hoeflea(1);						✓
Rhizobium sp. SD404	Proteobacteria(1); Alphaproteobacteria(1); Rhizobiales(1); Rhizobiaceae(0.96); Rhizobium(0.92);	✓	✓	✓			✓
Roseovarius sp. EC-HK134	Proteobacteria(1); Alphaproteobacteria(1); Rhodobacterales(1); Rhodobacteraceae(1); Roseovarius(1);				✓	✓	✓
Roseovarius sp. SD190	Proteobacteria(1); Alphaproteobacteria(1); Rhodobacterales(1); Rhodobacteraceae(1); Roseovarius(1);				✓	✓	✓
Erythrobacter sp. HK427	Proteobacteria(1); Alphaproteobacteria(1); Sphingomonadales(1); Erythrobacteraceae(1); Erythrobacter(1);	✓					✓
Novosphingobium sp. 9U	Proteobacteria(1); Alphaproteobacteria(1); Sphingomonadales(1); Sphingomonadaceae(1); Novosphingobium(1);	✓					✓
Sphingomonas sp. 8AM	Proteobacteria(1); Alphaproteobacteria(1); Sphingomonadales(1); Sphingomonadaceae(1); Sphingomonas(1);						✓
Sphingomonas sp. AX6	Proteobacteria(1); Alphaproteobacteria(1); Sphingomonadales(1); Sphingomonadaceae(1); Sphingomonas(1);						✓
Sphingomonas sp. EC-HK361	Proteobacteria(1); Alphaproteobacteria(1); Sphingomonadales(1); Sphingomonadaceae(1); Sphingomonas(1);						✓
Sphingomonas sp. EC-SD391	Proteobacteria(1); Alphaproteobacteria(1); Sphingomonadales(1); Sphingomonadaceae(1); Sphingomonas(1);				✓	✓	✓
Sphingomonas sp. T1	Proteobacteria(1); Alphaproteobacteria(1); Sphingomonadales(1); Sphingomonadaceae(1); Sphingomonas(1);					✓	✓
Sphingorhabdus sp. 109	Proteobacteria(1); Alphaproteobacteria(1); Sphingomonadales(1); Sphingomonadaceae(1); Sphingorhabdus(1);					✓	✓
Burkholderia sp. 8Y	Proteobacteria(1); Betaproteobacteria(1); Burkholderiales(1); Burkholderiaceae(1); Burkholderia(1);				✓		✓
Burkholderiales bacterium 8X	Proteobacteria(1); Betaproteobacteria(1); Burkholderiales(1); Comamonadaceae(1); Variovorax(1);				✓	✓	✓
Massilia sp. 9I	Proteobacteria(1); Betaproteobacteria(1); Burkholderiales(1); Oxalobacteraceae(1); Massilia(1);					✓	✓
Aeromonas sp. 8C	Proteobacteria(1); Gammaproteobacteria(1); Aeromonadales(1); Aeromonadaceae(1); Aeromonas(1);	✓	✓			✓	✓
Aeromonas sp. 9A	Proteobacteria(1); Gammaproteobacteria(1); Aeromonadales(1); Aeromonadaceae(1); Aeromonas(1);	✓	✓		✓		✓
Alteromonas sp. 154	Proteobacteria(1); Gammaproteobacteria(1); Aeromonadales(1); Aeromonadaceae(1); Aeromonas(1);					✓	
Alteromonas sp. 38	Proteobacteria(1); Gammaproteobacteria(1); Alteromonadales(1); Alteromonadaceae(1); Alteromonas(1);					✓	
Marinobacter sp. HK377	Proteobacteria(1); Gammaproteobacteria(1); Alteromonadales(1); Alteromonadaceae(1); Marinobacter(1);				✓	✓	✓
Marinobacter sp. N1	Proteobacteria(1); Gammaproteobacteria(1); Alteromonadales(1); Alteromonadaceae(1); Marinobacter(1);				✓	✓	✓
Pantoea sp. 111	Proteobacteria(1); Gammaproteobacteria(1); Enterobacteriales(1); Enterobacteriaceae(1); Pantoea(1);	✓				✓	
Enterobacterales bacterium 8AC	Proteobacteria(1); Gammaproteobacteria(1); Enterobacteriales(1); Enterobacteriaceae(1); Yersinia(0.68);			✓	✓	✓	✓
Halomonas sp. 113	Proteobacteria(1); Gammaproteobacteria(1); Oceanospirillales(1); Halomonadaceae(1); Halomonas(1);			✓	✓		✓
Halomonas sp. 153	Proteobacteria(1); Gammaproteobacteria(1); Oceanospirillales(1); Halomonadaceae(1); Halomonas(1);			✓	✓	✓	✓
Halomonas sp. 156	Proteobacteria(1); Gammaproteobacteria(1); Oceanospirillales(1); Halomonadaceae(1); Halomonas(1);			✓	✓	✓	✓
Halomonas sp. 59	Proteobacteria(1); Gammaproteobacteria(1); Oceanospirillales(1); Halomonadaceae(1); Halomonas(1);			✓	✓	✓	✓
Halomonas sp. 98	Proteobacteria(1); Gammaproteobacteria(1); Oceanospirillales(1); Halomonadaceae(1); Halomonas(1);			✓	✓	✓	✓
Halomonas sp. I3	Proteobacteria(1); Gammaproteobacteria(1); Oceanospirillales(1); Halomonadaceae(1); Halomonas(1);			✓	✓		✓
Acinetobacter sp. 8BE	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Moraxellaceae(1); Acinetobacter(1);			✓		✓	
Acinetobacter sp. 8I-beige	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Moraxellaceae(1); Acinetobacter(1);					✓	✓
Enhydrobacter sp. AX1	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Moraxellaceae(1); Enhydrobacter(0.85);					✓	
Moraxellaceae bacterium 17A	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Moraxellaceae(1); Enhydrobacter(0.91);				✓	✓	
Enhydrobacter sp. 8BJ	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Moraxellaceae(1); Enhydrobacter(1);				✓	✓	
Pseudomonas sp. 8AS	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Pseudomonadaceae(1); Pseudomonas(1);			✓		✓	✓
Pseudomonas sp. 8BK	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Pseudomonadaceae(1); Pseudomonas(1);					✓	
Pseudomonas sp. 8O	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Pseudomonadaceae(1); Pseudomonas(1);	✓				✓	
Pseudomonas sp. 8Z	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Pseudomonadaceae(1); Pseudomonas(1);				✓	✓	✓
Pseudomonas sp. 9Ag	Proteobacteria(1); Gammaproteobacteria(1); Pseudomonadales(1); Pseudomonadaceae(1); Pseudomonas(1);					✓	✓
Number of compounds producible by the community;		506	484	549	492	504	590
Size of the minimal solution:		10	7	13	10	9	14

| Total number of bacteria found in at least one solution (union): | 15 | 17 | 21 | 28 | 63 | 46 |