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-03778267v2

HAL Id: hal-03778267
https://hal.science/hal-03778267v2
Preprint submitted on 2 May 2023

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

Arnaud Belcour\textsuperscript{1,*}, Jeanne Got\textsuperscript{1,*}, Méziane Aite\textsuperscript{1,*}, Ludovic Delage\textsuperscript{2}, Jonas Collén\textsuperscript{2}, Clémence Frioux\textsuperscript{3}, Catherine Leblanc\textsuperscript{2}, Simon M. Dittami\textsuperscript{2}, Samuel Blanquart\textsuperscript{1}, Gabriel V. Markov\textsuperscript{2,†} and Anne Siegel\textsuperscript{1,†}

March 23, 2023

1. Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
2. Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680 Roscoff, France
3. Inria, INRAE, Université de Bordeaux, France

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

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

Running Title: AuCoMe: Automatic Comparison of Metabolism

Character count (including spaces): 59743

Abstract

Comparative analysis of Genome-Scale Metabolic Networks (GSMNs) may yield important information on the biology, evolution, and adaptation of species. However, it is impeded by the high
heterogeneity of the quality and completeness of structural and functional genome annotations, which may bias the results of such comparisons. To address this issue, we developed AuCoMe – a pipeline to automatically reconstruct homogeneous GSMNs from a heterogeneous set of annotated genomes without discarding available manual annotations. We tested AuCoMe with three datasets, one bacterial, one fungal, and one algal, and demonstrated that it successfully reduces technical biases while capturing the metabolic specificities of each organism. Our results also point out shared metabolic traits and divergence points among evolutionarily distant algae, underlining the potential of AuCoMe to accelerate the broad exploration of metabolic evolution across the tree of life.
Introduction

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

For most GSMN analyses, some limitations still need to be addressed (Bernstein et al., 2021). When comparing different GSMNs, two main biases concern the variable quality of genome annotations and the multitude of reconstruction approaches. A variety of methods exists to perform structural (gene structure prediction) and functional (association of functions to genes) annotation steps (Yandell and Ence, 2012) and the method choice has previously been shown to have
direct effects on the reconstructed GSMNs (Karimi et al., 2021). Similarly, numerous methods for GSMN reconstruction have been developed, e.g. Pathway Tools (Karp et al., 2019), RAVEN (Wang et al., 2018), merlin (Dias et al., 2015; Capela et al., 2022), KBase (Arkin et al., 2018), ModelSEED (Devoid et al., 2013), AuReMe (Aite et al., 2018), AutoKEGGRec (Karlsen et al., 2018), CarVeMe (Machado et al., 2018), and gapseq (Zimmermann et al., 2021). They rely on one or several metabolic databases such as MetaCyc (Caspi et al., 2020), KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2017), ModelSEED (Seaver et al., 2021) or BiGG (King et al., 2016). Despite efforts in the direction of database reconciliation (Moretti et al., 2021), the heterogeneity of metabolic databases requires time-consuming matching of their respective identifiers and may thus impede the comparison of the GSMNs.

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

A second strategy to resolve GSMN comparison issues is to adapt the GSMN reconstruction method. This strategy aims at reducing annotation biases through the reconstruction of GSMNs from homogeneously annotated genomes using the same method and database, possibly followed by the propagation of annotations with sequence alignments (Vieira et al., 2011; Prigent et al., 2018). This strategy was pushed forward and automatized in the tool CoReCo, which enabled the reconstruction of gap-less metabolic networks from several non-annotated genomes (Pitkänen et al., 2014; Castillo et al., 2016). The main limitation of such approaches is that the re-annotation of the genomes supplants the previous genome annotation.
Annotations of genomes in databases also reflect the expertise of scientists. Their quality and precision, ranging from structural features, such as accuracy of intron-exon boundaries or functional inferences, like the assignation to a specific catalytic activity based on previous biochemical evidence, highly depend on the amount of curation effort done after the initial automated steps. Such valuable information is lost during a systematic re-annotation step. For a reliable interpretation of data, expert annotations therefore ought to be preserved while automatically inferring metabolic networks from any type of genomic resource. In this article, we introduce a new method, *AuCoMe* (Automated Comparison of Metabolism) that creates a set of homogenized GSMNs from heterogeneously-annotated genomes. This enables a less biased functional comparison of the networks and the determination of metabolic distances using the presence/absence of reactions. Our objective was to develop an efficient and robust approach, which does not depend on the quality of the initial annotations and is able to aggregate heterogeneous information in both prokaryote and eukaryote datasets. *AuCoMe* combines metabolic network reconstruction, propagation, and verification of annotations. The method automatizes the strategy of transferring information from the annotations of the genomes and complements this information transfer with local searches of missing structural annotations. *AuCoMe* was applied to three heterogeneous datasets composed of fungal, algal, and bacterial genomes. Our results demonstrate that *AuCoMe* succeeds at propagating missing reactions to degraded metabolic networks while capturing the metabolic specificities of organisms despite profound differences in the quality of genome annotations. This provides a knowledge base for the comparison of metabolisms between different organisms.

**Results**

**A tool for homogenizing metabolism inference**

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

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

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

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

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

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

Finally, the *spontaneous completion* step (Fig. 1D) adds spontaneous reactions to each metabolic network if these reactions complete BioCyc pathways (red bars in Fig. 2, Supplemental Fig. S1,
S2, and S3). For the fungal dataset, this step added between two and 23 spontaneous reactions, leading to two to 27 additional MetaCyc pathways that achieved a completion rate equal to 100%.

For the algae, the same step added between 4 and 36 spontaneous reactions yielding two to 31 additional pathways. The fewer reactions were inferred at the draft reconstruction step, the more spontaneous reactions were added to complete pathways (Pearson R = -0.83 and -0.84 for the fungal and algal datasets, respectively). The addition of these spontaneous reactions to the ones predicted by Pathway Tools (only other step predicting this type of reactions) lead to the prediction of less than a hundred spontaneous reactions per GSMN.

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

**Validation of AuCoMe predictions**

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

In the first experiment, we compared the GSMNs created by AuCoMe to those created by CarveMe, ModelSEED and gapseq on the bacterial dataset (Supplemental Fig. S4). On this dataset, AuCoMe performed well regarding the recovery of EC numbers, although, it does not reconstruct the largest GSMNs, limiting the inference of reactions to those associated with genes.

The ECs inferred by the different tools for *E. coli* K–12 MG1655 were compared with a reference containing ECs from EoCyc, KEGG, BiGG and ModelSEED associated with *E. coli* K–12 MG1655.
In this comparison, AuCoMe predicted the highest number of true positives (Supplemental Fig. S5). Then, a second comparison on the eukaryotes was performed with AuCoMe, gapseq find module and ModelSEED on 5 fungal genomes. Results on the eukaryotic genomes showed that AuCoMe predicts the most EC numbers, reactions, and pathways in species distant from the model ones (Supplemental Table S5 and Figure S7). A comparison with metabolic pathways contained in YeastCyc for the genome of *Saccharomyces cerevisiae S288C* was done to estimate the quality of the predicted pathways. AuCoMe predicted a high number of pathways with low completion rate not found in YeastCyc (Supplemental Figure S8). For pathways with completion rate above 70%, AuCoMe and gapseq exhibited similar performance (Supplemental Fig. S9). Although these experiments should be confirmed by an exhaustive comparative study, these results suggest that AuCoMe is suitable for the study of the metabolism of multiple eukaryotic genomes by predicting robust gene-reactions associations.

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

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

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

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

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

The accuracy of the annotation transfer procedure by AuCoMe was further assessed using two pathways where there where clear biological expectations in the algal dataset. The Calvin cycle is a biochemical pathway present in photosynthetic organisms to fix CO₂ into three-carbon sugars
composed of 13 reactions (MetaCyc identifier: CALVIN-PWY, Fig. 4).

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

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

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

AuCoMe GSMNs are consistent with species phylogeny

To further assess the predictions of AuCoMe and to explore biological features, we clustered the GSMNs of the algal dataset after the draft reconstruction as well as at the end of the pipeline by using the presence or absence of reactions in the GSMNs (see Fig. 5A). We compared these
clusterings with a phylogeny compiled from Strassert et al. (2021). The initial GSMNs produced from the annotations exhibited low consistency with the phylogenetic relationships. Even well-established phylogenetic groups like red algae or brown algae were not recovered. At this step, the principal factor leading to the repartition of points in the MDS was the heterogeneity of genome annotations. An ANOSIM test supports this as it was not able to differentiate the main phylogenetic groups (R=0, P-value=0.45). However, in the MDS made from the GSMNs after the final step of AuCoMe, we observed a clear separation between the known phylogenetic groups, supported by an ANOSIM test (R=0.811, P-value=1e-04). This is also visible in the dendrograms clustering the GSMNs generated by the complete AuCoMe pipeline, which was broadly consistent with the reference species phylogeny (Fig. 5B). There were only three higher order inconsistencies concerning C. paradoxa, for which the genome version deposited in GenBank fully lacked expert annotations (Price et al., 2012), G. theta, which belongs to Cryptophytes, for which the phylogenetic position is controversial (Strassert et al., 2021), and N. gaditana, which was the only representative of eustigmatophycean stramenopiles. The two other stramenopile groups, diatoms and brown algae, were represented by multiple species which likely minimizes errors linked with peculiarities of a single genome. There were also some minor inconsistencies in intra-group relationships, in green algae, diatoms, brown algae, and opisthokonts.

An illustration of the efficiency of AuCoMe was the de novo reconstruction of the GSMN of the glaucophyte C. paradoxa. For the reconstruction of this GSMN, we used the initially published genome sequence, which contained only two functionally-annotated genes (Price et al., 2012). The draft reconstruction by AuCoMe enabled us to retrieve 1,675 GPRs, a number within the same range as the other species from the dataset. Accordingly, C. paradoxa branched at the basis of the dendrogram after the draft reconstruction step, whereas it moved to the archeplastids after the orthology propagation step. Even if the grouping of C. paradoxa within archeplastids with the streptophytes Chara braunii and Klebsormidium nitens does not reflect the phylogenetic relationships, this shows that AuCoMe is a reasonable proxy for handling nearly unannotated genome sequences.
By exploring cluster of reactions shared in phylogenetic groups (as shown in Supplemental Fig S12), results of AuCoMe could pave the way to the identification of gene candidates for enzymatic reactions. We analyzed a cluster of fourteen reactions present in *Cladosiphon okamuranus* and *Saccharina japonica* but absent in other brown algae (see Supplemental Table S8). Among those fourteen reactions, twelve were enzymatic reactions assigned based on annotations, but orthology propagation in the AuCoMe pipeline identified only a subset of the potential orthologs (see Supplemental Table S9). A focus was made on the o-aminophenol oxidases. Comparative genomics analysis using sequences from additional BLASTP searches showed that potential homologs were present for the other brown algae (see Supplemental Fig. S13). The o-aminophenol oxidase family proteins present in the genome of *E. siliculosus* are predicted to be cytoplasmic, extracellular, or to target the membrane (see Supplemental Table S10), suggesting different roles depending on their subcellular localization. In this case, AuCoMe, with the support of more focused analyses, led to the identification of numerous candidate o-aminophenol oxidases in stramenopiles.

By exploring the group of stramenopiles in the final GSMN dendrogram (Fig. 5B), we noticed that it grouped with the small unicellular alga *G. theta*, which belongs to the cryptophytes, usually grouping with the archeplastids or the haptophytes. Its plastid is derived from a secondary endosymbiosis event with a red alga (Curtis et al., 2012). The phylogenetic position of cryptophytes is unclear, but they have been suggested to be phylogenetically separate from haptophytes closer to the green algae lineage (Burki et al., 2012). To further examine the position of *G. theta* in our metabolic trees, we analyzed the presence/absence matrix of metabolic reactions to determine which of them most clearly linked *G. theta* to each of the three groups in question (stramenopiles, archeplastids, haptophytes). To this means we focused on reactions that distinguished at least two of these groups, i.e. that were present in at least 80% of the networks of at least one group, and absent from at least one other group (Supplemental Table S11). A total of 216 reactions met this criterion, 109 of which were found in *G. theta* and 107 were absent. We found that the network of *G. theta* shared the presence or absence of a similar number of distinctive reactions with all three groups: 120 with stramenopiles, 112 with haptophytes, and 101 with archeplastids.
Next, we examined the metabolic pathways represented by the reactions that associated \textit{G. theta} with the three groups, focusing on pathways that were more than 50% complete. The metabolic networks showed, for instance, that \textit{G. theta}, (i) like haptophytes in our dataset, possess parts of the mitochondrial L-carnitine shuttle pathway, (ii) like the stramenopiles, comprises the complete pathway of glycine betaine synthesis, and, (iii) like terrestrial plants, can synthesize carnosine. We also manually examined the genes associated with these reactions, and found that in all cases, their sequences differed strongly from other sequences in the database, and could not be clearly associated with either archeplastids, stramenopiles, or haptophytes (see Supplemental Table S12).

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

**Discussion**

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

**Accounting for existing annotations in the inference of homogenized GSMNs**

Automatic inference of single species GSMNs is now routinely achieved, especially for prokaryotic species, and is often systematically performed for multiple genomes. With such data at hand, one may compare the predicted metabolism among related species from a given clade and subsequently identify metabolic specificities or putative functional interactions in microbial communities (Machado \textit{et al.}, 2018; Frioux \textit{et al.}, 2018). Such applications require consistent genome quality and
similar data treatment (genome annotation, metabolic network reconstruction) to minimize biases in predictions. However, ensuring the latter is complex for eukaryotic genomes, as their enzymatic functions are difficult to characterize automatically and they often need expert annotation. Moreover, annotation efforts can greatly vary between genomes, resulting in heterogeneous annotation and metabolic prediction quality. As the automatization of both (meta)genome reconstruction and annotation is now routinely applied, it is likely that efforts toward manual annotation will decline. However, we believe the need to manually curate annotations will remain [Karimi et al., 2021]. In addition, AuCoMe could also be used to homogenize annotations in several genome versions of the same species, or to reconcile several annotations performed on the same genome.

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

Method limitations and improvements

AuCoMe incorporates several strategies to optimize the method’s selectivity and sensitivity. Together these strategies collectively achieve comparable GSMN reconstruction with two objectives: having comparisons as homogeneous as possible given the initial heterogeneity and incompleteness.
of databases, and thus identifying errors that can be corrected during further analysis.

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

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

Finally, we want to emphasize that our attempts to limit the inference of false positive reactions also directed the choice of method for the initial draft metabolic inference. We used Pathway Tools because of its several advantages such as the capacity to work with eukaryotic genomes, the
suitability for parallel computing (Belcour et al., 2020a), and the possibility to limit gap-filling of
metabolic networks. However, metabolic pathway completion performed by Pathway Tools does
not systematically extend to ensuring the production of biomass. Pathway Tools was therefore
adapted to our objective of avoiding to go beyond the strict interpretation of genome annotations.
This goal was fulfilled, as attested by the benchmark shown in Supplemental Fig. S4 which confirms
that AuCoMe GSMNs have by design no reaction lacking gene association.

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

Biological insights on the comparison of metabolic networks across species

Evolution

Our examples of the Calvin cycle and phycobiliprotein synthesis demonstrate that, once all steps
of the AuCoMe pipeline have been executed, the predicted metabolic capacities of the analyzed
genomes reflect the biological knowledge we have of the corresponding organisms. Our approach,
therefore, enables GSMNs to be compared in the light of evolutionary biology. The metabolic
dendrograms calculated from final AuCoMe reconstruction are mostly consistent with reference
species phylogeny. Indeed, numerous studies have shown that comparing GSMNs by computing a
metabolic distance and arranging them into a dendrogram allows clustering organisms into groups
close to the ones known by phylogenetic analysis. However, the position of species inside these
groups is often different from the one of the phylogenetic groups (Vieira et al., 2011; Bauer et al., 2015; Prigent et al., 2018; Schulz and Almaas, 2020). It furthermore gives support to the hypothesis of a metabolic clock based on the congruence between molecular and metabolomic divergence in phytoplankton (Marcellin-Gros et al., 2020). The difference observed in the tanglegram (Fig. 5B) between phylogeny and metabolic distances could be further explored. One possibility could be to look at different similarity measures for the clustering. In this work, the Jaccard distance has been used but other measures could be used. For example, if we consider an absence of a reaction in two organisms as a similarity (to represent the loss of a function) then other measures could be envisaged such as the Simple Matching Coefficient. This also opens the perspective of inferring ancestral metabolic networks to better understand the dynamics of character evolution across time (Psomopoulos et al., 2020).

Adaptation

The second aim of reconstructing comparable GSMNs is to determine to what extent metabolic changes are the result of or the prerequisite for adaptation. In our study, we made a first attempt at this question regarding the cryptophyte *Guillardia theta*. This species has several potentially plesiomorphic metabolic traits in common with other marine lineages, that may constitute adaptations to their shared marine environment. Glycine-betaine, for instance, is known to be an osmoregulator or osmoprotectant in green plants (Di Martino et al., 2003), and carnosine has been proposed to function as an antioxidant in red algae (Tamura et al., 1998). Regarding carnitine, its physiological significance in photosynthetic organisms is still largely unknown, but antioxidant and osmolyte properties along with signaling functions have also been suggested (Jacques et al., 2018). However, for now, all of this remains purely hypothetical. To dig deeper into such questions in the future, we need to be able to distinguish changes that simply result from random processes such as metabolic drift (Belcour et al., 2020b) from changes that have an adaptive value. Currently, we envision two approaches that will help with this distinction. The first approach will be to further increase the number of species and lineages included in order to identify adaptive patterns, for
example to among organisms occupying similar ecological niches. In phylogenomics, wide taxon
sampling is recognized as one of the key features for reliable comparisons (Young and Gillung
2020), whereas pairwise genomic comparisons across species are generally viewed as problematic
(Dunn et al. 2018). Given that, as demonstrated above, phylogenetic signals in metabolism are
stronger than the adaptive signals we can expect, this approach would also benefit from the devel-
opment or adaptation of statistical models that could help detect signals of adaptation in an overall
noisy dataset. Such models exist, for instance, to detect selective signatures in the evolution of
protein-coding gene (Shapiro and Alm 2008), but to our knowledge, have not been developed for
metabolic networks or presence/absence signatures of genes. The second related strategy consists
in focusing on phylogenetically closely related species that have only recently diverged and adapted
to different environments. In such cases, we anticipate that the relative importance of drift along
with the noise from the phylogenetic signal will be reduced due to the short evolutionary time since
the separation. With such datasets, we may be able to reduce the level of replication required to
find biologically relevant metabolic adaptations. The range of questions that could be addressed
with the appropriate dataset is long and includes metabolic adaptations to different environments
(Xu et al. 2020), food sources and domestication (Giannakou et al. 2020), multicellularity (Cock
et al. 2010), or even life-history transitions to endophytism (Bernard et al. 2019).

Interactions

Lastly, we anticipate that AuCoMe will provide new opportunities to study metabolic interactions
between symbiotic organisms. For example, the tentative o-aminophenol oxidase activities pointed
out by AuCoMe in brown algae could be involved in the protection against pathogen attacks at
the cell surface. Indeed, a molecular oxygen-scavenging function in the chloroplast (Constabel
et al. 1995) and a defense role (Gandía-Herrero et al. 2005) have been suggested for these enzymes
in terrestrial plants. An o-Aminophenol Oxidase Streptomyces griseus is known to be involved
in the grixazole biosynthesis, i.e. an antibiotic (Suzuki et al. 2006). Similarly, brown algal o-
aminophenol oxidases or tyrosinases might be involved in the production of specific antibiotics.
The o-aminophenol oxidase enzymes resemble laccases or tyrosinases. They can be involved in catechol or pigment production by oxidation ([Le Roes-Hill et al., 2009]). Numerous references have also shown that tyrosinases are efficiently inhibited by some phlorotannins, antioxidant compounds specific to the brown algae ([Kang et al., 2004] [Manandhar et al., 2019]) suggesting there might be a regulation of polyphenol oxidation in certain conditions.

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

Methods

Genomes and models

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

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

The algal dataset contains 36 algal genomes selected to represent a wide diversity of photosyn-
thetic eukaryotes and downloaded from public databases. The dataset includes 16 Viridiplantae (green algae), 5 Phaeophyceae (brown algae), 5 Rhodophyceae (red algae), 4 diatoms, 3 haptotytes, 1 cryptophyte (Guillardia theta), 1 Eustigmatophyceae (Nannochloropsis gaditana), 1 Glaucophyceae (Cyanophora paradoxa). The genomes of C. elegans (Witting et al., 2018), Muco circinelloides (Vongsangnak et al., 2016), N. crassa (Dreyfuss et al., 2013), and S. cerevisiae (Lu et al., 2019) were selected as outgroup genomes (see Supplemental Table S3).

Each annotated genome of the datasets was curated manually in order to make it compatible with Pathway Tools v23.5. Curated genomes are available at https://zenodo.org/record/7752449#.ZBhOpiOZN-E.

AuCoMe, a method to reconstruct genome-scale metabolic networks homogenized across related species

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

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

The aucome reconstruction command runs the draft reconstruction step, which consists in reconstructing draft GSMNs according to the set of available genome annotations. During this step, the pipeline first checks the input GenBank files using Biopython (Cock et al., 2009). Then
using the mpwt package \cite{Belcour2020}, AuCoMe launches parallel processes of the Patho-
Logic algorithm of Pathway Tools \cite{Karp2019}. Pathway Tools creates Pathway/Genome
Databases (PGDB) for all genomes. The resulting PGDBs are converted into PADMET and SBML
files \cite{Hucka2003, Hucka2018} using the PADMet package \cite{Aite2018}. During this con-
version, pathway hole reactions (reactions predicted by Pathway Tools for which no enzymes were
detected in the genomes) are removed as they are not associated with a gene and are not sponta-
eous reactions. For example, in Fig. 1A, the draft reconstruction step generates 6 GPRs in total
for the 3 considered genomes.

The \texttt{aucome orthology} command runs the \textbf{orthology propagation step}, which complements
the previous GSMNs with GPRs associations whose genes are predicted to be orthologs to genes
from GPR relations of other GSMNs of the dataset (Fig. 1B). To that purpose, the pipeline relies
on OrthoFinder \cite{Emms2015, Emms2019} for the inference of orthologs defined as clusters
of homologous proteins shared across species. For each pair of orthologous genes shared between
two species, the pipeline checks whether one of the genes is associated with an existing GPR
association. If so, a putative GPR association with the orthologous gene is added to the GSMN. At
the end of the analysis of all genomes, a robustness score is calculated for assessing the confidence
of each putative GPR association based on the number of annotated GPRs associations between
the orthologs (see below). Non-robust GPR associations are not integrated in the final GSMNs. In
the example shown in Fig. 1B, applying the robustness criteria leads to generating a putative new
GPR association in the GSMN 2 (see the green orthogroup). In this example, the pipeline does
not validate the GPR association related to the blue orthogroup because of insufficient annotation
support.

The \texttt{aucome structural} command runs the \textbf{structural verification step} to identify GPRs
associated with missing structural annotations of the input genomes. This pipeline step comple-
ments GSMNs with GPR associations from other GSMNs according to protein-against-genome
alignment criteria. This enables the identification of reactions which are associated with gene se-
quences absent from the initial structural annotations of the input genomes. A pairwise comparison
For each protein sequence associated with a GPR relation in a GSMN, a TBLASTN (Altschul et al., 1990; Camacho et al., 2009) with Biopython (Cock et al., 2009) is performed against the other genome. If a match (e-value < 1e-20) is found, the gene prediction tool Exonerate (Slater and Birney, 2005) is run on the region linked to the best match (region +- 10 KB). If Exonerate finds a match, then the reaction associated with the protein sequence is added. In Fig. 1C, one reaction is added to the GSMN 2.

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

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

The robustness score of GPR associations of the pan-metabolic network after the orthology propagation was defined as illustrated in Algorithm 1 and detailed in the following. We denote by \( \text{org}(g) \) the organism of a gene \( g \). For every pair of genes \( g_1, g_2 \) of two different organisms, we denote \( \text{orth}(g_1, g_2) = 1 \) if the genes are predicted to be orthologs. We denote by \( \text{association}(r, g) = 1 \) a GPR association between a reaction \( r \) and a gene \( g \) which is predicted by the AuCoMe algorithm. When the gene-association is predicted by the draft reconstruction step, we denote \( \text{annot.type}(r, g) = 1 \) (and zero otherwise). When the gene-association is predicted according to orthology criteria, we denote \( \text{ortho.type}(r, g) = 1 \) (and zero otherwise).

Let us consider now a reaction \( r \) of the pan-metabolic network. We denote by \( N_{\text{org}}(r) \) the number of organisms for which the reaction \( r \) has been associated with a GPR relationship with
any gene \( g \): \( N_{\text{org}}(r) = \#\{\text{org}(g), \text{association}(r,g) = 1\} \) (L2, Alg. 1). For every gene \( g \) with \( \text{annot.type}(r,g) = 1 \), we denote by \( N_{\text{prop}}(r,g) \) the number of organisms different from \( \text{org}(g) \) the GPR association between \( r \) and \( g \) has been propagated to according to an orthology relation with the gene \( g \) \( N_{\text{prop}}(r,g) = \#\{\text{org}(g1), \exists g1 \text{ s.t. org}(g1) \neq \text{org}(g), \text{orth}(g,g1) = 1, \text{association}(r,g1) = 1\} \). The GPR association between \( r \) and \( g \) is considered robust: \( \text{robust}(r,g) = 1 \) as long as \( \text{annot.type}(r,g) = 1 \).

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

In the second scenario the GPR association between \( r \) and \( g \) was propagated from a unique gene \( g1 \) with \( \text{annot.type}(r,g1) = 1 \) in the orthology cluster (L11, Alg. 1). For these genes our strategy is to be as stringent as possible and we introduce a robustness criterion to reduce the risk of propagating false-positive reactions. The GPR association is considered robust if the number of organisms to which the reaction is propagated according to the annotation of \( g1 \) remains low with respect to the total number of considered organisms. More precisely, \( \text{robust}(r,g) = 1 \) if \( N_{\text{prop}}(r,g1) \leq \lceil \text{robust.func}(N_{\text{org}}(r) - 1) \times (N_{\text{org}}(r) - 1) \rceil \) (L12-13, Alg. 1). The robustness function \( \text{robust.func}^{(t)}(x) = \min(1, \frac{1}{2} \max([t \times],[\frac{5}{2}])) \) was chosen such that it is 1 for low values of \( N_{\text{org}} \), and then decreases to a threshold value (by default \( t = 0.05 \)) for large values of \( N_{\text{org}} \) (see a plot in Supplemental Fig S14).

Altogether, the robustness criterion removes orthology predictions for GPR associations that are supported by a unique gene annotation and propagated to a large number of organisms. A toy example of the application of the algorithm is detailed in Supplemental Methods Section and Fig S15).
Algorithm 1 Robustness criterion algorithm

1: for $r$ in panmetabolism do
2:   $N_{\text{org}}(r) \leftarrow \#\{\text{org}(g), \exists g \text{ s.t. } \text{association}(r, g) = 1\}$ \Comment{Number of organisms with GPRs relations to $r$}
3: for all genes $g$ s.t. $\text{annot}_{\text{type}}(r, g) = 1$ do
4:   $\text{robust}(r, g) = 1$
5:   $N_{\text{prop}}(r, g) \leftarrow \#\{\text{org}(g_1), \exists g_1 \text{ s.t. } \text{org}(g_1) \neq \text{org}(g), \text{orth}(g, g_1) = 1, \text{association}(r, g_1) = 1\}$. \Comment{Number of organisms to which the GPR has been propagated to}
6: end for
7: for all genes $g$ s.t. $\text{annot}_{\text{type}}(r; g) = 0$ and $\text{orth}_{\text{type}}(r; g) = 1$ do \Comment{Only other way to add the reaction}
8:   if $\exists g_1, g_2 \text{ s.t. } \text{orth}(g, g_1) = \text{orth}(g, g_2) = 1$ and $\text{annot}_{\text{type}}(r, g_1) = \text{annot}_{\text{type}}(r, g_2) = 1$ then
9:     $\text{robust}(r, g) = 1$
10: else \Comment{Prevent the propagation of an isolated annotation to too many organisms}
11:     $g_1 \leftarrow \text{unique gene s.t. } \text{orth}(g, g_1) = 1 \text{ and } \text{annot}_{\text{type}}(r; g_1) = 1$
12:     if $N_{\text{prop}}(r, g_1) \leq \text{robust}_{\text{func}}(N_{\text{org}}(r) - 1) \times (N_{\text{org}}(r) - 1)$ then
13:         $\text{robust}(r, g) = 1$
14:     else
15:         $\text{robust}(r, g) = 0$
16:     end if
17: end if
18: end for
19: end for

Validation of AuCoMe predictions

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

A second comparison was made on the eukaryotes and especially the fungal dataset (using five organisms: *Laccaria bicolor*, *Neurospora crassa*, *Rhizopus oryzae*, *Saccharomyces cerevisiae S288C* and *Schizosaccharomyces pombe*). We used Kbase [Arkin et al., 2018] and gapseq 1.2 [Zimmermann et al., 2021]. The genomes were imported into Kbase and the metabolic networks were reconstructed with ‘Build Fungal Model’ 1.0.0 (with gap-filling). We also used gapseq to predict the metabolic
pathways present in an organism using its *find* module associated with the option '-t Fungi'. We did not use CarveMe as it has been developed for Bacteria or Archaea (Capela et al., 2022). We compared the completion rate of metabolic pathways predicted by AuCoMe and gapseq. Then for *Saccharomyces cerevisiae S288C*, we used the reference network YeastCyc to estimate the quality of the pathways predicted by both gapseq and AuCoMe.

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

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

Finally, the complementarity between the orthology propagation step (second step) and the structural verification step (third step) was assessed using the *E. coli* K–12 MG1655 genome modified to generate replicates with randomly degraded annotations associated with GPR of the non-degraded *E. coli* K–12 MG1655 GSMN. Two degradation types were simulated, (i) a degradation of the functional annotations of the genes, where all the annotations like GO Terms, EC numbers,
gene names, etc. associated with a reaction were removed, and (ii) a degradation of the structural annotation of the genes, where gene positions and functional annotations were removed from the genome annotations. A third type of replicate was considered including the degradation of both structural and functional annotations. Replicates with increasing percentages of degraded annotations were generated for each of the three types of degradation. Details on the degradation algorithm are shown in the Supplementary file (section Methods). Furthermore the taxonomic ID associated with the E. coli K–12 MG1655 genome was degraded to cellular organism, to focus on the impact of genome annotations on GSMN reconstructions by AuCoMe, rather than on the effect of the automatic completion by the EcoCyc source performed by Pathway Tools when analyzing E. coli K–12 MG1655 . Each degraded replicate was associated with the 28 other E. coli and Shigella genomes, generating 31 synthetic bacterial datasets, plus the dataset with non-degraded E. coli K–12 MG1655 genome, which was called dataset 0. Their characteristics are detailed in Supplemental Table S4. For each E. coli K–12 MG1655 replicate in a dataset, AuCoMe produced a GSMN, which was compared to EcoCyc, considered as ground truth (Karp et al., 2002; 2018a; Keseler et al., 2021). For more information on the computation of the F-measure, see Supplemental file (section Methods).

Phylogenetic analysis of the brown algal o-aminophenol oxidases

A dataset of 193 protein sequences was constructed using the closest homologs of the S. japonica o-aminophenol oxidase (SJ09941) in brown algae and extended to more distant sequences present in other organisms. Sequences were submitted to NGPhylogeny.fr via the "A la carte" (Lemoine et al., 2019) pipeline. The alignment was carried out by MAFFT (Katoh and Standley, 2013) using default parameters and automatically cleaned with trimAl (Capella-Gutiérrez et al., 2009) to obtain 372 informative positions. Then a maximum likelihood phylogenetic reconstruction was carried out using default parameters of the PhyML-SMS tool (Guindon et al., 2010; Lefort et al., 2017) allowing the best substitution model selection. Bootstrap analysis (Lemoine et al., 2018) with 100 replicates was used to provide estimates for the phylogenetic tree topology. The Newick
file (Junier and Zdobnov, 2010) was further formatted by MEGA v10.1.1 (Tamura et al., 2021) to obtain the simplified dendrogram (see Supplemental Fig. S13).

Supplemental Files and Software availability

Supplemental Files

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

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

Software availability

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

Running AuCoMe on the datasets studied in the paper required as dependencies BLAST v2.6.0
Altschul et al. (1990), Diamond v0.9.35 (Buchfink et al., 2015), Exonerate v2.2.0 (Slater and Birney, 2005), FastME v2.1.15 (Lefort et al., 2015), MCL (Enright et al., 2002), MMseqs2 v11-e1a1c (Steinegger and Söding, 2017), OrthoFinder v2.3.3 (Emms and Kelly, 2015, 2019), Pathway Tools v23.5 (Karp et al., 2019). The following Python packages are needed to install AuCoMe: Matplotlib, mpwt v0.6.3 (Belcour et al., 2020a), padmet v5.0.1 (Aite et al., 2018), rpy2 v3.0.5, seaborn, supervenn, and tzlocal. The pvclust R package is also required.

A docker or a singularity container can be created and enriched according to the dockerfile available on [https://github.com/AuReMe/aucome/blob/master/recipes/Dockerfile](https://github.com/AuReMe/aucome/blob/master/recipes/Dockerfile).

**Acknowledgements**

We acknowledge the GenOuest bioinformatics core facility [https://www.genouest.org](https://www.genouest.org) for providing the computing infrastructure. We also thank Erwan Corre (ABiMS Platform) and Pauline Hamon-Giraud for fruitful discussions. This work benefited from the support of the French Government via the National Research Agency investment expenditure program IDEALG (ANR-10-BTBR-04) and from Région Bretagne via the grant ≪SAD 2016 - METALG (9673)≫.

**Author Contributions**

**AB**: Conceptualization, Data curation, Methodology, Formal Analysis, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **JG**: Data curation, Formal Analysis, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **MA**: Conceptualization, Data curation, Methodology, Software. **LD**: Formal Analysis, Validation, Writing – original draft, Writing – review & editing. **JC**: Formal Analysis, Validation, Writing – review & editing. **CF**: Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **CL**: Funding acquisition, Writing – original draft, Writing – review & editing. **SD**: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Validation, Writing – original draft, Writing – review & editing. **SB**: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **GVM**: Data curation, Formal Analysis,
Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AS: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Conflict of interest

The authors declare no conflict of interest.

References


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


technology reconstruction of metabolic pathways across pangenome ensembles. *Microbial genomics*
DOI[10.1099/mgen.0.000429].

Rockwell NC and Lagarias JC. 2017. Ferredoxin-dependent bilin reductases in eukaryotic algae:

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


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

Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, Mundy M, Chia N, Noor E,
annotations and the reconstruction, comparison and analysis of metabolic models for plants,

signatures. *PLOS Genetics* **4**: 1–12. DOI[10.1371/journal.pgen.0040023].

Slater GSC and Birney E. 2005. Automated generation of heuristics for biological sequence com-

Stamboulian M, Guerrero RF, Hahn MW, and Radivojac P. 2020. The ortholog conjecture revis-
DOI[10.1093/bioinformatics/btaa468].

Steinegger M and Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the


**Figure Legends**

**Figure 1** Reconstruction and homogenization of metabolisms with AuCoMe. Starting from a dataset of partially structurally- and functionally-annotated genomes, AuCoMe’s pipeline
performs the following four steps. **A. Draft reconstruction.** The reconstruction of draft genome-scale metabolic networks (GSMNs) is performed using Pathway Tools in a parallel implementation. **B. Orthology propagation.** OrthoFinder predicts orthologs by aligning protein sequences of all genomes. The robustness of orthology relationships is evaluated (see Methods), and GPRs of robust orthologs are propagated. **C. Structural verification.** The absence of a GPR in genomes is verified through pairwise alignments of the GPR-associated sequence to all genomes where it is missing. If the GPR-associated sequence is identified in other genomes, the gene is annotated, and the GPR is propagated. **D. Spontaneous completion.** Missing spontaneous reactions enabling the completion of metabolic pathways are added to the GSMNs. GSMN: Genome-scale metabolic network. OG: orthologs. GPR: Gene-protein-reaction relationship. Outlines around GPR or reactions indicate that the GPR or reaction is newly added during the corresponding step.

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

Figure 3 Efficiency of AuCoMe on degraded genome assemblies. (A) Number of reactions in *E. coli* K–12 MG1655 degraded networks after application of AuCoMe to 32 synthetic bacterial datasets. Each dataset consists of the genome of *E. coli* K–12 MG1655 to which degradation of the functional and/or structural annotations was applied, together with 28
bacterial genomes. Each vertical bar corresponds to the result on the *E. coli* K–12 MG1655 within a synthetic dataset, with the percentages of degraded annotations indicated below. The dataset labelled 0 was not subject to degradation of the *E. coli* K–12 MG1655 annotations. Three types of degradation have been performed: functional annotation degradation only (left side, datasets labelled 1 to 10), structural annotation degradation only (right side, datasets labelled 22 to 31) and both degradation types (middle, dataset labelled 11 to 21). The colored bars depict the number of reactions added to the degraded network at the different steps of the method (the blue, orange, green, grey and red color legends are as described in the figure [2]). The table shown as axis indicates the dataset number and the percentage of functional or structural annotation impacted by the degradation for the corresponding column in both subfigures. **(B) F-measures after comparison of the GSMNs recovered for each *E. coli* K–12 MG1655 genome replicates with a gold-standard network.** Reactions inferred by each AuCoMe step for each replicate were compared to the gold-standard EcoCyc GSMN, allowing for the computation of F-measures. F-measures obtained after the draft reconstruction step, the orthology propagation step, or the structural verification step are shown as blue circles, orange triangles, and green crosses, respectively. The hashed rectangle from F-measure 0.79 to 1 highlights the values of F-measure, which are unreachable because 1019 reactions in EcoCyc were not present in the panmetabolism of the 29 non-degraded bacteria.

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

**Figure 5: AuCoMe as a tool to improve taxonomic consistency of GSMNs.** A. MDS
plot for GSMNs calculated with the AuCoMe draft reconstruction step or after all AuCoMe steps. In both cases, ANOSIM values are indicated below (MDS and ANOSIM were computed using the vegan package (Oksanen et al., 2020)). B. Tanglegram evaluating the taxonomic consistency of AuCoMe dendrograms based on metabolic distances using the pvclust package (Suzuki and Shimodaira, 2006) with the Jaccard distance (right side) in comparison with reference phylogeny (left side), compiled from Strassert et al. (2021). Full lines join species for which the position in the AuCoMe dendrogram is consistent with the reference phylogeny. Dotted lines join species for which the metabolic dendrogram and the reference phylogeny diverge. A/C: Archeplastids/Cryptophytes, A: Archeplastids, R: Rodophytes, Gr: Green algae, M: Mamiellales, Chla: Chlamydomonadales, Sph: Sphaeropleales, T: Trebouxiophyceae, Chlo: Chlorellaceae, St: Streptophytes, Gl: Glauco-phytes, C: Cryptophytes, H: Haptophytes, I: Isochrysida, D: Diatoms, S: Stramenopiles, B: Brown algae, E: Ectocarpales, Ec: Ectocarpaceae, Ch: Chordariaceae, Op: Opistochonts, F: Fungi, As: Ascomycetes.
Figure 1: Reconstruction and homogenization of metabolisms with AuCoMe.
### Table 1: Summary of Reconstruction Metrics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Sd</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Sd</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Sd</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Draft reconstruction</strong></td>
<td>150</td>
<td>2568</td>
<td>2263</td>
<td>88.5</td>
<td>685</td>
<td>2691</td>
<td>2297</td>
<td>300</td>
<td>2273</td>
<td>2692</td>
<td>201.9</td>
<td></td>
</tr>
<tr>
<td><strong>Robust orthology</strong></td>
<td>172</td>
<td>2575</td>
<td>761</td>
<td>29.7</td>
<td>442</td>
<td>2520</td>
<td>1648</td>
<td>509</td>
<td>547</td>
<td>2704</td>
<td>1410</td>
<td>666</td>
</tr>
<tr>
<td><strong>Non-robust orthology</strong></td>
<td>0</td>
<td>940</td>
<td>21.6</td>
<td>29.7</td>
<td>0</td>
<td>2090</td>
<td>5.6</td>
<td>24.0</td>
<td>0</td>
<td>860</td>
<td>5.0</td>
<td>13.5</td>
</tr>
<tr>
<td><strong>Structural verification</strong></td>
<td>2.0</td>
<td>12.0</td>
<td>7.0</td>
<td>2.6</td>
<td>2.0</td>
<td>23.0</td>
<td>9.5</td>
<td>6.2</td>
<td>4.0</td>
<td>36.0</td>
<td>13.2</td>
<td>7.9</td>
</tr>
<tr>
<td><strong>Spontaneous completion</strong></td>
<td>2.0</td>
<td>12.0</td>
<td>7.0</td>
<td>2.6</td>
<td>2.0</td>
<td>23.0</td>
<td>9.5</td>
<td>6.2</td>
<td>4.0</td>
<td>36.0</td>
<td>13.2</td>
<td>7.9</td>
</tr>
<tr>
<td><strong>Final network</strong></td>
<td>2047</td>
<td>2568</td>
<td>2263</td>
<td>88.5</td>
<td>685</td>
<td>2691</td>
<td>2297</td>
<td>300</td>
<td>2273</td>
<td>2692</td>
<td>201.9</td>
<td></td>
</tr>
</tbody>
</table>

### Figure 2: Application of the AuCoMe pipeline to the bacterial, fungal and algal datasets of genomes.
Dataset of 29 bacteria

Structural annotation of K12 MG1655 (%)

<table>
<thead>
<tr>
<th></th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>90</th>
<th>80</th>
<th>70</th>
<th>60</th>
<th>50</th>
<th>40</th>
<th>30</th>
<th>20</th>
<th>10</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Functional annotation of K12 MG1655 (%) also impacted by the degradation of the structural annotation

<table>
<thead>
<tr>
<th></th>
<th>100</th>
<th>90</th>
<th>80</th>
<th>70</th>
<th>60</th>
<th>50</th>
<th>40</th>
<th>30</th>
<th>20</th>
<th>10</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.