



HAL
open science

Inferring and comparing metabolism across heterogeneous sets of annotated genomes using AuCoMe

Arnaud Belcour, Jeanne Got, Méziane Aite, Ludovic Delage, Jonas Collen, Clémence Frioux, Catherine Leblanc, Simon Dittami, Samuel Blanquart, Gabriel Markov, et al.

► To cite this version:

Arnaud Belcour, Jeanne Got, Méziane Aite, Ludovic Delage, Jonas Collen, et al.. Inferring and comparing metabolism across heterogeneous sets of annotated genomes using AuCoMe. 2023. hal-04192851v2

HAL Id: hal-04192851

<https://hal.science/hal-04192851v2>

Preprint submitted on 2 May 2023 (v2), last revised 31 Aug 2023 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 Inferring and comparing metabolism across heterogeneous sets of
2 annotated genomes using AuCoMe

3 Arnaud Belcour^{1,*}  Jeanne Got^{1,} , Méziane Aite^{1,} , Ludovic Delage², Jonas Collén²,
Clémence Frioux³, Catherine Leblanc², Simon M. Dittami², Samuel Blanquart¹,
Gabriel V. Markov^{2,†} and Anne Siegel^{1,†}

4 March 23, 2023

5 1. Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France

6 2. Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station
7 Biologique de Roscoff (SBR), 29680 Roscoff, France

8 3. Inria, INRAE, Université de Bordeaux, France

9  These authors contributed equally to this work. - † Co-last authors - * Corresponding authors:
10 Arnaud Belcour and Anne Siegel.

11 Keywords: genomics, metabolism, metabolic evolution, genomes, systems biology

12 Running Title: AuCoMe: Automatic Comparison of Metabolism

13 Character count (including spaces): 59743

14 **Abstract**

15 Comparative analysis of Genome-Scale Metabolic Networks (GSMNs) may yield important infor-
16 mation on the biology, evolution, and adaptation of species. However, it is impeded by the high

17 heterogeneity of the quality and completeness of structural and functional genome annotations,
18 which may bias the results of such comparisons. To address this issue, we developed AuCoMe –
19 a pipeline to automatically reconstruct homogeneous GSMNs from a heterogeneous set of anno-
20 tated genomes without discarding available manual annotations. We tested AuCoMe with three
21 datasets, one bacterial, one fungal, and one algal, and demonstrated that it successfully reduces
22 technical biases while capturing the metabolic specificities of each organism. Our results also point
23 out shared metabolic traits and divergence points among evolutionarily distant algae, underlining
24 the potential of AuCoMe to accelerate the broad exploration of metabolic evolution across the tree
25 of life.

26 Introduction

27 The comparison of genomic data gave rise to today’s view of the three domains of life: bacteria,
28 archaea, and eukaryotes, being divided into several supergroups (Burki et al., 2020). The evolution
29 of the organisms within these lineages is linked to their ability to adapt to their environment and,
30 therefore, to the plasticity of their metabolic responses. In this context, the analysis of Genome-
31 Scale Metabolic Networks (GSMNs) constitutes a powerful approach, both for graph-based and
32 metadata comparison and, when compatible, for flux-based approaches (Gu et al., 2019). The
33 number of sequences available in public databases is continuously rising, as illustrated by the Gen-
34 Bank database, which grew by 74.30% for Whole Genome Shotgun data in 2019 compared to 2018
35 (Sayers et al., 2019). GSMN reconstruction is theoretically possible for any genome and has already
36 been used to explore evolutionary questions. Metabolic relationships in 975 organisms from the
37 three domains of life showed that these domains were well-separated Schulz and Almaas (2020).
38 Using GSMN reconstruction in bacteria, metabolic and phylogenetic distances between *Escherichia*
39 *coli* and *Shigella* strains could be explained by the parasitic lifestyle of the latter (Vieira et al.,
40 2011). Another GSMN-based study of 301 genomes from the human gut microbiota identified
41 marginal metabolic differences at the microbiota family level but significant metabolic differences
42 between closely related species (Bauer et al., 2015). Analysis of fungal GSMNs additionally demon-
43 strated correlation between metabolic distances and the phylogeny of *Penicillium* species, even if
44 no connection was found between the metabolic distances and the species habitat (Prigent et al.,
45 2018). In brown algae, the GSMNs of *Saccharina japonica* and *Cladosiphon okamuranus* (Nègre
46 et al., 2019) were compared to the GSMN of *Ectocarpus siliculosus* revealing that heterogeneity of
47 genome annotations may have a stronger impact on GSMNs than genuine biological differences.

48 For most GSMN analyses, some limitations still need to be addressed (Bernstein et al., 2021).
49 When comparing different GSMNs, two main biases concern the variable quality of genome anno-
50 tations and the multitude of reconstruction approaches. A variety of methods exists to perform
51 structural (gene structure prediction) and functional (association of functions to genes) annota-
52 tion steps (Yandell and Ence, 2012) and the method choice has previously been shown to have

53 direct effects on the reconstructed GSMNs (Karimi et al., 2021). Similarly, numerous methods
54 for GSMN reconstruction have been developed, e.g. Pathway Tools (Karp et al., 2019), RAVEN
55 (Wang et al., 2018), merlin (Dias et al., 2015; Capela et al., 2022), KBase (Arkin et al., 2018),
56 ModelSEED (Devoid et al., 2013), AuReMe (Aite et al., 2018), AutoKEGGRec (Karlsen et al.,
57 2018), CarVeMe (Machado et al., 2018), and gapseq (Zimmermann et al., 2021). They rely on
58 one or several metabolic databases such as MetaCyc (Caspi et al., 2020), KEGG (Kanehisa and
59 Goto, 2000; Kanehisa et al., 2017), ModelSEED (Seaver et al., 2021) or BiGG (King et al., 2016).
60 Despite efforts in the direction of database reconciliation (Moretti et al., 2021), the heterogeneity
61 of metabolic databases requires time-consuming matching of their respective identifiers and may
62 thus impede the comparison of the GSMNs.

63 One strategy to resolve the issue of GSMN comparison is to work directly on GSMNs. A first
64 method is the *reconstruction annotation jamboree* (Thiele and Palsson, 2010), a community effort to
65 curate pathway discrepancies by examining reactions, Gene-Protein-Reaction (GPR) associations,
66 and metabolites in GSMNs in order to create a consensus GSMN for an organism. This is relevant
67 for organisms for which multiple GSMNs exist, in order to establish a reference one. This strategy
68 was successfully applied to *Salmonella typhimurium* LT2 (Thiele et al., 2011) as well as *Saccha-*
69 *romyces cerevisiae* (Herrgård et al., 2008), and later multiple organisms to create a panmetabolism
70 of 33 fungi (Correia and Mahadevan, 2020). Although platforms now facilitate such community
71 efforts (Cottret et al., 2018), these methods are costly in terms of the manpower involved.

72 A second strategy to resolve GSMN comparison issues is to adapt the GSMN reconstruction
73 method. This strategy aims at reducing annotation biases through the reconstruction of GSMNs
74 from homogeneously annotated genomes using the same method and database, possibly followed
75 by the propagation of annotations with sequence alignments (Vieira et al., 2011; Prigent et al.,
76 2018). This strategy was pushed forward and automatized in the tool CoReCo, which enabled
77 the reconstruction of gap-less metabolic networks from several non-annotated genomes (Pitkänen
78 et al., 2014; Castillo et al., 2016). The main limitation of such approaches is that the re-annotation
79 of the genomes supplants the previous genome annotation.

80 Annotations of genomes in databases also reflect the expertise of scientists. Their quality and
81 precision, ranging from structural features, such as accuracy of intron-exon boundaries or func-
82 tional inferences, like the assignation to a specific catalytic activity based on previous biochemical
83 evidence, highly depend on the amount of curation effort done after the initial automated steps.
84 Such valuable information is lost during a systematic re-annotation step. For a reliable interpre-
85 tation of data, expert annotations therefore ought to be preserved while automatically inferring
86 metabolic networks from any type of genomic resource. In this article, we introduce a new method,
87 *AuCoMe* (Automated Comparison of Metabolism) that creates a set of homogenized GSMNs from
88 heterogeneously-annotated genomes. This enables a less biased functional comparison of the net-
89 works and the determination of metabolic distances using the presence/absence of reactions. Our
90 objective was to develop an efficient and robust approach, which does not depend on the quality
91 of the initial annotations and is able to aggregate heterogeneous information in both prokaryote
92 and eukaryote datasets. AuCoMe combines metabolic network reconstruction, propagation, and
93 verification of annotations. The method automatizes the strategy of transferring information from
94 the annotations of the genomes and complements this information transfer with local searches of
95 missing structural annotations. AuCoMe was applied to three heterogeneous datasets composed of
96 fungal, algal, and bacterial genomes. Our results demonstrate that AuCoMe succeeds at propagat-
97 ing missing reactions to degraded metabolic networks while capturing the metabolic specificities
98 of organisms despite profound differences in the quality of genome annotations. This provides a
99 knowledge base for the comparison of metabolisms between different organisms.

100 **Results**

101 **A tool for homogenizing metabolism inference**

102 AuCoMe is a Python package that aims to build homogeneous metabolic networks and pan-
103 metabolisms starting from genomes with heterogeneous functional and structural annotations. Au-
104 CoMe propagates annotation information among organisms through a four-step pipeline (Fig. 1).

105 The AuCoMe pipeline was tested on three datasets composed of genomes that offer different

106 levels of phylogenetic diversity. The *bacterial dataset* includes 29 genomes belonging to different
107 species of *Escherichia* and closely related *Shigella*, the *fungus dataset* (74 fungal genomes and 3 out-
108 group genomes) covers a range of different phyla within this kingdom, and finally the *algal dataset*
109 (36 algal genomes and 4 outgroup genomes) exhibits the highest phylogenetic diversity including
110 eukaryotes from the supergroups SAR (Stramenopiles, Alveolata and Rhizaria), Haptophyta, Cryp-
111 tophyta, and Archaeplastidia. For all species included in the three datasets, we used annotated
112 genomes publicly available (see Supplemental Tables S1, S2, S3). Run times of AuCoMe on a
113 cluster were 7 hours (10 CPUs), 25 hours (40 CPUs), and 45 hours (40 CPUs) for the bacterial,
114 fungal, and algal datasets, respectively. Details for individual steps are reported in Supplemental
115 file, section 2.

116 In the first step, the *draft reconstruction* step, draft metabolic networks are automatically in-
117 ferred from the original annotations (especially Gene Ontology (GO) terms and Enzyme Commis-
118 sion (EC) numbers) using Pathway Tools (Fig. 1A). Only reactions supported by gene associations
119 or spontaneous reactions were kept in the draft metabolic networks (see Methods). The GSMNs
120 reconstructed at this step from the three datasets exhibit highly heterogeneous reactions (Fig. 2A
121 and blue bars in Fig. 2 B, C, D, see also Supplemental Fig. S1, S2, and S3). Notably in the
122 fungal dataset, no reactions were inferred from annotations in seven species, and 12 draft GSMNs
123 contained less than ten reactions. For the latter, their respective genome annotations included no
124 EC number, and eleven did not include any GO term.

125 Similar observations were also made, although to a lesser extent, for the algal genome dataset,
126 with seven genomes having more than 2,000 reactions and seven genomes less than 500 reactions.
127 At this step, high heterogeneity in the number of reactions can be attributed mainly to differences in
128 the quality and quantity of the functional annotations provided, precluding biologically meaningful
129 comparisons of the GSMNs obtained at the draft reconstruction step. Those initial results from
130 Pathway Tools are a good proxy for the quality of initial genome annotations.

131 The resulting GSMNs and their proteomes were then subjected to comparative genomic analyses
132 in the *orthology propagation* step. During this process, GPR associations are propagated across

133 GSMNs according to orthology relations established using OrthoFinder (Fig. 1B). A robustness
134 filter (see Methods) then selects the robust GPR relationships among all propagated associations.
135 After this step, we observed an homogenization of the number of reactions in the datasets (orange
136 bars in Fig. 2, Supplemental Fig. S1, S2, and S3). The fungal dataset exhibits an outlier at this step;
137 the GSMN of *Encephalitozoon cuniculi* contained only 681 reactions compared to over thousand
138 reactions in the other fungal GSMNs. This is consistent with this species being a microsporidian
139 parasite with a strong genome and gene compaction (Gridale et al., 2013). In all datasets, among
140 the reactions propagated by orthology, a few hundred were removed because they did not fulfill the
141 robustness score criterion (see Methods).

142 A third step (the *structural verification*) consists in checking for the presence of additional
143 GPR associations by finding missing structural annotations in all genomes (Fig. 1C). Compared
144 to the orthology propagation, the *structural verification* step had a smaller impact on the size
145 of the final networks (green bars in Fig. 2, Supplemental Fig. S1, S2, and S3). Ninety-five
146 percent of the GSMNs received less than 28 reactions during this step, and the maximum was
147 209. In the bacterial dataset, the six *Shigella* received more reactions at this step compared to
148 the other strains (on average 76.2 vs. 7.4). After a manual examination, a majority of these
149 reactions were associated with pseudogenes. For the fungal dataset, AuCoMe added 209 reactions
150 for *Saccharomyces kudriavzevii*. These reactions were associated with 192 sequences recovered
151 during the structural step. For all of these sequences, we found corresponding transcripts in a
152 published transcriptome dataset (Blevins et al., 2021). As for the algal dataset, 86 reactions were
153 added for *Ectocarpus subulatus*. We validated the presence of 59 out of 65 genes (83 out of 86
154 reactions) by associating them with existing transcripts. The remaining six genes (three reactions)
155 corresponded to plastid sequences that had remained in the nuclear genome assembly. In both
156 fungal and algal datasets, the structural completion step was, therefore, able to recover sequences
157 likely to correspond to functional genes.

158 Finally, the *spontaneous completion* step (Fig. 1D) adds spontaneous reactions to each metabolic
159 network if these reactions complete BioCyc pathways (red bars in Fig. 2, Supplemental Fig. S1,

160 S2, and S3). For the fungal dataset, this step added between two and 23 spontaneous reactions,
161 leading to two to 27 additional MetaCyc pathways that achieved a completion rate equal to 100%.
162 For the algae, the same step added between 4 and 36 spontaneous reactions yielding two to 31
163 additional pathways. The fewer reactions were inferred at the draft reconstruction step, the more
164 spontaneous reactions were added to complete pathways (Pearson R = -0.83 and -0.84 for the fungal
165 and algal datasets, respectively). The addition of these spontaneous reactions to the ones predicted
166 by Pathway Tools (only other step predicting this type of reactions) lead to the prediction of less
167 than a hundred spontaneous reactions per GSMN.

168 When looking at the size of the final networks, overall, in the three datasets, the final GSMNs
169 were of similar size after applying AuCoMe regardless of the quantity and quality of their corre-
170 sponding genome annotations. In the bacterial dataset (Fig. 2B and Supplemental Fig. S1) the
171 networks of *Shigella* strains comprised fewer reactions than the rest (average of 2,148 reactions vs.
172 2,294, Wilcoxon rank-sum test $W = 138$, $P = 2e-4$). This is consistent with the results of Vieira
173 et al. (2011). On the other hand, *E. coli* K-12 MG1655 stood out with 2,568 reactions compared
174 to 2,047 to 2,342 for the other strains. This can be explained by the curation on this strain and
175 the fact that reactions propagated from *E. coli* K-12 MG1655 to the other strains were frequently
176 supported by only one gene predicted at the draft reconstruction step, and were removed after the
177 orthology propagation (see Methods).

178 **Validation of AuCoMe predictions**

179 To estimate the quality of the predictions made by AuCoMe, experiments were performed.

180 In the first experiment, we compared the GSMNs created by AuCoMe to those created by
181 CarveMe, ModelSEED and gapseq on the bacterial dataset (Supplemental Fig. S4). On this
182 dataset, AuCoMe performed well regarding the recovery of EC numbers, although, it does not
183 reconstruct the largest GSMNs, limiting the inference of reactions to those associated with genes.
184 The ECs inferred by the different tools for *E. coli* K-12 MG1655 were compared with a reference
185 containing ECs from EoCyc, KEGG, BiGG and ModelSEED associated with *E. coli* K-12 MG1655

186 (Supplemental Fig. S5). In this comparison, AuCoMe predicted the highest number of true positives
187 (Supplemental Fig. S6).

188 Then, a second comparison on the eukaryotes was performed with AuCoMe, gapseq find module
189 and ModelSEED on 5 fungal genomes. Results on the eukaryotic genomes showed that AuCoMe
190 predicts the most EC numbers, reactions, and pathways in species distant from the model ones
191 (Supplemental Table S5 and Figure S7). A comparison with metabolic pathways contained in
192 YeastCyc for the genome of *Saccharomyces cerevisiae S288C* was done to estimate the quality
193 of the predicted pathways. AuCoMe predicted a high number of pathways with low completion
194 rate not found in YeastCyc (Supplemental Figure S8). For pathways with completion rate above
195 70%, AuCoMe and gapseq exhibited similar performance (Supplemental Fig. S9). Although these
196 experiments should be confirmed by an exhaustive comparative study, these results suggest that
197 AuCoMe is suitable for the study of the metabolism of multiple eukaryotic genomes by predicting
198 robust gene-reactions associations.

199 The third evaluation of the reliability of the reconstruction process was performed on the final
200 algal dataset. We manually examined 100 random GPR associations across the metabolic networks
201 generated by AuCoMe: 50 reactions that were predicted to be present and 50 reactions that were
202 predicted to be absent (see methods). Not counting spontaneous reactions, manual annotations
203 and automatic predictions corresponded in 86% of all cases (42/49) for the reactions predicted to
204 be present and in 91% (40/44) for the reactions predicted to be absent (see Supplemental Tables
205 S6 and S7). These data underline the robustness of the AuCoMe pipeline.

206 For the fourth verification, we extracted the EC numbers of all reactions of the fungal and the
207 algal dataset GSMNs for which GPR associations were only predicted by orthology. For each EC
208 number, we extracted the associated protein sequence and used DeepEC (Ryu et al., 2019) to infer
209 EC numbers and compared them to the EC numbers linked to the reaction by the pipeline. An
210 enrichment of sequences confirmed by DeepEC is observed in robust GPR associations compared
211 to those discarded by the filter: 26% vs. 4.8 % in the fungal dataset and 13.6% vs. 1.4% in
212 the algal dataset (see Supplemental Fig S10). This confirms that the robustness filter removed

213 predominantly poorly supported reactions.

214 In the fifth experiment, thirty-two datasets were formed, each containing the 29 bacterial *E.*
215 *coli* and *Shigella* strains studied in Vieira et al. (2011), among them a replicate of the *E. coli* K-12
216 MG1655 genome degraded to a variable extent in its functional and/or structural annotations
217 (see Methods and Supplemental Table S4). The manually-curated EcoCyc database (Karp et al.,
218 2018a) was used to check the reliability of the GSMN reconstructed for each corresponding degraded
219 genome. For each of the 32 datasets, F-measures were computed at each AuCoMe step according to
220 comparisons of the reconstructed GSMN with the gold-standard EcoCyc database (see Methods).
221 Fig. 3A illustrates the number of reactions predicted by AuCoMe for the *E. coli* K-12 MG1655
222 GSMN in each of the 32 synthetic bacterial datasets to assess the importance of each step in the
223 homogenization of the GSMN sizes. Fig. 3B represents the F-measure for the corresponding dataset.
224 As expected, the more the genomes were degraded, the lower the F-measures were. The orthology
225 propagation alleviated this degradation for functionally degraded genomes (dataset labeled 1 to 10).
226 And the structural verification step compensated the loss of annotation in structurally degraded
227 genomes (datasets labeled 22 to 31). With both types of degradation (datasets 11 to 21), the
228 combination of the two steps recovered lost reactions.

229 Notably, even when 100% of the *E. coli* K-12 MG1655 functional and structural annotations
230 are degraded, the information from the other 28 non-altered genomes enabled the recovery of 2,244
231 reactions (Fig. 3A, dataset 31) and a F-measure of 0.60. Altogether, these results demonstrate that,
232 by taking advantage of the annotations present in the other genomes of the considered dataset, Au-
233 CoMe builds GSMNs with reactions even for genomes completely missing functional and structural
234 annotations.

235 **Exploration of Calvin cycle and pigment pathways in algae**

236 The accuracy of the annotation transfer procedure by AuCoMe was further assessed using two
237 pathways where there were clear biological expectations in the algal dataset. The Calvin cycle
238 is a biochemical pathway present in photosynthetic organisms to fix CO_2 into three-carbon sugars

239 composed of 13 reactions (MetaCyc identifier: CALVIN-PWY, Fig. 4).

240 The three main AuCoMe steps are required to obtain a homogeneous view of this pathway in all
241 organisms. The draft reconstruction (blue) and the orthology propagation (orange) steps provide
242 most of the reactions. The robustness criterion (grey) applied during the orthology propagation step
243 removed a GPR association with the reaction RIBULOSE-BISPHOSPHATE-CARBOXYLASE-
244 RXN for the non-photosynthetic fungus *Neurospora crassa*. The structural verification step added
245 one reaction (RIBULP3EPIM-RXN) for *Porphyra umbilicalis* (green square in Fig. 4). The G3P
246 dehydrogenase reaction (1.2.1.13-RXN) had to be added manually in brown algae, diatoms and
247 haptophytes because the canonical plastidial gene has been replaced by a cytosolic paralog (Liaud
248 et al., 1997). Similarly, the EC number associated with the reaction SEDOBISALDOL-RXN was
249 incomplete (only three digits) in the MetaCyc version used and not found in the 40 GSMNs, and
250 therefore manually added to the 40 GSMNs (GPR associations are indicated in yellow in Fig. 4,
251 for details, see Supplementary data).

252 A similar analysis was performed on pathways producing phycobilins in five brown algae (Sup-
253 plemental Fig S11). As for the Calvin cycle, reactions in the pathways were added during draft
254 reconstruction, orthology propagation and spontaneous completion. The finding of those path-
255 ways in brown algae may appear contradictory with the loss of associated phycobiliproteins during
256 evolution (Bhattacharya et al., 2004). However, the retention of enzymes related to phycobin
257 biosynthesis is linked with their cooption from a role as photosynthetic pigments to a function of
258 signaling within photoreceptors (Rockwell and Lagarias, 2017).

259 Both of these analyses highlight the potential of AuCoMe to help understand metabolism and
260 its evolution in a group of non-model organisms by predicting candidate GPRs and pathways.

261 **AuCoMe GSMNs are consistent with species phylogeny**

262 To further assess the predictions of AuCoMe and to explore biological features, we clustered the
263 GSMNs of the algal dataset after the draft reconstruction as well as at the end of the pipeline
264 by using the presence or absence of reactions in the GSMNs (see Fig. 5A). We compared these

265 clusterings with a phylogeny compiled from Strassert et al. (2021). The initial GSMNs produced
266 from the annotations exhibited low consistency with the phylogenetic relationships. Even well-
267 established phylogenetic groups like red algae or brown algae were not recovered. At this step, the
268 principal factor leading to the repartition of points in the MDS was the heterogeneity of genome
269 annotations. An ANOSIM test supports this as it was not able to differentiate the main phylogenetic
270 groups ($R=0$, $P\text{-value}=0.45$). However, in the MDS made from the GSMNs after the final step
271 of AuCoMe, we observed a clear separation between the known phylogenetic groups, supported
272 by an ANOSIM test ($R=0.811$, $P\text{-value}=1e-04$). This is also visible in the dendrograms clustering
273 the GSMNs generated by the complete AuCoMe pipeline, which was broadly consistent with the
274 reference species phylogeny (Fig. 5B). There were only three higher order inconsistencies concerning
275 *C. paradoxa*, for which the genome version deposited in GenBank fully lacked expert annotations
276 (Price et al., 2012), *G. theta*, which belongs to Cryptophytes, for which the phylogenetic position
277 is controversial (Strassert et al., 2021), and *N. gaditana*, which was the only representative of
278 eustigmatophycean stramenopiles. The two other stramenopile groups, diatoms and brown algae,
279 were represented by multiple species which likely minimizes errors linked with peculiarities of a
280 single genome. There were also some minor inconsistencies in intra-group relationships, in green
281 algae, diatoms, brown algae, and opisthokonts.

282 An illustration of the efficiency of AuCoMe was the *de novo* reconstruction of the GSMN of
283 the glaucophyte *C. paradoxa*. For the reconstruction of this GSMN, we used the initially published
284 genome sequence, which contained only two functionally-annotated genes (Price et al., 2012). The
285 draft reconstruction by AuCoMe enabled us to retrieve 1,675 GPRs, a number within the same
286 range as the other species from the dataset. Accordingly, *C. paradoxa* branched at the basis of
287 the dendrogram after the draft reconstruction step, whereas it moved to the archeplastids after
288 the orthology propagation step. Even if the grouping of *C. paradoxa* within archeplastids with
289 the streptophytes *Chara braunii* and *Klebsormidium nitens* does not reflect the phylogenetic rela-
290 tionships, this shows that AuCoMe is a reasonable proxy for handling nearly unannotated genome
291 sequences.

292 By exploring cluster of reactions shared in phylogenetic groups (as shown in Supplemental Fig
293 S12), results of AuCoMe could pave the way to the identification of gene candidates for enzymatic
294 reactions. We analyzed a cluster of fourteen reactions present in *Cladosiphon okamuranus* and
295 *Saccharina japonica* but absent in other brown algae (see Supplemental Table S8). Among those
296 fourteen reactions, twelve were enzymatic reactions assigned based on annotations, but orthology
297 propagation in the AuCoMe pipeline identified only a subset of the potential orthologs (see Sup-
298 plemental Table S9). A focus was made on the o-aminophenol oxidases. Comparative genomics
299 analysis using sequences from additional BLASTP searches showed that potential homologs were
300 present for the other brown algae (see Supplemental Fig. S13). The o-aminophenol oxidase family
301 proteins present in the genome of *E. siliculosus* are predicted to be cytoplasmic, extracellular, or to
302 target the membrane (see Supplemental Table S10), suggesting different roles depending on their
303 subcellular localization. In this case, AuCoMe, with the support of more focused analyses, led to
304 the identification of numerous candidate o-aminophenol oxidases in stramenopiles.

305 By exploring the group of stramenopiles in the final GSMN dendrogram (Fig. 5B), we noticed
306 that it grouped with the small unicellular alga *G. theta*, which belongs to the cryptophytes, usually
307 grouping with the archeplastids or the haptophytes. Its plastid is derived from a secondary en-
308 dosymbiosis event with a red alga (Curtis et al., 2012). The phylogenetic position of cryptophytes
309 is unclear, but they have been suggested to be phylogenetically separate from haptophytes closer
310 to the green algae lineage (Burki et al., 2012). To further examine the position of *G. theta* in
311 our metabolic trees, we analyzed the presence/absence matrix of metabolic reactions to determine
312 which of them most clearly linked *G. theta* to each of the three groups in question (stramenopiles,
313 archeplastids, haptophytes). To this means we focused on reactions that distinguished at least two
314 of these groups, i.e. that were present in at least 80% of the networks of at least one group, and
315 absent from at least one other group (Supplemental Table S11). A total of 216 reactions met this
316 criterion, 109 of which were found in *G. theta* and 107 were absent. We found that the network of
317 *G. theta* shared the presence or absence of a similar number of distinctive reactions with all three
318 groups: 120 with stramenopiles, 112 with haptophytes, and 101 with archeplastids.

319 Next, we examined the metabolic pathways represented by the reactions that associated *G. theta*
320 with the three groups, focusing on pathways that were more than 50% complete. The metabolic
321 networks showed, for instance, that *G. theta*, (i) like haptophytes in our dataset, possess parts of
322 the mitochondrial L-carnitine shuttle pathway, (ii) like the stramenopiles, comprises the complete
323 pathway of glycine betaine synthesis, and, (iii) like terrestrial plants, can synthesize carnosine. We
324 also manually examined the genes associated with these reactions, and found that in all cases,
325 their sequences differed strongly from other sequences in the database, and could not be clearly
326 associated with either archeplastids, stramenopiles, or haptophytes (see Supplemental Table S12).

327 These examples underline the fact that cryptophytes diverged from the other lineages early
328 in the history of eukaryotes and support the hypothesis that the metabolic capacities of extant
329 cryptophytes might reflect adaptation to their specific environment more clearly than their ancient
330 evolutionary history.

331 Discussion

332 Numerous sequencing projects and available annotation approaches generate heterogeneously an-
333 notated data. There is currently a need to homogenize annotations to make them comparable for
334 wider scale studies. In this work we introduced a method to automatically homogenize functional
335 predictions across heterogeneously-annotated genomes for large-scale metabolism comparisons be-
336 tween species across the tree of life. We illustrated how the tool can be applied both to prokaryotes
337 and eukaryotes, even with high levels of annotation degradation.

338 Accounting for existing annotations in the inference of homogenized GSMNs

339 Automatic inference of single species GSMNs is now routinely achieved, especially for prokaryotic
340 species, and is often systematically performed for multiple genomes. With such data at hand,
341 one may compare the predicted metabolism among related species from a given clade and subse-
342 quently identify metabolic specificities or putative functional interactions in microbial communities
343 (Machado et al., 2018; Frioux et al., 2018). Such applications require consistent genome quality and

344 similar data treatment (genome annotation, metabolic network reconstruction) to minimize biases
345 in predictions. However, ensuring the latter is complex for eukaryotic genomes, as their enzymatic
346 functions are difficult to characterize automatically and they often need expert annotation. More-
347 over, annotation efforts can greatly vary between genomes, resulting in heterogeneous annotation
348 and metabolic prediction quality. As the automatization of both (meta)genome reconstruction and
349 annotation is now routinely applied, it is likely that efforts toward manual annotation will decline.
350 However, we believe the need to manually curate annotations will remain (Karimi et al., 2021). In
351 addition, AuCoMe could also be used to homogenize annotations in several genome versions of the
352 same species, or to reconcile several annotations performed on the same genome.

353 We have shown above that the performance of AuCoMe is superior to or on par with other com-
354 monly used reconstruction pipelines, notably GapSeq, ModelSEED, and CarveMe. The originality
355 of our metabolic inference method resides in the possibility to account for, and preserve, avail-
356 able expert genome annotations. Not considering the genome annotations performed by specialists
357 may lead to the omission of unique metabolic functions that are not well described in reference
358 databases. On the other hand, comparing metabolic networks built from well-curated annotations
359 to those built from poorly or automatically-annotated genomes will result in biases. In such cases,
360 real metabolic differences between species cannot be distinguished from missing annotations in
361 some genomes. AuCoMe constitutes a solution to such challenges through the propagation of ex-
362 pert annotations to less characterized genomes in the process of metabolic network reconstruction.
363 By accounting for possibly missing functional but also structural annotations in the input genomes,
364 the resulting metabolic networks are homogeneous and can therefore be directly compared in both
365 prokaryotes and eukaryotes.

366 **Method limitations and improvements**

367 AuCoMe incorporates several strategies to optimize the method’s selectivity and sensitivity. To-
368 gether these strategies collectively achieve comparable GSMN reconstruction with two objectives:
369 having comparisons as homogeneous as possible given the initial heterogeneity and incompleteness

370 of databases, and thus identifying errors that can be corrected during further analysis.

371 A first limitation is illustrated by the comparison of AuCoMe reconstructions to the EcoCyc
372 database considered as ground truth in our experiment. We observed that the GSMN automatically
373 reconstructed from the reference genome substantially differs from the database. Extensive and
374 systematic manual curation has been performed on this database since its creation in 1998 and we
375 hypothesize that these efforts have not been all translated in the *E. coli* K-12 MG1655 annotations.
376 As a result, several reactions were systematically missing from the automatic inferences provided
377 by AuCoMe. This example illustrates the role of curation in producing high quality models. The
378 homogenization of metabolic inference proposed by AuCoMe does not aim at replacing this step
379 but rather enable an unbiased metabolic comparison between species.

380 Running AuCoMe on the bacterial dataset highlighted the impact of a single highly-annotated
381 genome on metabolic inference. This dataset included a single well-annotated reference genome
382 of the *E. coli* K-12 MG1655 strain, which caused a number of reactions initially propagated by
383 orthology from the *E. coli* K-12 MG1655 genome to others to be discarded by the AuCoMe fil-
384 ter. Reasoning on ortholog clusters, the filter implies that several congruent genome sources are
385 mandatory to confidently achieve an annotation propagation. While the relevance of the filter was
386 demonstrated on the algal dataset by avoiding the propagation of annotations related to photo-
387 synthesis to non-photosynthetic organisms, it may be too stringent in some applications. Several
388 improvements of the filtering approach could be devised. For example, the structural annotation
389 step could be improved: the annotation of pseudogenes in *Shigella* species would have been avoided
390 by considering the annotations as pseudogenes available for the identified loci. More generally, in
391 addition to the difficulties of automatically estimating protein homology, the link between orthol-
392 ogy and conservation of function is still a matter of active investigation and methodological debate
393 (Stambouliau et al., 2020; Begum et al., 2021).

394 Finally, we want to emphasize that our attempts to limit the inference of false positive reactions
395 also directed the choice of method for the initial draft metabolic inference. We used Pathway
396 Tools because of its several advantages such as the capacity to work with eukaryotic genomes, the

397 suitability for parallel computing (Belcour et al., 2020a), and the possibility to limit gap-filling of
398 metabolic networks. However, metabolic pathway completion performed by Pathway Tools does
399 not systematically extend to ensuring the production of biomass. Pathway Tools was therefore
400 adapted to our objective of avoiding to go beyond the strict interpretation of genome annotations.
401 This goal was fulfilled, as attested by the benchmark shown in Supplemental Fig. S4 which confirms
402 that AuCoMe GSMNs have by design no reaction lacking gene association.

403 A typical use for genome-scale metabolic networks is their simulation, generally with flux-based
404 approaches. As AuCoMe performs an homogenization step on GSMNs but does not provide de-novo
405 annotation, using AuCoMe without further curation might lead to missing reactions in organisms.
406 In addition, the complexity of eukaryotes and their strong dependency on their environment makes
407 it difficult to provide a flux-based simulation-ready gap-filled model that would minimize the risk
408 of adding false positives. For further simulation studies, GSMNs built with AuCoMe therefore
409 still need to be gap-filled and curated (Karp et al., 2018b; Latendresse and Karp, 2018). However,
410 regarding the reactions that are present in at least one GSMN reconstructed by AuCoMe, the tool
411 ensures that their absence in other organisms is true. In that sense, AuCoMe reduces the need for
412 curation.

413 **Biological insights on the comparison of metabolic networks across species**

414 **Evolution**

415 Our examples of the Calvin cycle and phycobiliprotein synthesis demonstrate that, once all steps
416 of the AuCoMe pipeline have been executed, the predicted metabolic capacities of the analyzed
417 genomes reflect the biological knowledge we have of the corresponding organisms. Our approach,
418 therefore, enables GSMNs to be compared in the light of evolutionary biology. The metabolic
419 dendrograms calculated from final AuCoMe reconstruction are mostly consistent with reference
420 species phylogeny. Indeed, numerous studies have shown that comparing GSMNs by computing a
421 metabolic distance and arranging them into a dendrogram allows clustering organisms into groups
422 close to the ones known by phylogenetic analysis. However, the position of species inside these

423 groups is often different from the one of the phylogenetic groups (Vieira et al., 2011; Bauer et al.,
424 2015; Prigent et al., 2018; Schulz and Almaas, 2020). It furthermore gives support to the hypothesis
425 of a metabolic clock based on the congruence between molecular and metabolomic divergence in
426 phytoplankton (Marcellin-Gros et al., 2020). The difference observed in the tanglegram (Fig. 5B)
427 between phylogeny and metabolic distances could be further explored. One possibility could be to
428 look at different similarity measures for the clustering. In this work, the Jaccard distance has been
429 used but other measures could be used. For example, if we consider an absence of a reaction in
430 two organisms as a similarity (to represent the loss of a function) then other measures could be
431 envisaged such as the Simple Matching Coefficient. This also opens the perspective of inferring
432 ancestral metabolic networks to better understand the dynamics of character evolution across time
433 (Psomopoulos et al., 2020).

434 **Adaptation**

435 The second aim of reconstructing comparable GSMNs is to determine to what extent metabolic
436 changes are the result of or the prerequisite for adaptation. In our study, we made a first attempt
437 at this question regarding the cryptophyte *Guillardia theta*. This species has several potentially
438 plesiomorphic metabolic traits in common with other marine lineages, that may constitute adap-
439 tations to their shared marine environment. Glycine-betaine, for instance, is known to be an
440 osmoregulator or osmoprotectant in green plants (Di Martino et al., 2003), and carnosine has been
441 proposed to function as an antioxidant in red algae (Tamura et al., 1998). Regarding carnitine, its
442 physiological significance in photosynthetic organisms is still largely unknown, but antioxidant and
443 osmolyte properties along with signaling functions have also been suggested (Jacques et al., 2018).
444 However, for now, all of this remains purely hypothetical. To dig deeper into such questions in the
445 future, we need to be able to distinguish changes that simply result from random processes such
446 as metabolic drift (Belcour et al., 2020b) from changes that have an adaptive value. Currently, we
447 envision two approaches that will help with this distinction. The first approach will be to further
448 increase the number of species and lineages included in order to identify adaptive patterns, for

449 example to among organisms occupying similar ecological niches. In phylogenomics, wide taxon
450 sampling is recognized as one of the key features for reliable comparisons (Young and Gillung,
451 2020), whereas pairwise genomic comparisons across species are generally viewed as problematic
452 (Dunn et al., 2018). Given that, as demonstrated above, phylogenetic signals in metabolism are
453 stronger than the adaptive signals we can expect, this approach would also benefit from the devel-
454 opment or adaptation of statistical models that could help detect signals of adaptation in an overall
455 noisy dataset. Such models exist, for instance, to detect selective signatures in the evolution of
456 protein-coding gene (Shapiro and Alm, 2008), but to our knowledge, have not been developed for
457 metabolic networks or presence/absence signatures of genes. The second related strategy consists
458 in focusing on phylogenetically closely related species that have only recently diverged and adapted
459 to different environments. In such cases, we anticipate that the relative importance of drift along
460 with the noise from the phylogenetic signal will be reduced due to the short evolutionary time since
461 the separation. With such datasets, we may be able to reduce the level of replication required to
462 find biologically relevant metabolic adaptations. The range of questions that could be addressed
463 with the appropriate dataset is long and includes metabolic adaptations to different environments
464 (Xu et al., 2020), food sources and domestication (Giannakou et al., 2020), multicellularity (Cock
465 et al., 2010), or even life-history transitions to endophytism (Bernard et al., 2019).

466 **Interactions**

467 Lastly, we anticipate that AuCoMe will provide new opportunities to study metabolic interactions
468 between symbiotic organisms. For example, the tentative o-aminophenol oxidase activities pointed
469 out by AuCoMe in brown algae could be involved in the protection against pathogen attacks at
470 the cell surface. Indeed, a molecular oxygen-scavenging function in the chloroplast (Constabel
471 et al., 1995) and a defense role (Gandía-Herrero et al., 2005) have been suggested for these enzymes
472 in terrestrial plants. An o-Aminophenol Oxidase *Streptomyces griseus* is known to be involved
473 in the grixazone biosynthesis, i.e. an antibiotic (Suzuki et al., 2006). Similarly, brown algal o-
474 aminophenol oxidases or tyrosinases might be involved in the production of specific antibiotics.

475 The o-aminophenol oxidase enzymes resemble laccases or tyrosinases. They can be involved in
476 catechol or pigment production by oxidation (Le Roes-Hill et al., 2009). Numerous references have
477 also shown that tyrosinases are efficiently inhibited by some phlorotannins, antioxidant compounds
478 specific to the brown algae (Kang et al., 2004; Manandhar et al., 2019) suggesting there might be
479 a regulation of polyphenol oxidation in certain conditions.

480 In the same vein, metabolic complementarity has previously been used to predict potentially
481 beneficial metabolic interaction between a host and its associated microbiome (Frioux et al., 2018),
482 and to successfully predict metabolic traits of the communities (Burgunter-Delamare et al., 2020).
483 These studies have, so far, examined large numbers of symbionts (all sequenced and annotated with
484 identical pipelines), but usually consider one specific host whose metabolic network was manually
485 curated. With AuCoMe, these previous efforts could be expanded to incorporate a range of different
486 hosts with their associated microbiota, thus facilitating the identification of common patterns in
487 host-symbiont metabolic complementarity as well as their differences in these complementarities
488 across different species and lineages. Just as for the question of adaptation, we believe this new
489 scale of comparisons enabled by tools such as AuCoMe, will enable researchers to move from the
490 study of specific examples to the identification of general trends, thus approaching the biologically
491 most relevant evolutionary constraints.

492 **Methods**

493 **Genomes and models**

494 The *bacterial dataset* includes the 29 bacterial *Escherichia coli* and *Shigella* strains studied in
495 (Vieira et al., 2011), downloaded from public databases (see Supplemental Table S1).

496 The *fungus dataset* includes 74 fungal genomes which were selected according to Wang et al.
497 (2009) as representative of the fungal diversity, together with 3 outgroup genomes: *Caenorhabdi-*
498 *tis elegans*, *Drosophila melanogaster*, and *Monosiga brevicollis*. All proteomes and genomes were
499 downloaded from the NCBI Assembly Database (Kitts et al., 2016). See Supplemental Table S2.

500 The *algal dataset* contains 36 algal genomes selected to represent a wide diversity of photosyn-

501 thetic eukaryotes and downloaded from public databases. The dataset includes 16 Viridiplantae
502 (green algae), 5 Phaeophyceae (brown algae), 5 Rhodophyceae (red algae), 4 diatoms, 3 hap-
503 tophytes, 1 cryptophyte (*Guillardia theta*), 1 Eustigmatophyceae (*Nannochloropsis gaditana*), 1
504 Glaucophyceae (*Cyanophora paradoxa*). The genomes of *C. elegans* (Witting et al., 2018), *Muco*
505 *circinelloides* (Vongsangnak et al., 2016), *N. crassa* (Dreyfuss et al., 2013), and *S. cerevisiae* (Lu
506 et al., 2019) were selected as outgroup genomes (see Supplemental Table S3).

507 Each annotated genome of the datasets was curated manually in order to make it compati-
508 ble with Pathway Tools v23.5. Curated genomes are available at [https://zenodo.org/record/](https://zenodo.org/record/7752449#.ZBh0pi0ZN-E)
509 [7752449#.ZBh0pi0ZN-E](https://zenodo.org/record/7752449#.ZBh0pi0ZN-E).

510 **AuCoMe, a method to reconstruct genome-scale metabolic networks homoge-** 511 **nized across related species**

512 AuCoMe is a Python package implementing a pipeline whose steps are described in Fig. 1. The
513 method aims at producing homogenized genome scale metabolic networks (GSMNs) for a set of
514 heterogeneously-annotated genomes containing closely related or outlier species of a taxonomic
515 group. AuCoMe takes as input GenBank files containing the genome sequences, the structural
516 annotation of the genomes (gene and protein locations), the functional annotations (especially with
517 GO terms and EC numbers) and the protein sequences. The output of AuCoMe is a set of GSMNs,
518 provided in SBML and PADMET formats (Hucka et al., 2018; Aite et al., 2018). AuCoMe also
519 produces a global report describing the sets of reactions added at all steps of the pipeline. The
520 global panmetabolism, which is the complete family of metabolic reactions included in at least one
521 GSMN of the set of genomes, is described in a tabulated file.

522 At **the initialization step** the command `aucome init` creates a template folder in which the
523 user puts the input GenBank files.

524 The `aucome reconstruction` command runs **the draft reconstruction step**, which consists
525 in reconstructing draft GSMNs according to the set of available genome annotations. During this
526 step, the pipeline first checks the input GenBank files using Biopython (Cock et al., 2009). Then

527 using the mpwt package (Belcour et al., 2020a), AuCoMe launches parallel processes of the Patho-
528 Logic algorithm of Pathway Tools (Karp et al., 2019). Pathway Tools creates Pathway/Genome
529 Databases (PGDB) for all genomes. The resulting PGDBs are converted into PADMET and SBML
530 files (Hucka et al., 2003, 2018) using the PADMet package (Aite et al., 2018). During this con-
531 version, pathway hole reactions (reactions predicted by Pathway Tools for which no enzymes were
532 detected in the genomes) are removed as they are not associated with a gene and are not sponta-
533 neous reactions. For example, in Fig. 1A, the draft reconstruction step generates 6 GPRs in total
534 for the 3 considered genomes.

535 The `aucome orthology` command runs the **orthology propagation step**, which complements
536 the previous GSMNs with GPRs associations whose genes are predicted to be orthologs to genes
537 from GPR relations of other GSMNs of the dataset (Fig. 1B). To that purpose, the pipeline relies
538 on OrthoFinder (Emms and Kelly, 2015, 2019) for the inference of *orthologs* defined as clusters
539 of homologous proteins shared across species. For each pair of orthologous genes shared between
540 two species, the pipeline checks whether one of the genes is associated with an existing GPR
541 association. If so, a putative GPR association with the orthologous gene is added to the GSMN. At
542 the end of the analysis of all genomes, a robustness score is calculated for assessing the confidence
543 of each putative GPR association based on the number of annotated GPRs associations between
544 the orthologs (see below). Non-robust GPR associations are not integrated in the final GSMNs. In
545 the example shown in Fig. 1B, applying the robustness criteria leads to generating a putative new
546 GPR association in the GSMN 2 (see the green orthogroup). In this example, the pipeline does
547 not validate the GPR association related to the blue orthogroup because of insufficient annotation
548 support.

549 The `aucome structural` command runs the **structural verification step** to identify GPRs
550 associated with missing structural annotations of the input genomes. This pipeline step comple-
551 ments GSMNs with GPR associations from other GSMNs according to protein-against-genome
552 alignment criteria. This enables the identification of reactions which are associated with gene se-
553 quences absent from the initial structural annotations of the input genomes. A pairwise comparison

554 of the reactions in the GSMNs produced during the previous step is performed (Fig. 1C). In this
555 comparison, if a reaction is missing in an organism, a structural verification will be performed.
556 For each protein sequence associated with a GPR relation in a GSMN, a TBLASTN (Altschul
557 et al., 1990; Camacho et al., 2009) with Biopython (Cock et al., 2009) is performed against the
558 other genome. If a match (evalue<1e-20) is found, the gene prediction tool Exonerate (Slater and
559 Birney, 2005) is run on the region linked to the best match (region +- 10 KB). If Exonerate finds
560 a match, then the reaction associated with the protein sequence is added. In Fig. 1C, one reaction
561 is added to the GSMN 2.

562 The command `aucome spontaneous` runs the **spontaneous completion step** to fill metabolic
563 pathways with spontaneous reactions, in order to complement each GSMN obtained after the
564 structural-completion step with spontaneous reactions. For each pathway of the MetaCyc database
565 (Caspi et al., 2020) that was incomplete in a GSMN, AuCoMe checks whether adding spontaneous
566 reactions could complete the pathway. When this is the case, the spontaneous reaction is added to
567 the GSMN. In Fig. 1D, two spontaneous reactions are added to the GSMN 1 and GSMN 3. Then
568 the final PADMET and SBML files are created for each studied organism.

569 **Robustness criteria for GPR association predicted by orthology**

570 The robustness score of GPR associations of the pan-metabolic network after the orthology propa-
571 gation was defined as illustrated in Algorithm 1 and detailed in the following. We denote by $org(g)$
572 the organism of a gene g . For every pair of genes $g1, g2$ of two different organisms, we denote
573 $orth(g1, g2) = 1$ if the genes are predicted to be orthologs. We denote by $association(r, g) = 1$
574 a GPR association between a reaction r and a gene g which is predicted by the AuCoMe al-
575 gorithm. When the gene-association is predicted by the draft reconstruction step, we denote
576 $annot_type(r, g) = 1$ (and zero otherwise). When the gene-association is predicted according to
577 orthology criteria, we denote $ortho_type(r, g) = 1$ (and zero otherwise).

578 Let us consider now a reaction r of the pan-metabolic network. We denote by $N_org(r)$ the
579 number of organisms for which the reaction r has been associated with a GPR relationship with

580 any gene g : $N_org(r) = \#\{org(g), association(r, g) = 1\}$ (L2, Alg. 1). For every gene g with
581 $annot_type(r, g) = 1$, we denote by $N_prop(r, g)$ the number of organisms different from $org(g)$ the
582 GPR association between r and g has been propagated to according to an orthology relation with the
583 gene g $N_prop(r, g) = \#\{org(g1), \exists g1 \text{ s.t. } org(g1) \neq org(g), orth(g, g1) = 1, association(r, g1) =$
584 $1\}$. The GPR association between r and g is considered robust: $robust(r, g) = 1$ as long as
585 $annot_type(r, g) = 1$.

586 The robustness assessment of a GPR between r and g propagated by orthology (L7, Alg. 1)
587 distinguishes two scenarios. In the first scenario g belongs to an orthology cluster which is supported
588 by at least two annotations. Formally this means that there exist two genes $g1$ to $g2$, both orthologs
589 to g , such that $annot_type(r, g1) = 1$ and $annot_type(r, g2) = 1$. The presence of these genes leads
590 us to consider g robustly associated with r (L8-9, Alg. 1).

591 In the second scenario the GPR association between r and g was propagated from a unique
592 gene $g1$ with $annot_type(r, g1) = 1$ in the orthology cluster (L11, Alg. 1). For these genes our
593 strategy is to be as stringent as possible and we introduce a robustness criterion to reduce the risk
594 of propagating false-positive reactions. The GPR association is considered robust if the number
595 of organisms to which the reaction is propagated according to the annotation of $g1$ remains low
596 with respect to the total number of considered organisms. More precisely, $robust(r, g) = 1$ if
597 $N_prop(r, g1) \leq \lceil robust_func(N_org(r) - 1) \times (N_org(r) - 1) \rceil$ (L12-13, Alg. 1). The robustness
598 function $robust_func^{(t)}(x) = \min(1, \frac{1}{x} \max(\lceil tx \rceil, \lceil \frac{5}{x} \rceil))$ was chosen such that it is 1 for low values
599 of N_org , and then decreases to a threshold value (by default $t = 0.05$) for large values of N_org
600 (see a plot in Supplemental Fig S14).

601 Altogether, the robustness criterion removes orthology predictions for GPR associations that
602 are supported by a unique gene annotation and propagated to a large number of organisms. A toy
603 example of the application of the algorithm is detailed in Supplemental Methods Section and Fig
604 S15).

Algorithm 1 Robustness criterion algorithm

```
1: for  $r$  in panmetabolism do
2:    $N\_org(r) \leftarrow \#\{org(g), \exists g1 \text{ s.t. } association(r, g) = 1\}$   $\triangleright$  Number of organisms with GPRs relations to  $r$ 
3:   for all genes  $g$  s.t.  $annot\_type(r, g) = 1$  do
4:      $robust(r, g) = 1$ 
5:      $N\_prop(r, g) \leftarrow \#\{org(g1), \exists g1 \text{ s.t. } org(g1) \neq org(g), orth(g, g1) = 1, association(r, g1) = 1\}$ .  $\triangleright$  Number
of organisms to which the GPR has been propagated to
6:   end for
7:   for all genes  $g$  s.t.  $annot\_type(r, g) = 0$  and  $orth\_type(r, g) = 1$  do  $\triangleright$  Only other way to add the reaction
8:     if  $\exists g1, g2$  s.t.  $orth(g, g1) = orth(g, g2) = 1$  and  $annot\_type(r, g1) = annot\_type(r, g2) = 1$  then
9:        $robust(r, g) = 1$   $\triangleright$  At least two annotations support the GPR relation
10:    else  $\triangleright$  Prevent the propagation of an isolated annotation to too many organisms
11:       $g1 \leftarrow$  unique gene s.t.  $orth(g, g1) = 1$  and  $annot\_type(r, g1) = 1$ 
12:      if  $N\_prop(r, g1) \leq robust\_func(N\_org(r) - 1) \times (N\_org(r) - 1)$  then
13:         $robust(r, g) = 1$ 
14:      else
15:         $robust(r, g) = 0$ 
16:      end if
17:    end if
18:  end for
19: end for
```

605 **Validation of AuCoMe predictions**

606 A first experiment was performed on the bacterial dataset, for which we reconstructed the metabolic
607 networks (29 bacteria containing strains of *Escherichia coli*) using CarveMe 1.5.1 (Machado et al.,
608 2018) with default parameters, gapseq 1.2 (Zimmermann et al., 2021) with default parameters
609 and ModelSEED with Kbase. For the latter, we first imported the genomes and annotated them
610 with 'Bulk Annotate Genomes/Assemblies with RASTtk - v1.073' (Aziz et al., 2008; Overbeek
611 et al., 2014; Brettin et al., 2015) and then reconstructed the models with 'Build Multiple Metabolic
612 Models' 2.0.0 (Henry et al., 2010). We compared the ECs predicted by these methods to the
613 ones contained in a reference EC catalog for *E. coli* K-12 MG1655 created from 4 databases
614 (KEGG, EcoCyc, ModelSEED and BiGG). For more information on the reference EC catalog, see
615 Supplementary file (section Methods).

616 A second comparison was made on the eukaryotes and especially the fungal dataset (using five
617 organisms: *Laccaria bicolor*, *Neurospora crassa*, *Rhizopus oryzae*, *Saccharomyces cerevisiae* S288C
618 and *Schizosaccharomyces pombe*). We used Kbase (Arkin et al., 2018) and gapseq 1.2 (Zimmermann
619 et al., 2021). The genomes were imported into Kbase and the metabolic networks were reconstructed
620 with 'Build Fungal Model' 1.0.0 (with gap-filling). We also used gapseq to predict the metabolic

621 pathways present in an organism using its *find* module associated with the option '-t Fungi'. We
622 did not use CarveMe as it has been developed for Bacteria or Archaea (Capela et al., 2022). We
623 compared the completion rate of metabolic pathways predicted by AuCoMe and gapseq. Then for
624 *Saccharomyces cerevisiae S288C*, we used the reference network YeastCyc to estimate the quality
625 of the pathways predicted by both gapseq and AuCoMe.

626 In a third evaluation, one hundred random GPR associations were randomly selected and ex-
627 amined across the metabolic networks generated by AuCoMe for the algal dataset. Among them,
628 50 reactions that were predicted to be present and 50 reactions that were predicted to be absent in
629 the metabolic networks. Regarding the former, their first associated gene was manually annotated
630 based on reciprocal BLAST searches against UniProt (Bateman et al., 2021) and the presence
631 of conserved domains and the result of this manual annotation was compared to the predicted
632 metabolic reaction. For absent reactions, we searched for characterized proteins known to catalyze
633 the reaction in question, and then performed reciprocal BLASTP searches with the corresponding
634 algal proteome.

635 A fourth experiment was performed to analyse the results of the orthology propagation and the
636 robustness filter. DeepEC (version 0.4.0) (Ryu et al., 2019) was applied both to fungal and algal
637 protein sequences. This tool predicts EC numbers for protein sequences. We extracted the EC
638 numbers of reactions for which at least one GPR association was predicted according to orthology
639 propagation for all reactions of the fungal and the algal datasets. For each EC number, we extracted
640 the protein sequences associated with the considered reaction in the GSMNs, and we used DeepEC
641 to infer an EC number for these proteins. Then we compared the EC number found by DeepEC
642 (if found) to the EC number linked to the reaction by the pipeline.

643 Finally, the complementarity between the orthology propagation step (second step) and the
644 structural verification step (third step) was assessed using the *E. coli* K-12 MG1655 genome mod-
645 ified to generate replicates with randomly degraded annotations associated with GPR of the non
646 degraded *E. coli* K-12 MG1655 GSMN. Two degradation types were simulated, (i) a degradation
647 of the functional annotations of the genes, where all the annotations like GO Terms, EC numbers,

648 gene names, etc. associated with a reaction were removed, and (ii) a degradation of the struc-
649 tural annotation of the genes, where gene positions and functional annotations were removed from
650 the genome annotations. A third type of replicate was considered including the degradation of
651 both structural and functional annotations. Replicates with increasing percentages of degraded
652 annotations were generated for each of the three types of degradation. Details on the degradation
653 algorithm are shown in the Supplementary file (section Methods). Furthermore the taxonomic ID
654 associated with the *E. coli* K-12 MG1655 genome was degraded to *cellular organism*, to focus on
655 the impact of genome annotations on GSMN reconstructions by AuCoMe, rather than on the effect
656 of the automatic completion by the EcoCyc source performed by Pathway Tools when analyzing
657 *E. coli* K-12 MG1655 . Each degraded replicate was associated with the 28 other *E. coli* and
658 *Shigella* genomes, generating 31 synthetic bacterial datasets, plus the dataset with non-degraded
659 *E. coli* K-12 MG1655 genome, which was called dataset 0. Their characteristics are detailed in
660 Supplemental Table S4. For each *E. coli* K-12 MG1655 replicate in a dataset, AuCoMe produced
661 a GSMN, which was compared to EcoCyc, considered as ground truth (Karp et al., 2002, 2018a;
662 Keseler et al., 2021). For more information on the computation of the F-measure, see Supplemental
663 file (section Methods).

664 **Phylogenetic analysis of the brown algal o-aminophenol oxidases**

665 A dataset of 193 protein sequences was constructed using the closest homologs of the *S. japonica*
666 o-aminophenol oxidase (SJ09941) in brown algae and extended to more distant sequences present
667 in other organisms. Sequences were submitted to NGPhylogeny.fr via the "A la carte" (Lemoine
668 et al., 2019) pipeline. The alignment was carried out by MAFFT (Katoh and Standley, 2013)
669 using default parameters and automatically cleaned with trimAl (Capella-Gutiérrez et al., 2009)
670 to obtain 372 informative positions. Then a maximum likelihood phylogenetic reconstruction was
671 carried out using default parameters of the PhyML-SMS tool (Guindon et al., 2010; Lefort et al.,
672 2017) allowing the best substitution model selection. Bootstrap analysis (Lemoine et al., 2018)
673 with 100 replicates was used to provide estimates for the phylogenetic tree topology. The Newick

674 file (Junier and Zdobnov, 2010) was further formatted by MEGA v10.1.1 (Tamura et al., 2021) to
675 obtain the simplified dendrogram (see Supplemental Fig. S13).

676 **Supplemental Files and Software availability**

677 **Supplemental Files**

678 **Supplemental File** The supplemental file contains the description of the datasets, additional
679 details on the results on running times of the AuCoMe pipeline, the three panels of B, C, D, of
680 Fig. 2, a detailed comparison with gapseq, ModelSEED and CarveMe on bacterial and fungal
681 datasets (if it is feasible). It also includes information about validation of filtering steps and GPR
682 associations, validation of EC numbers with deep-learning approaches, and two relevant biological
683 analyses: to two pathways, to the consistency between AuCoMe GSMNs and species phylogeny.
684 Moreover, it contains methodological details on the robustness criteria applied to a toy example,
685 on the comparison to EcoCyc, and on the degradation of *E. coli* K-12 MG1655 genome to generate
686 32 synthetic datasets. It also includes a description of the Zenodo archive.

687 **Additional file** The associated archive contains analyses (all tabulated files used to create the
688 figures and results of the paper), the datasets on which AuCoMe was run: the bacterial, fungal,
689 and algal datasets, and the 32 synthetic datasets, which contain an *E. coli* K-12 MG1655 genome
690 to which various degradations were applied, together with 28 other bacterial genomes. It contains
691 a version of AuCoMe, PADMet source code, and the scripts used to run some figures. It is available
692 at <https://zenodo.org/record/7752449#.ZBh0pi0ZN-E>.

693 **Software availability**

694 AuCoMe is a Python package under GPL-3.0 license, available through the Python Package Index
695 at <https://pypi.org/project/aucome>. The source code and the complete documentation are
696 freely available at <https://github.com/AuReMe/aucome> and as a supplementary zip file.

697 Running AuCoMe on the datasets studied in the paper required as dependencies BLAST v2.6.0

698 (Altschul et al., 1990), Diamond v0.9.35 (Buchfink et al., 2015), Exonerate v2.2.0 (Slater and
699 Birney, 2005), FastME v2.1.15 (Lefort et al., 2015), MCL (Enright et al., 2002), MMseqs2 v11-
700 e1a1c (Steinegger and Söding, 2017), OrthoFinder v2.3.3 (Emms and Kelly, 2015, 2019), Pathway
701 Tools v23.5 (Karp et al., 2019). The following Python packages are needed to install AuCoMe:
702 Matplotlib, mpwt v0.6.3 (Belcour et al., 2020a), padmet v5.0.1 (Aite et al., 2018), rpy2 v3.0.5,
703 seaborn, supervenn, and tzlocal. The pvclust R package is also required.

704 A docker or a singularity container can be created and enriched according to the dockerfile
705 available on <https://github.com/AuReMe/aucome/blob/master/recipes/Dockerfile>.

706 Acknowledgements

707 We acknowledge the GenOuest bioinformatics core facility <https://www.genouest.org> for pro-
708 viding the computing infrastructure. We also thank Erwan Corre (ABiMS Platform) and Pauline
709 Hamon-Giraud for fruitful discussions. This work benefited from the support of the French Gov-
710 ernment via the National Research Agency investment expenditure program IDEALG (ANR-10-
711 BTBR-04) and from Région Bretagne via the grant «SAD 2016 - METALG (9673)».

712 Author Contributions

713 **AB**: Conceptualization, Data curation, Methodology, Formal Analysis, Software, Validation, Visu-
714 alization, Writing – original draft, Writing – review & editing. **JG**: Data curation, Formal Analysis,
715 Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
716 **MA**: Conceptualization, Data curation, Methodology, Software. **LD**: Formal Analysis, Validation,
717 Writing – original draft, Writing – review & editing. **JC**: Formal Analysis, Validation, Writing –
718 review & editing. **CF**: Methodology, Software, Visualization, Writing – original draft, Writing –
719 review & editing. **CL**: Funding acquisition, Writing – original draft, Writing – review & editing.
720 **SD**: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Vali-
721 dation, Writing – original draft, Writing – review & editing. **SB**: Conceptualization, Methodology,
722 Writing – original draft, Writing – review & editing. **GVM**: Data curation, Formal Analysis,

723 Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review &
724 editing. **AS**: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Supervision,
725 Writing – original draft, Writing – review & editing.

726 **Conflict of interest**

727 The authors declare no conflict of interest.

728 **References**

729 Aite M, Chevallier M, Frioux C, Trottier C, Got J, Cortés MP, Mendoza SN, Carrier G, Dameron
730 O, Guillaudeau N, et al.. 2018. Traceability, reproducibility and wiki-exploration for “à-la-carte”
731 reconstructions of genome-scale metabolic models. *PLoS Computational Biology* **14**: e1006146.
732 DOI:10.1371/journal.pcbi.1006146.

733 Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search
734 tool. *Journal of Molecular Biology* **215**: 403–410. DOI:10.1016/S0022-2836(05)80360-2.

735 Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez
736 F, Canon S, et al.. 2018. KBase: The united states department of energy systems biology
737 knowledgebase. *Nature Biotechnology* **36**: 566–569. DOI:10.1038/nbt.4163.

738 Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass
739 EM, Kubal M, et al.. 2008. The RAST server: Rapid annotations using subsystems technology.
740 *BMC Genomics* **9**: 75.

741 Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett
742 EH, Britto R, Bursteinas B, et al.. 2021. UniProt: the universal protein knowledgebase in 2021.
743 *Nucleic Acids Res* **49**: D480–D489. DOI:10.1093/nar/gkaa1100.

744 Bauer E, Laczny CC, Magnúsdóttir S, Wilmes P, and Thiele I. 2015. Phenotypic differentiation of

745 gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome* **3**: 55.
746 DOI:10.1186/s40168-015-0121-6.

747 Begum T, Serrano-Serrano ML, and Robinson-Rechavi M. 2021. Performance of a phylogenetic
748 independent contrast method and an improved pairwise comparison under different scenarios of
749 trait evolution after speciation and duplication. *Methods in Ecology and Evolution* **12**: 1875–
750 1887. DOI:10.1111/2041-210x.13680.

751 Belcour A, Frioux C, Aite M, Bretaudeau A, Hildebrand F, and Siegel A. 2020a. Metage2metabo,
752 microbiota-scale metabolic complementarity for the identification of key species. *eLife* **9**: e61968.
753 DOI:10.7554/eLife.61968.

754 Belcour A, Girard J, Aite M, Delage L, Trottier C, Marteau C, Leroux C, Dittami SM, Sauleau P,
755 Corre E, et al.. 2020b. Inferring biochemical reactions and metabolite structures to understand
756 metabolic pathway drift. *iScience* **23**. DOI:10.1016/j.isci.2020.100849.

757 Bernard MS, Strittmatter M, Murúa P, Heesch S, Cho GY, Leblanc C, and Peters AF. 2019.
758 Diversity, biogeography and host specificity of kelp endophytes with a focus on the genera *Lam-*
759 *inarionema* and *Laminariocolax* (Ectocarpales, Phaeophyceae). *European Journal of Phycology*
760 **54**: 39–51. DOI:10.1080/09670262.2018.1502816.

761 Bernstein DB, Sulheim S, Almaas E, and Segrè D. 2021. Addressing uncertainty in genome-scale
762 metabolic model reconstruction and analysis. *Genome Biology* **22**. DOI:10.1186/s13059-021-
763 02289-z.

764 Bhattacharya D, Yoon HS, and Hackett JD. 2004. Photosynthetic eukaryotes unite: endosymbiosis
765 connects the dots. *Bioessays* **26**: 50–60. DOI:10.1002/bies.10376.

766 Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez
767 J, Carey LB, and Albà MM. 2021. Uncovering de novo gene birth in yeast using deep transcrip-
768 tomics. *Nat Commun* **12**: 604. DOI:10.1038/s41467-021-20911-3.

769 Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello
770 B, Pusch GD, et al.. 2015. RASTtk: a modular and extensible implementation of the RAST
771 algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific*
772 *Reports* **5**: 8365.

773 Buchfink B, Xie C, and Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND.
774 *Nat Methods* **12**: 59–60. DOI:10.1038/nmeth.3176.

775 Burgunter-Delamare B, KleinJan H, Frioux C, Fremy E, Wagner M, Corre E, Le Salver A, Leroux
776 C, Leblanc C, Boyen C, et al.. 2020. Metabolic complementarity between a brown alga and
777 associated cultivable bacteria provide indications of beneficial interactions. *Frontiers in Marine*
778 *Science* **7**. DOI:10.3389/fmars.2020.00085.

779 Burki F, Okamoto N, Pombert JF, and Keeling PJ. 2012. The evolutionary history of haptophytes
780 and cryptophytes: phylogenomic evidence for separate origins. *Proc Biol Sci* **279**: 2246–2254.
781 DOI:10.1098/rspb.2011.2301.

782 Burki F, Roger AJ, Brown MW, and Simpson AG. 2020. The new tree of eukaryotes. *Trends in*
783 *Ecology & Evolution* **35**: 43–55. DOI:10.1016/j.tree.2019.08.008.

784 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. 2009.
785 BLAST+: architecture and applications. *BMC bioinformatics* **10**: 421. DOI:10.1186/1471-2105-
786 10-421.

787 Capela J, Lagoa D, Rodrigues R, Cunha E, Cruz F, Barbosa A, Bastos J, Lima D, Ferreira EC,
788 Rocha M, et al.. 2022. merlin, an improved framework for the reconstruction of high-quality
789 genome-scale metabolic models. *Nucleic Acids Research* **50**: 6052–6066.

790 Capella-Gutiérrez S, Silla-Martínez JM, and Gabaldón T. 2009. trimAl: a tool for auto-
791 mated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.
792 DOI:10.1093/bioinformatics/btp348.

793 Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S,
794 Subhraveti P, and Karp PD. 2020. The MetaCyc database of metabolic pathways and enzymes
795 - a 2019 update. *Nucleic Acids Research* **48**: D445–D453. DOI:10.1093/nar/gkz862.

796 Castillo S, Barth D, Arvas M, Pakula TM, Pitkänen E, Blomberg P, Seppanen-Laakso T, Nygren H,
797 Sivasiddharthan D, Penttilä M, et al.. 2016. Whole-genome metabolic model of *Trichoderma reesei*
798 built by comparative reconstruction. *Biotechnology for Biofuels* **9**: 252. DOI:10.1186/s13068-
799 016-0665-0.

800 Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F,
801 Aury JM, Badger JH, et al.. 2010. The *Ectocarpus* genome and the independent evolution of
802 multicellularity in brown algae. *Nature* **465**: 617–621. DOI:10.1038/nature09016.

803 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
804 Wilczynski B, et al.. 2009. Biopython: freely available python tools for computational molecular
805 biology and bioinformatics. *Bioinformatics* **25**: 1422–1423. DOI:10.1093/bioinformatics/btp163.

806 Constabel CP, Bergey DR, and Ryan CA. 1995. Systemin activates synthesis of wound-inducible
807 tomato leaf polyphenol oxidase via the octadecanoid defense signaling pathway. *Proc Natl Acad*
808 *Sci U S A* **92**: 407–411. DOI:10.1073/pnas.92.2.407.

809 Correia K and Mahadevan R. 2020. Pan-genome-scale network reconstruction: Harnessing phy-
810 logenomics increases the quantity and quality of metabolic models. *Biotechnology Journal* **15**:
811 1900519. DOI:10.1002/biot.201900519.

812 Cottret L, Frainay C, Chazalviel M, Cabanettes F, Gloaguen Y, Camenen E, Merlet B, Heux
813 S, Portais JC, Poupin N, et al.. 2018. MetExplore: collaborative edition and exploration of
814 metabolic networks. *Nucleic Acids Research* **46**: W495–W502. DOI:10.1093/nar/gky301.

815 Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile
816 GH, Hirakawa Y, et al.. 2012. Algal genomes reveal evolutionary mosaicism and the fate of
817 nucleomorphs. *Nature* **492**: 59–65. DOI:10.1038/nature11681.

818 Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, and Henry C. 2013. Automated genome
819 annotation and metabolic model reconstruction in the SEED and model SEED. *Methods in*
820 *Molecular Biology (Clifton, N.J.)* **985**: 17–45. DOI:10.1007/978-1-62703-299-5_2.

821 Di Martino C, Delfine S, Pizzuto R, Loreto F, and Fuggi A. 2003. Free amino acids and glycine
822 betaine in leaf osmoregulation of spinach responding to increasing salt stress. *New Phytologist*
823 **158**: 455–463. DOI:10.1046/j.1469-8137.2003.00770.x.

824 Dias O, Rocha M, Ferreira EC, and Rocha I. 2015. Reconstructing genome-scale metabolic models
825 with merlin. *Nucleic Acids Research* **43**: 3899–3910. DOI:10.1093/nar/gkv294.

826 Dreyfuss JM, Zucker JD, Hood HM, Ocasio LR, Sachs MS, and Galagan JE. 2013. Reconstruction
827 and validation of a genome-scale metabolic model for the filamentous fungus *Neurospora crassa*
828 using FARM. *PLoS Comput Biol* **9**: e1003126. DOI:10.1371/journal.pcbi.1003126.

829 Dunn CW, Zapata F, Munro C, Siebert S, and Hejnlol A. 2018. Pairwise comparisons across species
830 are problematic when analyzing functional genomic data. *Proceedings of the National Academy*
831 *of Sciences* **3**: E409–E417. DOI:10.1073/pnas.1707515115.

832 Emms DM and Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome com-
833 parisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**: 157.
834 DOI:10.1186/s13059-015-0721-2.

835 Emms DM and Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative
836 genomics. *Genome Biology* **20**: 238. DOI:10.1186/s13059-019-1832-y.

837 Enright AJ, Van Dongen S, and Ouzounis CA. 2002. An efficient algorithm for large-scale detection
838 of protein families. *Nucleic Acids Res* **30**: 1575–1584. DOI:10.1093/nar/30.7.1575.

839 Frioux C, Fremy E, Trottier C, and Siegel A. 2018. Scalable and exhaustive screening
840 of metabolic functions carried out by microbial consortia. *Bioinformatics* **34**: i934–i943.
841 DOI:10.1093/bioinformatics/bty588.

842 Gandía-Herrero F, Escribano J, and García-Carmona F. 2005. Betaxanthins as substrates for
843 tyrosinase. An approach to the role of tyrosinase in the biosynthetic pathway of betalains. *Plant*
844 *Physiol* **138**: 421–432. DOI:10.1104/pp.104.057992.

845 Giannakou K, Cotterrell M, and Delneri D. 2020. Genomic adaptation of *Saccharomyces* species
846 to industrial environments. *Frontiers in Genetics* **11**. DOI:10.3389/fgene.2020.00916.

847 Gridale CJ, Bowers LC, Didier ES, and Fast NM. 2013. Transcriptome analysis of the parasite
848 *Encephalitozoon cuniculi*: an in-depth examination of pre-mRNA splicing in a reduced eukaryote.
849 *BMC Genomics* **14**: 207. DOI:10.1186/1471-2164-14-207.

850 Gu C, Kim GB, Kim WJ, Kim HU, and Lee SY. 2019. Current status and applications of genome-
851 scale metabolic models. *Genome Biology* **121**. DOI:10.1186/s13059-019-1730-3.

852 Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, and Gascuel O. 2010. New algorithms
853 and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML
854 3.0. *Systematic Biology* **59**: 307–321. DOI:10.1093/sysbio/syq010.

855 Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, and Stevens RL. 2010. High-throughput
856 generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*
857 **28**: 977–982.

858 Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S,
859 Costenoble R, Heinemann M, et al.. 2008. A consensus yeast metabolic network reconstruction
860 obtained from a community approach to systems biology. *Nature Biotechnology* **26**: 1155–1160.
861 DOI:10.1038/nbt1492.

862 Hucka M, Bergmann FT, Dräger A, Hoops S, Keating SM, Le Novère N, Myers CJ, Olivier BG,
863 Sahle S, Schaff JC, et al.. 2018. The Systems Biology Markup Language (SBML): Language Spec-
864 ification for Level 3 Version 2 Core. *Journal of integrative bioinformatics* **15**. DOI:10.1515/jib-
865 2017-0081.

866 Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray
867 D, Cornish-Bowden A, et al.. 2003. The systems biology markup language (SBML): A medium
868 for representation and exchange of biochemical network models. *Bioinformatics* **19**: 524–531.
869 DOI:10.1093/bioinformatics/btg015.

870 Jacques F, Rippa S, and Perrin Y. 2018. Physiology of L-carnitine in plants in light of the knowledge
871 in animals and microorganisms. *Plant Sci* **274**: 432–440. DOI:10.1016/j.plantsci.2018.06.020.

872 Junier T and Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree process-
873 ing in the unix shell. *Bioinformatics* **26**: 1669–1670. DOI:10.1093/bioinformatics/btq243.

874 Kanehisa M, Furumichi M, Tanabe M, Sato Y, and Morishima K. 2017. KEGG: new perspec-
875 tives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**: D353–D361.
876 DOI:10.1093/nar/gkw1092.

877 Kanehisa M and Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
878 *Research* **28**: 27–30. DOI:10.1093/nar/28.1.27.

879 Kang HS, Kim HR, Byun DS, Son BW, Nam TJ, and Choi JS. 2004. Tyrosinase inhibitors
880 isolated from the edible brown alga *Ecklonia stolonifera*. *Arch Pharm Res* **27**: 1226–1232.
881 DOI:10.1007/BF02975886.

882 Karimi E, Geslain E, Belcour A, Frioux C, Aïte M, Siegel A, Corre E, and Dittami SM. 2021.
883 Robustness analysis of metabolic predictions in algal microbial communities based on different
884 annotation pipelines. *PeerJ* **9**: e11344. DOI:10.7717/peerj.11344.

885 Karlsen E, Schulz C, and Almaas E. 2018. Automated generation of genome-scale metabolic draft
886 reconstructions based on KEGG. *BMC Bioinformatics* **19**: 467. DOI:10.1186/s12859-018-2472-z.

887 Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, Ong WK,
888 Subhraveti P, Caspi R, Fulcher C, et al.. 2019. Pathway Tools version 23.0 update: soft-
889 ware for pathway/genome informatics and systems biology. *Briefings in Bioinformatics* Bbz104,
890 DOI:10.1093/bib/bbz104.

891 Karp PD, Ong WK, Paley S, Billington R, Caspi R, Fulcher C, Kothari A, Krummenacker
892 M, Latendresse M, Midford PE, et al.. 2018a. The EcoCyc database. *EcoSal Plus* **8**.
893 DOI:10.1128/ecosalplus.ESP-0006-2018.

894 Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bona-
895 vides C, and Gama-Castro S. 2002. The EcoCyc database. *Nucleic Acids Research* **30**: 56–58.
896 DOI:10.1093/nar/30.1.56.

897 Karp PD, Weaver D, and Latendresse M. 2018b. How accurate is automated gap filling of metabolic
898 models? *BMC Syst Biol* **12**: 73.

899 Katoh K and Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
900 Improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.
901 DOI:10.1093/molbev/mst010.

902 Keseler IM, Gama-Castro S, Mackie A, Billington R, Bonavides-Martínez C, Caspi R, Kothari A,
903 Krummenacker M, Midford PE, Muñoz-Rascado L, et al.. 2021. The EcoCyc database in 2021.
904 *Frontiers in Microbiology* **12**: 2098. DOI:10.3389/fmicb.2021.711077.

905 King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, Ebrahim A, Palsson BO, and Lewis
906 NE. 2016. BiGG models: A platform for integrating, standardizing and sharing genome-scale
907 models. *Nucleic Acids Research* **44**: D515–D522. DOI:10.1093/nar/gkv1049.

908 Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T,
909 Xiang C, Zherikov A, et al.. 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic*
910 *Acids Research* **44**: 73–80. DOI:10.1093/nar/gkv1226.

911 Latendresse M and Karp PD. 2018. Evaluation of reaction gap-filling accuracy by randomization.
912 *BMC Bioinformatics* **19**: 53.

913 Le Roes-Hill M, Goodwin C, and Burton S. 2009. Phenoxazinone synthase: what’s in a name?
914 *Trends Biotechnol* **27**: 248–258. DOI:10.1016/j.tibtech.2009.01.001.

915 Lefort V, Desper R, and Gascuel O. 2015. FastME 2.0: A Comprehensive, Accurate,
916 and Fast Distance-Based Phylogeny Inference Program. *Mol Biol Evol* **32**: 2798–2800.
917 DOI:10.1093/molbev/msv150.

918 Lefort V, Longueville JE, and Gascuel O. 2017. SMS: Smart model selection in PhyML. *Molecular*
919 *Biology and Evolution* **34**: 2422–2424. DOI:10.1093/molbev/msx149.

920 Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S, and Gascuel
921 O. 2019. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic Acids*
922 *Research* **47**: W260–W265. DOI:10.1093/nar/gkz303.

923 Lemoine F, Domelevo Entfellner JB, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, and
924 Gascuel O. 2018. Renewing Felsenstein’s phylogenetic bootstrap in the era of big data. *Nature*
925 **556**: 452–456. DOI:10.1038/s41586-018-0043-0.

926 Liaud MF, Brandt U, Scherzinger M, and Cerff R. 1997. Evolutionary origin of cryptomonad
927 microalgae: Two novel chloroplast/cytosol-specific GAPDH genes as potential markers of an-
928 cestral endosymbiont and host cell components. *Journal of Molecular Evolution* **44**: S28–S37.
929 DOI:10.1007/PL00000050.

930 Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, Marcišauskas S, Anton PM, Lappa D, Lieven C,
931 et al. 2019. A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehen-
932 sively probing cellular metabolism. *Nat Commun* **10**: 3586. DOI:10.1038/s41467-019-11581-3.

933 Machado D, Andrejev S, Tramontano M, and Patil KR. 2018. Fast automated reconstruction of
934 genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*
935 **46**: 7542–7553. DOI:10.1093/nar/gky537.

936 Manandhar B, Wagle A, Seong SH, Paudel P, Kim HR, Jung HA, and Choi JS. 2019. Phlorotan-
937 nins with Potential Anti-Tyrosinase and Antioxidant Activity Isolated from the Marine Seaweed
938 *Ecklonia stolonifera*. *Antioxidants (Basel)* **8**. DOI:10.3390/antiox8080240.

939 Marcellin-Gros R, Piganeau G, and Stien D. 2020. Metabolomic insights into marine phytoplankton
940 diversity. *Marine Drugs* **18**. DOI:10.3390/md18020078.

941 Moretti S, Tran V, Mehl F, Ibberson M, and Pagni M. 2021. MetaNetX/MNXref: unified namespace
942 for metabolites and biochemical reactions in the context of metabolic models. *Nucleic Acids*
943 *Research* **49**: D570–D574. DOI:10.1093/nar/gkaa992.

944 Nègre D, Aite M, Belcour A, Frioux C, Brillet-Guéguen L, Liu X, Bordron P, Godfroy O, Lipinska
945 AP, Leblanc C, et al.. 2019. Genome-scale metabolic networks shed light on the carotenoid
946 biosynthesis pathway in the brown algae *Saccharina japonica* and *Cladosiphon okamuranus*.
947 *Antioxidants* **8**: 564. DOI:10.3390/antiox8110564.

948 Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O’Hara RB,
949 Simpson GL, Solymos P, et al.. 2020. vegan: Community ecology package.

950 Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello
951 B, Shukla M, et al.. 2014. The SEED and the rapid annotation of microbial genomes using
952 subsystems technology (RAST). *Nucleic Acids Research* **42**: D206–214.

953 Pitkänen E, Jouhten P, Hou J, Syed MF, Blomberg P, Kludas J, Oja M, Holm L, Pent-
954 tilä M, Rousu J, et al.. 2014. Comparative genome-scale reconstruction of gapless metabolic
955 networks for present and ancestral species. *PLOS Computational Biology* **10**: e1003465.
956 DOI:10.1371/journal.pcbi.1003465.

957 Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber APM, Schwacke R, Gross J, Blouin NA,
958 Lane C, et al.. 2012. *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae
959 and plants. *Science* **335**: 843–847. DOI:10.1126/science.1213561.

960 Prigent S, Nielsen JC, Frisvad JC, and Nielsen J. 2018. Reconstruction of 24 *Penicillium* genome-
961 scale metabolic models shows diversity based on their secondary metabolism. *Biotechnology and*
962 *Bioengineering* **115**: 2604–2612. DOI:10.1002/bit.26739.

963 Psomopoulos FE, van Helden J, Médigue C, Chasapi A, and Ouzounis CA. 2020. Ancestral
964 state reconstruction of metabolic pathways across pangenome ensembles. *Microbial genomics*
965 DOI:10.1099/mgen.0.000429.

966 Rockwell NC and Lagarias JC. 2017. Ferredoxin-dependent bilin reductases in eukaryotic algae:
967 Ubiquity and diversity. *J Plant Physiol* **217**: 57–67. DOI:10.1016/j.jplph.2017.05.022.

968 Ryu JY, Kim HU, and Lee SY. 2019. Deep learning enables high-quality and high-throughput
969 prediction of enzyme commission numbers. *Proceedings of the National Academy of Sciences*
970 **116**: 13996–14001. DOI:10.1073/pnas.1821905116.

971 Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, and Karsch-Mizrachi I. 2019. GenBank.
972 *Nucleic Acids Research* **47**: D94–D99. DOI:10.1093/nar/gky989.

973 Schulz C and Almaas E. 2020. Genome-scale reconstructions to assess metabolic phylogeny and
974 organism clustering. *PLOS ONE* **15**: e0240953. DOI:10.1371/journal.pone.0240953.

975 Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, Mundy M, Chia N, Noor E,
976 Beber M, et al.. 2021. The ModelSEED biochemistry database for the integration of metabolic
977 annotations and the reconstruction, comparison and analysis of metabolic models for plants,
978 fungi and microbes. *Nucleic Acids Research* **49**: D575–D588. DOI:10.1093/nar/gkaa746.

979 Shapiro BJ and Alm EJ. 2008. Comparing patterns of natural selection across species using selective
980 signatures. *PLOS Genetics* **4**: 1–12. DOI:10.1371/journal.pgen.0040023.

981 Slater GSC and Birney E. 2005. Automated generation of heuristics for biological sequence com-
982 parison. *BMC Bioinformatics* **6**: 31. DOI:10.1186/1471-2105-6-31.

983 Stambouliau M, Guerrero RF, Hahn MW, and Radivojac P. 2020. The ortholog conjecture revis-
984 ited: the value of orthologs and paralogs in function prediction. *Bioinformatics* **36**: i219–i226.
985 DOI:10.1093/bioinformatics/btaa468.

986 Steinegger M and Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the
987 analysis of massive data sets. *Nat Biotechnol* **35**: 1026–1028. DOI:10.1038/nbt.3988.

988 Strassert JFH, Irisarri I, Williams TA, and Burki F. 2021. A molecular timescale for eukaryote
989 evolution with implications for the origin of red algal-derived plastids. *Nature Communications*
990 **12**: 1–13. DOI:10.1038/s41467-021-22044.

991 Suzuki H, Furusho Y, Higashi T, Ohnishi Y, and Horinouchi S. 2006. A novel o-aminophenol
992 oxidase responsible for formation of the phenoxazinone chromophore of grixazone. *J Biol Chem*
993 **281**: 824–833. DOI:10.1074/jbc.M505806200.

994 Suzuki R and Shimodaira H. 2006. Pvcust: an r package for assessing the uncertainty in hierarchical
995 clustering. *Bioinformatics* **22**: 1540–1542. DOI:10.1093/bioinformatics/btl117.

996 Tamura K, Stecher G, and Kumar S. 2021. MEGA11: Molecular evolutionary genetics analysis
997 version 11. *Molecular Biology and Evolution* **38**: 3022–3027. DOI:10.1093/molbev/msab120.

998 Tamura Y, Takenaka S, Sugiyama S, and Nakayama R. 1998. Occurrence of Anserine as an Antioxi-
999 dative Dipeptide in a Red Alga, *Porphyra yezoensis*. *Biosci Biotechnol Biochem* **62**: 561–563.
1000 DOI:10.1271/bbb.62.561.

1001 Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, Bazzani S, Charusanti P, Chen FC, Fleming
1002 RMT, Hsiung CA, et al.. 2011. A community effort towards a knowledge-base and mathemat-
1003 ical model of the human pathogen *Salmonella typhimurium* LT2. *BMC Systems Biology* **5**: 8.
1004 DOI:10.1186/1752-0509-5-8.

1005 Thiele I and Palsson BØ. 2010. Reconstruction annotation jamborees: a community approach to
1006 systems biology. *Molecular Systems Biology* **6**: 361. DOI:10.1038/msb.2010.15.

1007 Vieira G, Sabarly V, Bourguignon PY, Durot M, Le Fèvre F, Mornico D, Vallenet D, Bouvet O,
1008 Denamur E, Schachter V, et al.. 2011. Core and panmetabolism in *Escherichia coli*. *Journal of*
1009 *Bacteriology* **193**: 1461–1472. DOI:10.1128/JB.01192-10.

1010 Vongsangnak W, Klanchui A, Tawornsamretkit I, Tatiyaborwornchai W, Laoteng K, and Meechai
1011 A. 2016. Genome-scale metabolic modeling of *Mucor circinelloides* and comparative analysis
1012 with other oleaginous species. *Gene* **583**: 121–129. DOI:10.1016/j.gene.2016.02.028].

- 1013 Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, Nielsen J, and
1014 Kerkhoven EJ. 2018. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction
1015 and a case study on *Streptomyces coelicolor*. *PLOS Computational Biology* **14**: e1006541.
1016 DOI:10.1371/journal.pcbi.1006541.
- 1017 Wang H, Xu Z, Gao L, and Hao B. 2009. A fungal phylogeny based on 82 complete genomes using
1018 the composition vector method. *BMC Evolutionary Biology* **9**: 195. DOI:10.1186/1471-2148-9-
1019 195.
- 1020 Witting M, Hastings J, Rodriguez N, Joshi CJ, Hattwell JPN, Ebert PR, van Weeghel M, Gao AW,
1021 Wakelam MJO, Houtkooper RH, et al.. 2018. Modeling Meets Metabolomics—The WormJam
1022 Consensus Model as Basis for Metabolic Studies in the Model Organism *Caenorhabditis elegans*.
1023 *Front Mol Biosci* **5**: 96. DOI:10.3389/fmolb.2018.00096.
- 1024 Xu S, Wang J, Guo Z, He Z, and Shi S. 2020. Genomic convergence in the adaptation to extreme
1025 environments. *Plant Communications* **1**: 100117. DOI:10.1016/j.xplc.2020.100117.
- 1026 Yandell M and Ence D. 2012. A beginner’s guide to eukaryotic genome annotation. *Nature Reviews*.
1027 *Genetics* **13**: 329–342. DOI:10.1038/nrg3174.
- 1028 Young AD and Gillung JP. 2020. Phylogenomics — principles, opportunities and pitfalls of big-data
1029 phylogenetics. *Systematic Entomology* **45**: 225–247. DOI:10.1111/syen.12406.
- 1030 Zimmermann J, Kaleta C, and Waschina S. 2021. gapseq: informed prediction of bacterial
1031 metabolic pathways and reconstruction of accurate metabolic models. *Genome Biology* **22**: 81.
1032 DOI:10.1186/s13059-021-02295-1.

1033 **Figure Legends**

1034 **Figure 1 Reconstruction and homogenization of metabolisms with AuCoMe.** Starting
1035 from a dataset of partially structurally- and functionally-annotated genomes, AuCoMe’s pipeline

1036 performs the following four steps. **A. Draft reconstruction.** The reconstruction of draft genome-
1037 scale metabolic networks (GSMNs) is performed using Pathway Tools in a parallel implementation.
1038 **B. Orthology propagation.** OrthoFinder predicts orthologs by aligning protein sequences of
1039 all genomes. The robustness of orthology relationships is evaluated (see Methods), and GPRs of
1040 robust orthologs are propagated. **C. Structural verification.** The absence of a GPR in genomes
1041 is verified through pairwise alignments of the GPR-associated sequence to all genomes where it is
1042 missing. If the GPR-associated sequence is identified in other genomes, the gene is annotated, and
1043 the GPR is propagated. **D. Spontaneous completion.** Missing spontaneous reactions enabling
1044 the completion of metabolic pathways are added to the GSMNs. GSMN: Genome-scale metabolic
1045 network. OG: orthologs. GPR: Gene-protein-reaction relationship. Outlines around GPR or
1046 reactions indicate that the GPR or reaction is newly added during the corresponding step.

1047 **Figure 2 Application of the AuCoMe pipeline to the *bacterial, fungal and algal***
1048 **datasets of genomes.** The summary table (A.) depicts the number of reactions identified for
1049 each species at each step of the AuCoMe pipeline: reactions recovered by the *draft reconstruc-*
1050 *tion* step (blue), unreliable reactions predicted by orthology propagation and removed by the filter
1051 (gray), robust reactions predicted by *orthology propagation* that passed the filter (orange), addi-
1052 tional reactions predicted by the *structural verification* step (green), and *spontaneous completion*
1053 (red). The final metabolic networks encompass all these reactions except the non-reliable ones.
1054 Panels B., C., D. illustrate the results for each genome of the three datasets. The panmetabolism
1055 of each dataset (all the reactions occurring in any of the organisms after the final step of Au-
1056 CoMe) is presented in brown in B, C and D. Organisms with gray labels are outgroups. See also
1057 Supplemental Fig. S1, S2, and S3.

1058 **Figure 3 Efficiency of AuCoMe on degraded genome assemblies. (A) Number of**
1059 **reactions in *E. coli* K-12 MG1655 degraded networks after application of AuCoMe to**
1060 **32 synthetic bacterial datasets.** Each dataset consists of the genome of *E. coli* K-12 MG1655
1061 to which degradation of the functional and/or structural annotations was applied, together with 28

1062 bacterial genomes. Each vertical bar corresponds to the result on the *E. coli* K-12 MG1655 within
1063 a synthetic dataset, with the percentages of degraded annotations indicated below. The dataset
1064 labelled 0 was not subject to degradation of the *E. coli* K-12 MG1655 annotations. Three types
1065 of degradation have been performed: functional annotation degradation only (left side, datasets
1066 labelled 1 to 10), structural annotation degradation only (right side, datasets labelled 22 to 31)
1067 and both degradation types (middle, dataset labelled 11 to 21). The colored bars depict the
1068 number of reactions added to the degraded network at the different steps of the method (the
1069 blue, orange, green, grey and red color legends are as described in the figure 2). The table shown
1070 as axis indicates the dataset number and the percentage of functional or structural annotation
1071 impacted by the degradation for the corresponding column in both subfigures. **(B) F-measures**
1072 **after comparison of the GSMNs recovered for each *E. coli* K-12 MG1655 genome**
1073 **replicates with a gold-standard network.** Reactions inferred by each AuCoMe step for each
1074 replicate were compared to the gold-standard EcoCyc GSMN, allowing for the computation of F-
1075 measures. F-measures obtained after the draft reconstruction step, the orthology propagation step,
1076 or the structural verification step are shown as blue circles, orange triangles, and green crosses,
1077 respectively. The hashed rectangle from F-measure 0.79 to 1 highlights the values of F-measure,
1078 which are unreachable because 1019 reactions in EcoCyc were not present in the panmetabolism of
1079 the 29 non-degraded bacteria.

1080 Figure 4 **AuCoMe results on the Calvin cycle pathway in the algal dataset.** AuCoMe
1081 was applied to the dataset of 36 algae and 4 outgroup species (columns). Each row represents a
1082 MetaCyc reaction of the pathway, the table shows whether it is predicted by AuCoMe: blue - draft
1083 reconstruction, orange - robust reactions predicted by orthology propagation that passed the filter,
1084 green - structural verification, and gray - non-robust reactions predicted by orthology propagation
1085 and removed by the filter, black - not predicted, yellow - manually added because the MetaCyc
1086 database 23.5 does not contain a reference gene-reaction association for this reaction.

1087 Figure 5 **AuCoMe as a tool to improve taxonomic consistency of GSMNs.** A. MDS

1088 plot for GSMNs calculated with the AuCoMe draft reconstruction step or after all AuCoMe steps.
1089 In both cases, ANOSIM values are indicated below (MDS and ANOSIM were computed using the
1090 vegan package (Oksanen et al., 2020)). B. Tanglegram evaluating the taxonomic consistency of
1091 AuCoMe dendrograms based on metabolic distances using the pvclust package (Suzuki and Shi-
1092 modaira, 2006) with the Jaccard distance (right side) in comparison with reference phylogeny (left
1093 side), compiled from Strassert et al. (2021). Full lines join species for which the position in the
1094 AuCoMe dendrogram is consistent with the reference phylogeny. Dotted lines join species for which
1095 the metabolic dendrogram and the reference phylogeny diverge. A/C: Archeplastids/Cryptophytes,
1096 A: Archeplastids, R: Rodophytes, Gr: Green algae, M: Mamiellales, Chla: Chlamydomonadales,
1097 Sph: Sphaeropleales, T: Trebouxioephyceae, Chlo: Chlorellaceae, St: Streptophytes, Gl: Glauco-
1098 phytes, C: Cryptophytes, H: Haptophytes, I: Isochrytida, D: Diatoms, S: Stramenopiles, B: Brown
1099 algae, E: Ectocarpales, Ec: Ectocarpaceae, Ch: Chordariaceae, Op: Opisthochonts, F: Fungi , As:
1100 Ascomycetes.

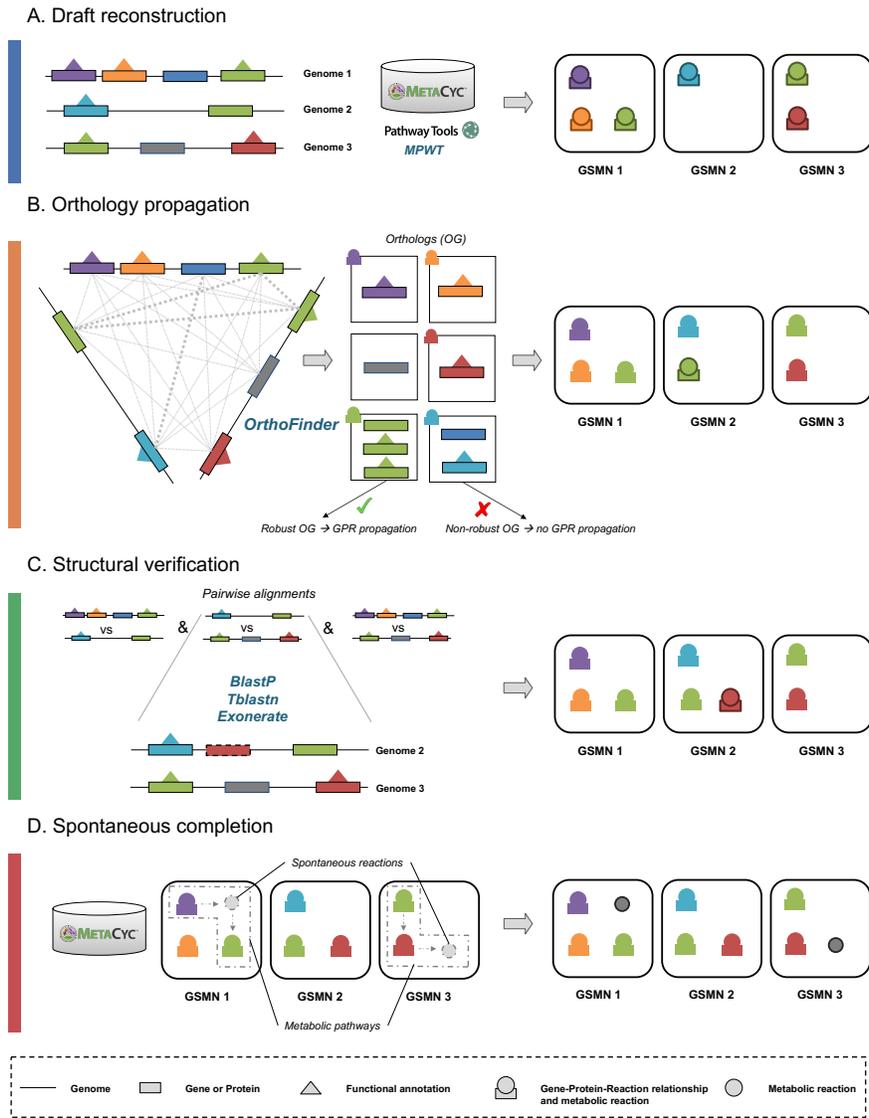
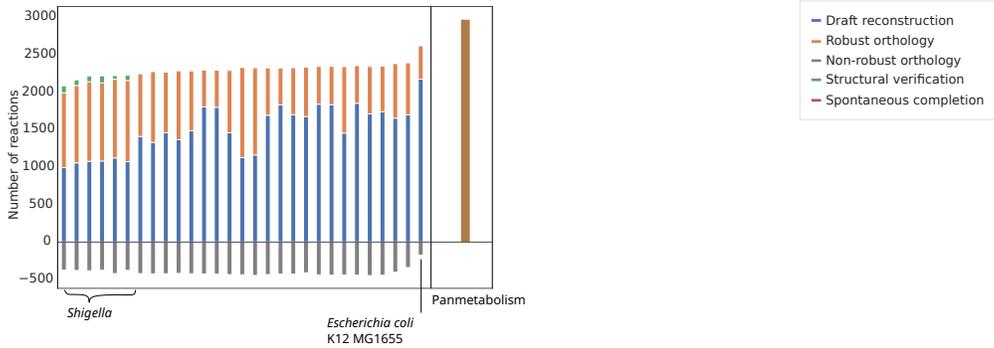


Figure 1: Reconstruction and homogenization of metabolisms with AuCoMe.

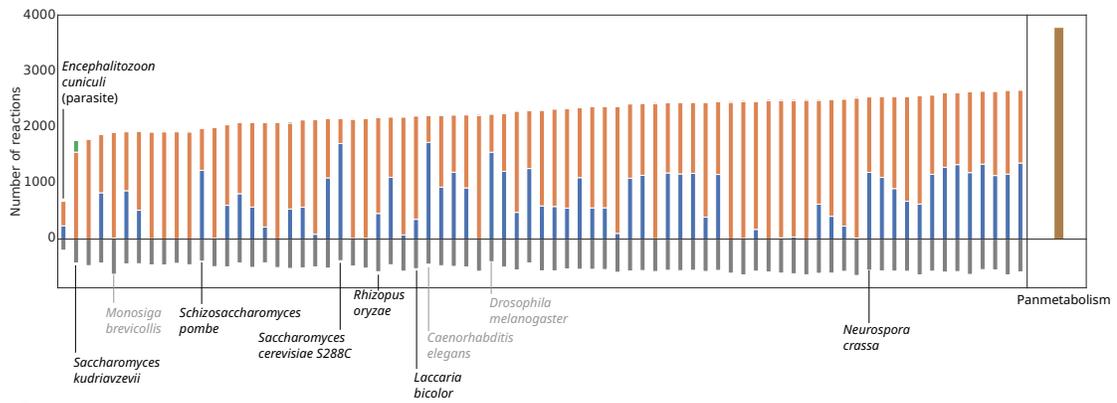
A.

Dataset	Bacterial Reactions				Fungal Reactions				Algal Reactions			
	Min	Max	Mean	Sd	Min	Max	Mean	Sd	Min	Max	Mean	Sd
Draft reconstruction	970.0	2123.0	1469.8	309.6	0.0	1742.0	635.1	520.0	0.0	2249.0	1263.4	695.8
Robust orthology	436.0	1175.0	765.2	227.4	442.0	2520.0	1648.9	509.3	547.0	2704.0	1410.8	666.6
Non robust orthology (not included in the GSMN)	(172.0)	(437.0)	(399.4)	(50.8)	(200.0)	(654.0)	(533.0)	(77.1)	(399.0)	(826.0)	(673.6)	(90.2)
Structural verification	0.0	94.0	21.6	29.7	0.0	209.0	5.6	24.0	0.0	86.0	5.0	13.5
Spontaneous completion	2.0	12.0	7.0	2.6	2.0	23.0	9.5	6.2	4.0	36.0	13.2	7.9
Final network	2047.0	2568.0	2263.6	88.5	685.0	2691.0	2297.0	300.1	2273.0	3052.0	2692.2	201.9
	Panmetabolism: 2903				Panmetabolism: 3810				Panmetabolism: 4880			

B.



C.



D.

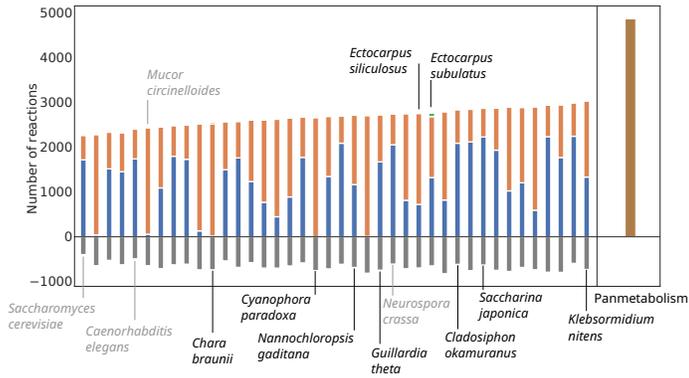


Figure 2: Application of the AuCoMe pipeline to the *bacterial*, *fungal* and *algal* datasets of genomes.

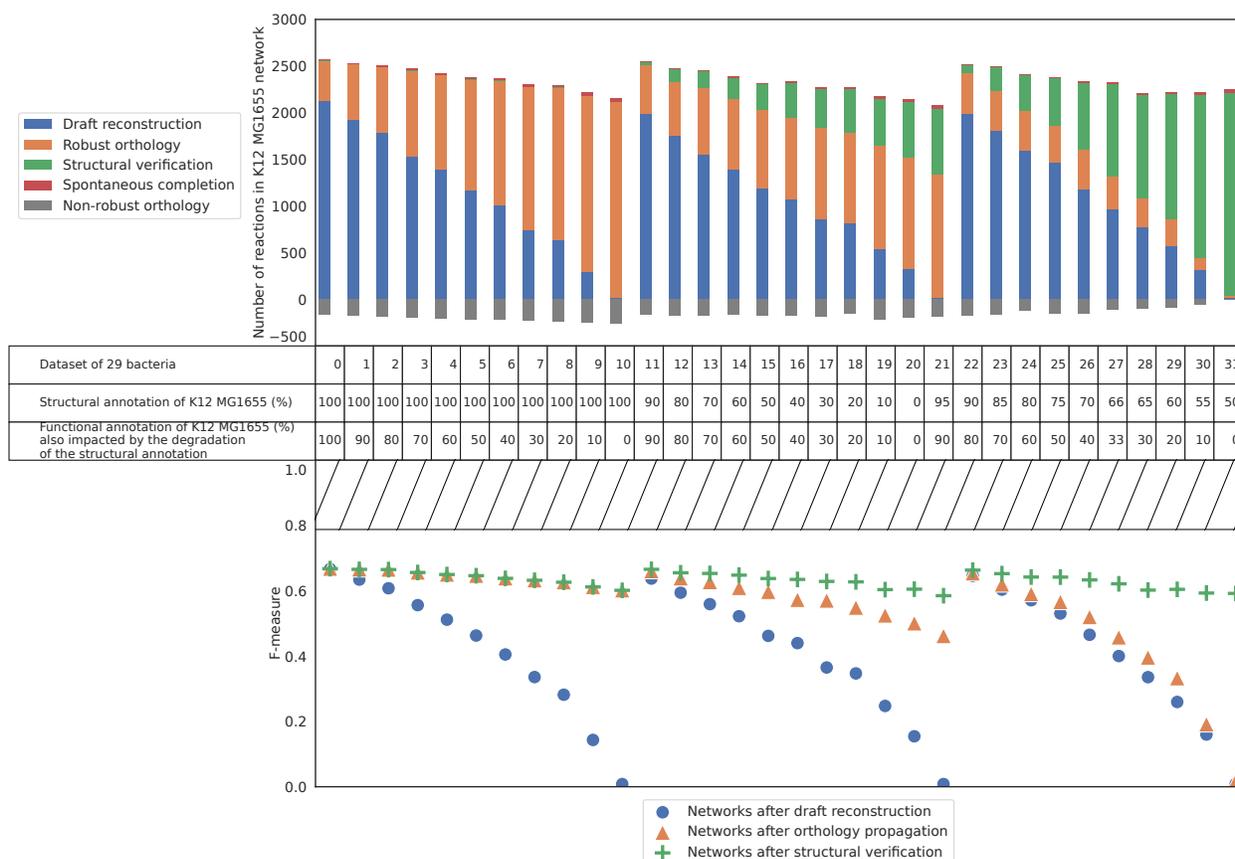


Figure 3: Efficiency of AuCoMe on degraded genome assemblies.

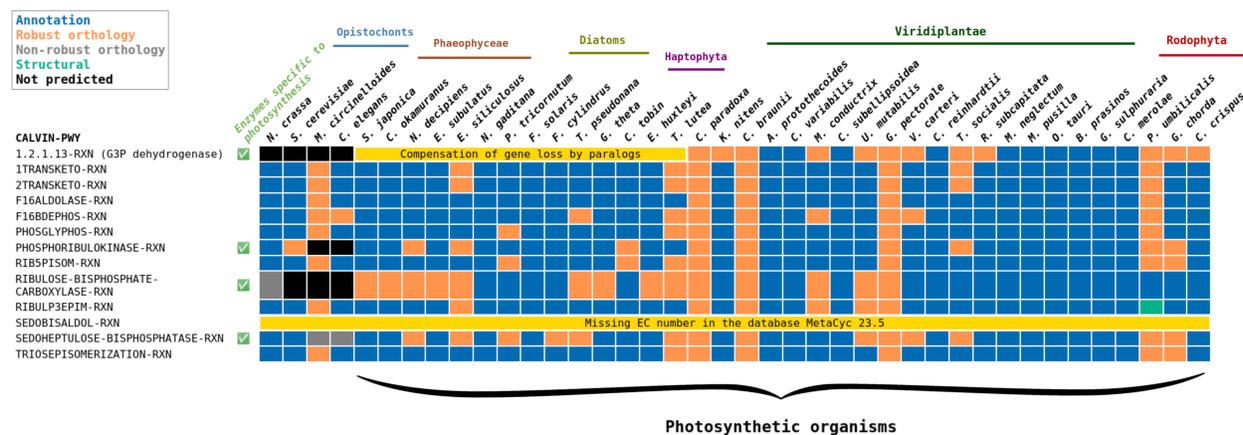


Figure 4: AuCoMe results on the Calvin cycle pathway in the algal dataset.

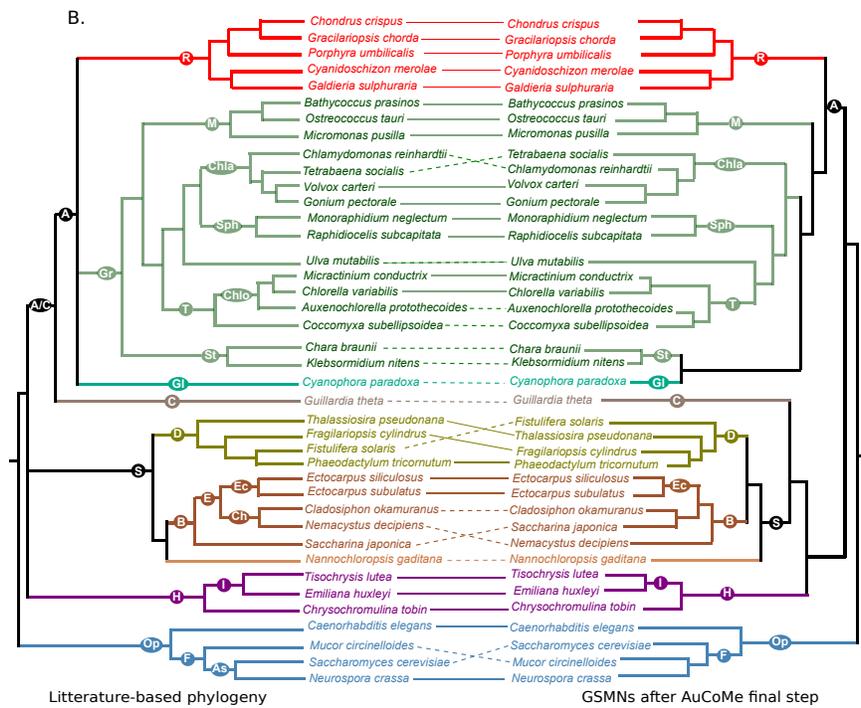
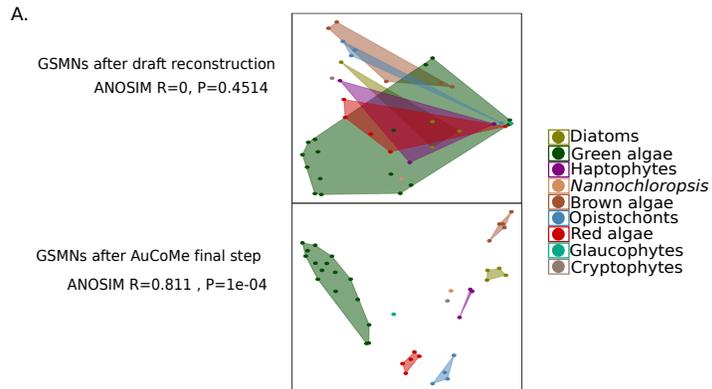


Figure 5: AuCoMe as a tool to improve taxonomic consistency of GSMNs.