

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

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

▶ **To cite this version:**

Arnaud Belcour, Jeanne Got, Méziane Aite, Ludovic Delage, Jonas Collén, et al.. Inferring and comparing metabolism across heterogeneous sets of annotated genomes using AuCoMe. Genome Research, 2023, 33, pp.972 - 987. 10.1101/gr.277056.122 . hal-04192851v3

HAL Id: hal-04192851
https://hal.science/hal-04192851v3

Submitted on 31 Aug 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

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

Method

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

Arnaud Belcour,[1,4] Jeanne Got,[1,4] Méziane Aite,[1] Ludovic Delage,[2] Jonas Collén,[2] Clémence Frioux,[3] Catherine Leblanc,[2] Simon M. Dittami,[2] Samuel Blanquart,[1] Gabriel V. Markov,[2,5] and Anne Siegel[1,5]

[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, 33400 Talence, France*

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 data sets, one bacterial, one fungal, and one algal, and showed that it successfully reduces technical biases while capturing the metabolic specificities of each organism. Our results also point out shared and divergent metabolic traits among evolutionarily distant algae, underlining the potential of AuCoMe to accelerate the broad exploration of metabolic evolution across the tree of life.

[Supplemental material is available for this article.]

The comparison of genomic data gave rise to today's view of the three domains of life: bacteria, archaea, and eukaryotes, which are 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 NCBI GenBank database, which grew by 74.30% for whole-genome shotgun data in 2019 compared with 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 showed 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 with 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 exist 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, for example, 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*

[4]These authors contributed equally to this work.
[5]These authors contributed equally to this work.
Corresponding authors: arnaud.belcour@inria.fr, anne.siegel@irisa.fr
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277056.122.

© 2023 Belcour et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see https://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

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* (Herrgård 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 nonannotated genomes (Pitkänen et al. 2014; Castillo et al. 2016). The main limitation of such approaches is that the reannotation 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 the accuracy of intron–exon boundaries, to functional inferences, like the assignation to a specific catalytic activity based on previous biochemical evidence, highly depend on the amount of curation effort performed after the initial automated steps. Such valuable information is lost during a systematic reannotation 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, automated comparison of metabolism (AuCoMe), 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 that does not depend on the quality of the initial annotations and is able to aggregate heterogeneous information in both prokaryote and eukaryote data sets. 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 data sets composed of fungal, algal, and bacterial genomes. Our results show 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 panmetabolisms, 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 data sets composed of genomes that offer different levels of phylogenetic diversity. The *bacterial data set* includes 29 genomes belonging to different species of *Escherichia* and the closely related *Shigella*, the *fungal data set* (74 fungal genomes and three outgroup genomes) covers a range of different phyla within this kingdom, and, finally, the *algal data set* (36 algal genomes and four outgroup genomes) shows the highest phylogenetic diversity, including eukaryotes from the supergroups Stramenopiles, Alveolata, and Rhizaria (SAR); Haptophyta; Cryptophyta; and Archaeplastidia. For all species included in the three data sets, we used publicly available annotated genomes (see Supplemental Tables S1–S3). Run times of AuCoMe on a cluster were 7 h (10 CPUs), 25 h (40 CPUs), and 45 h (40 CPUs) for the bacterial, fungal, and algal data sets, 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 data sets show highly heterogeneous reactions (Fig. 2A,B–D, blue bars; see also Supplemental Figs. S1–S3). Notably in the fungal data set, no reactions were inferred from annotations in seven species, and 12 draft GSMNs contained fewer than 10 reactions. For the latter, their respective genome annotations included no EC number, and 11 did not include any GO term.

Similar observations were also made, although to a lesser extent, for the algal genome data set, with seven genomes having more than 2000 reactions and seven genomes having fewer 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 a homogenization of the number of reactions in the data sets (Fig. 2, orange bars; Supplemental Figs. S1–S3). The fungal data set shows an outlier at this step; the GSMN of *Encephalitozoon cuniculi* contained only 681 reactions compared with more than 1000 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 data sets, 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 with the orthology propagation, the *structural verification* step had a smaller impact on the size of the final networks (Fig. 2, green

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

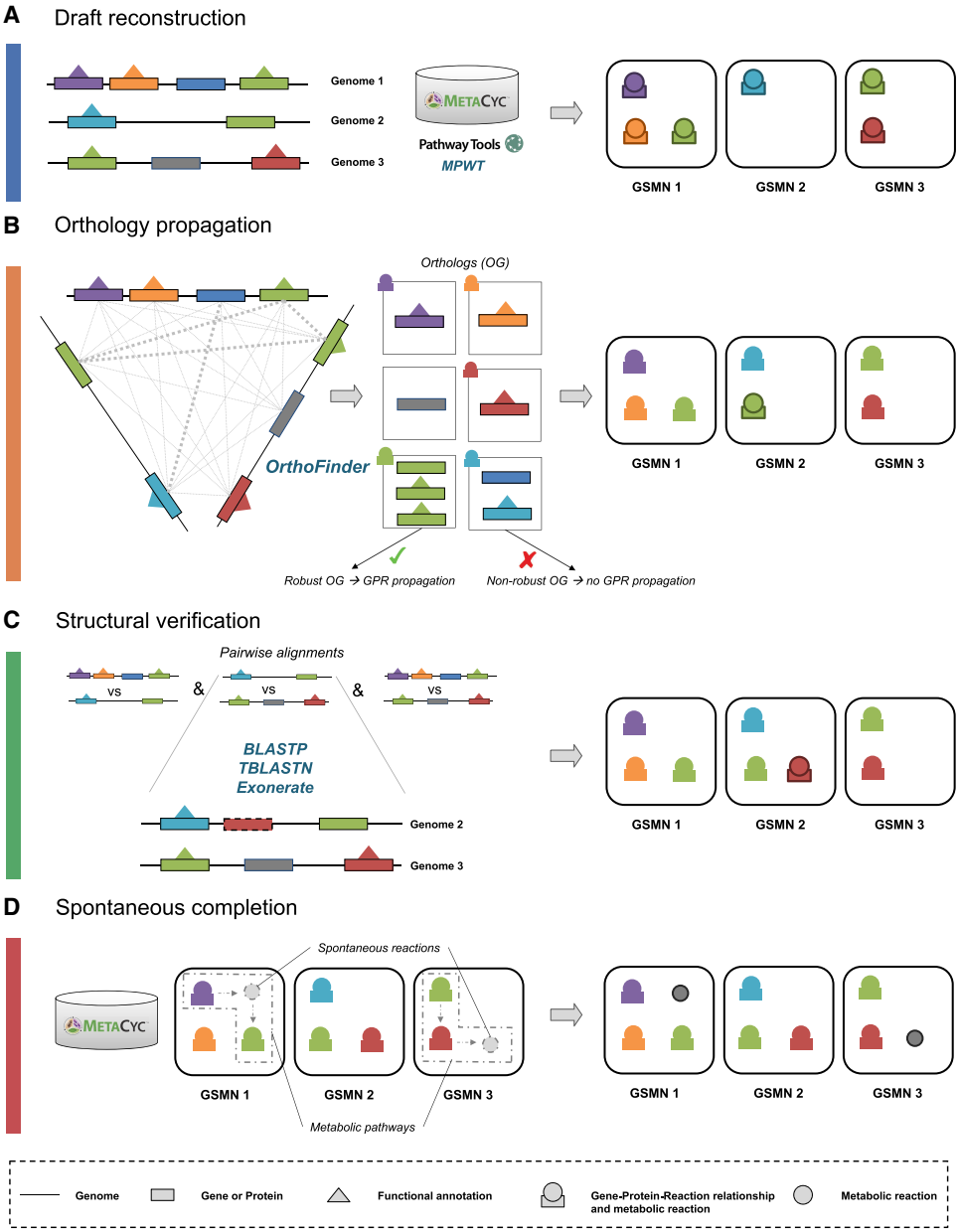


Figure 1. Reconstruction and homogenization of metabolisms with AuCoMe. Starting from a data set of partially structurally and functionally annotated genomes, AuCoMe's pipeline performs the following four steps. (*A*) Draft reconstruction. The reconstruction of draft genome-scale metabolic networks (GSMNs) 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 gene–protein–reaction (GPR) 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. (OG) Orthologs. Outlines around GPR or reaction indicate that the GPR or reaction is newly added during the corresponding step.

bars; Supplemental Figs. S1–S3). Ninety-five percent of the GSMNs received fewer than 28 reactions during this step, and the maximum was 209. In the bacterial data set, the six *Shigella* received more reactions at this step compared with 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 data set, 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 data set (Blevins et al. 2021). As for the algal data set, 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 data sets, the structural

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

A

Dataset	Bacterial Reactions Min	Bacterial Reactions Max	Bacterial Reactions Mean	Bacterial Reactions Sd	Fungal Reactions Min	Fungal Reactions Max	Fungal Reactions Mean	Fungal Reactions Sd	Algal Reactions Min	Algal Reactions Max	Algal Reactions Mean	Algal Reactions 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
Nonrobust 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			

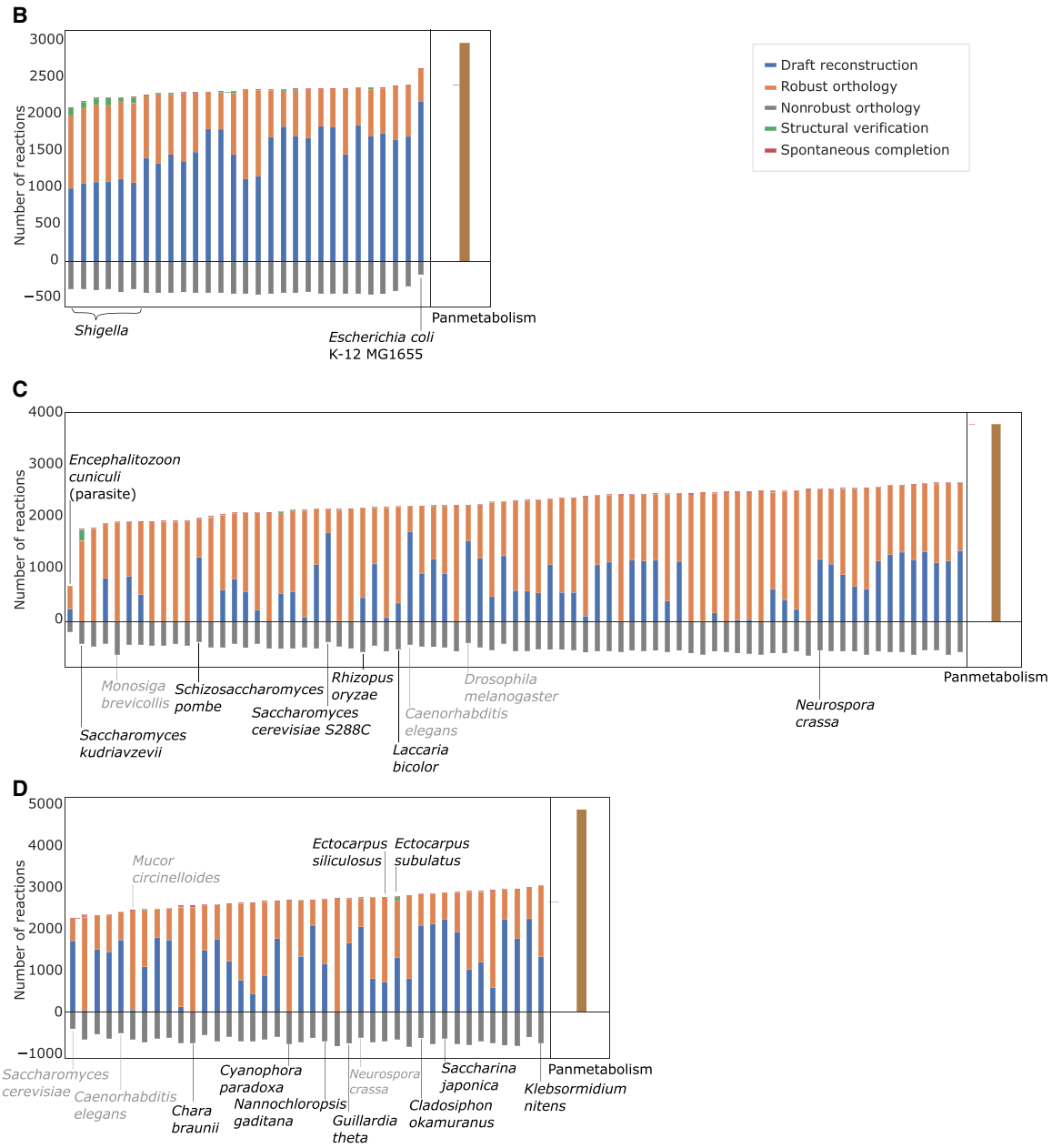


Figure 2. Application of the AuCoMe pipeline to the *bacterial*, *fungal*, and *algal* data sets of genomes. The summary table (*A*) depicts the number of reactions identified for each species at each step of the AuCoMe pipeline: reactions recovered by the *draft reconstruction* step (blue), unreliable 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 nonreliable ones. Panels *B–D* illustrate the results for each genome of the three data sets. The panmetabolism of each data set (all the reactions occurring in any of the organisms after the final step of AuCoMe) is presented in brown in *B–D*. Organisms with gray labels are outgroups. See also Supplemental Figures S1–S3.

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

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 MetaCyc pathways (Fig. 2, red bars; Supplemental Figs. S1–S3). For the fungal data set, 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 four and 36 spontaneous reactions, yielding two to 31 additional pathways. The fewer reactions that were inferred at the draft reconstruction step, the more spontaneous reactions that were added to complete pathways (Pearson's r = −0.83 and −0.84 for the fungal and algal data sets, respectively). The addition of these spontaneous reactions to the ones predicted by Pathway Tools (only other step predicting this type of reaction) led to the prediction of fewer than a hundred spontaneous reactions per GSMN.

When looking at the size of the final networks, overall, in the three data sets, the final GSMNs were of similar size after applying AuCoMe regardless of the quantity and quality of their corresponding genome annotations. In the bacterial data set (Fig. 2B; Supplemental Fig. S1), the networks of *Shigella* strains comprised fewer reactions than the rest (an average of 2148 reactions vs. 2294, Wilcoxon rank-sum test W = 138, $P = 2 \times 10^{-4}$). This is consistent with the results of Vieira et al. (2011). On the other hand, *E. coli* K–12 MG1655 stood out with 2568 reactions compared with 2047 to 2342 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 data set (Supplemental Fig. S4A–F). On this data set, 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 EcoCyc, KEGG, BiGG, and ModelSEED associated with *E. coli* K–12 MG1655 (Supplemental Fig. S5A–D). In this comparison, AuCoMe predicted the highest number of true positives (Supplemental Fig. S6).

Then, a second comparison on the eukaryotes was performed with AuCoMe, the gapseq find module, and ModelSEED on five 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; Supplemental Fig. S7). A comparison with metabolic pathways contained in YeastCyc for the genome of *S. cerevisiae S288C* was performed to estimate the quality of the predicted pathways. AuCoMe predicted a high number of pathways with a low completion rate not found in YeastCyc (Supplemental Fig. S8). For pathways with a completion rate >70%, AuCoMe and gapseq showed 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–reaction associations.

The third evaluation of the reliability of the reconstruction process was performed on the final algal data set. 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, 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 data set 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 with those discarded by the filter: 26% versus 4.8% in the fungal data set and 13.6% versus 1.4% in the algal data set (see Supplemental Fig. S10). This confirms that the robustness filter removed predominantly poorly supported reactions.

In the fifth experiment, 32 data sets were formed, each containing the 29 bacterial *E. coli* and *Shigella* strains studied by 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) (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 data sets, F-measures were computed at each AuCoMe step according to comparisons of the reconstructed GSMN with the gold-standard EcoCyc database (see Methods). Figure 3A illustrates the number of reactions predicted by AuCoMe for the *E. coli* K–12 MG1655 GSMN in each of the 32 synthetic bacterial data sets to assess the importance of each step in the homogenization of the GSMN sizes. Figure 3B represents the F-measure for the corresponding data set. As expected, the more the genomes were degraded, the lower the F-measures were. The orthology propagation alleviated this degradation for functionally degraded genomes (data sets labeled one to 10). And the structural verification step compensated the loss of annotation in structurally degraded genomes (data sets labeled 22 to 31). With both types of degradation (data sets 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 nonaltered genomes enabled the recovery of 2244 reactions (Fig. 3A, data set 31) and an F-measure of 0.60. Altogether, these results show that, by taking advantage of the annotations present in the other genomes of the considered data set, 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 in which there were clear biological expectations in the algal data set. The Calvin cycle is a biochemical pathway present in photosynthetic organisms to fix CO_2

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

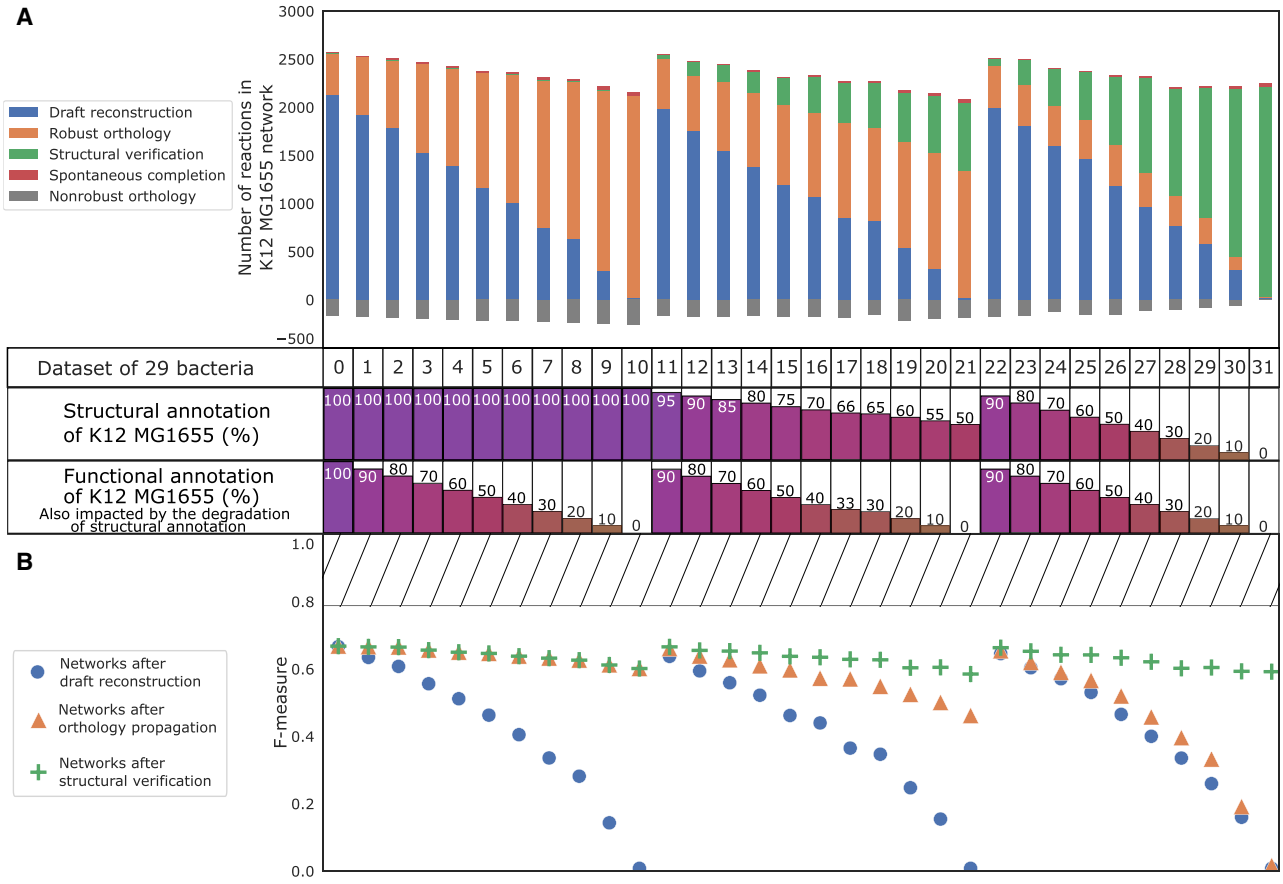


Figure 3. Efficiency of AuCoMe on degraded genome assemblies. (*A*) Number of reactions in *E. coli* K–12 MG1655 degraded networks after application of AuCoMe to 32 synthetic bacterial data sets. Each data set consists of the genome of *E. coli* K–12 MG1655, to which degradation of the functional and/or 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 data set, with the percentages of degraded annotations indicated *below*. The data set labeled "zero" 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; data sets labeled one to 10), structural annotation degradation only (*right* side; data sets labeled 22 to 31), and both degradation types (*middle*; data sets labeled 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, gray, and red color legends are as described in Fig. 2). The table shown as axis indicates the data set 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 with 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 one highlights the values of F-measure, which are unreachable because 1019 reactions in EcoCyc were not present in the panmetabolism of the 29 nondegraded bacteria.

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 (gray) applied during the orthology propagation step removed a GPR association with the reaction RIBULOSE-BISPHOSPHATE-CARBOXYLASE-RXN for the nonphotosynthetic fungus *Neurospora crassa*. The structural verification step added one reaction (RIBULP3EPIM-RXN) for *Porphyra umbilicalis* (Fig. 4, green square). 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, was manually added to the 40 GSMNs (for GPR associations, see Fig. 4, yellow; for details, see Supplemental 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 co-option 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.

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

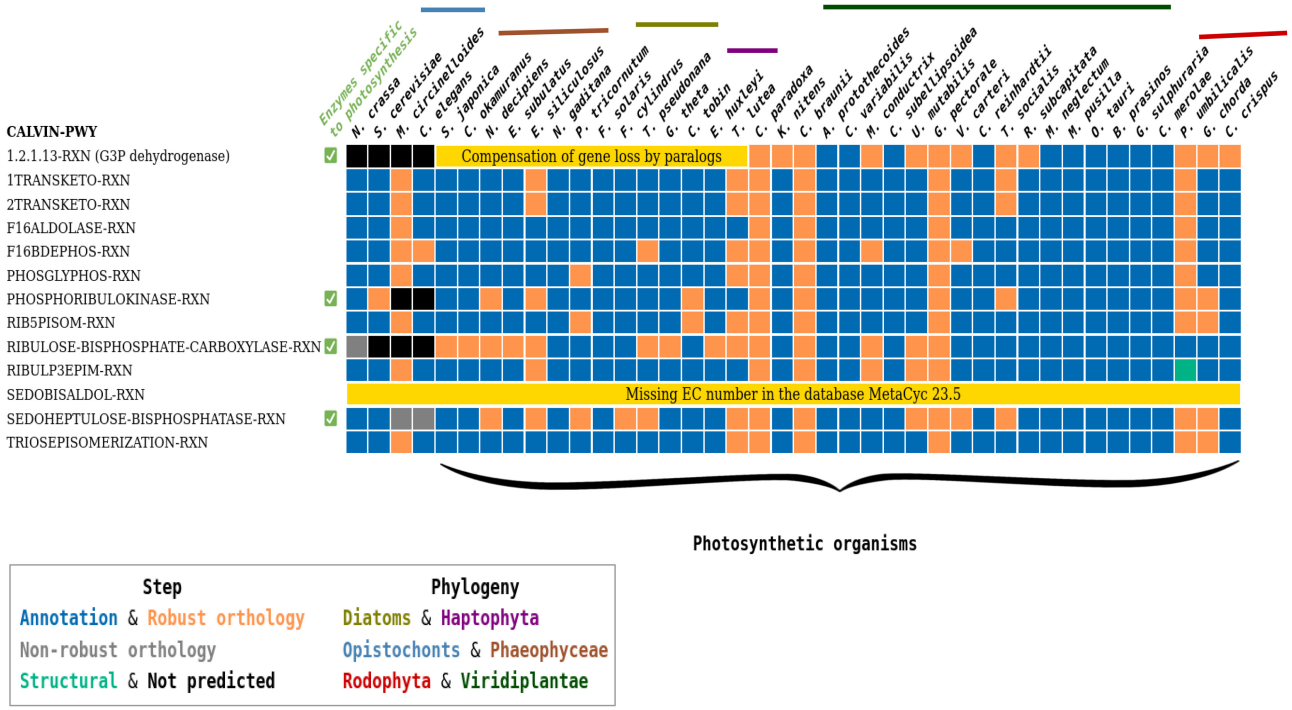


Figure 4. AuCoMe results on the Calvin cycle pathway in the algal data set. AuCoMe was applied to the data set of 36 algae and four outgroup species (columns). Each row represents a MetaCyc reaction of the pathway; the table shows whether it is predicted by AuCoMe: blue, draft reconstruction; orange,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

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 data set 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, B). We compared these clusterings with a phylogeny compiled from Strassert et al. (2021). The initial GSMNs produced from the annotations showed low consistency with the phylogenetic relationships (Fig. 5A). 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 (Fig. 5B), we observed a clear separation between the known phylogenetic groups, supported by an ANOSIM test ($R = 0.811$, P-value $= 1 \times 10^{-4}$). 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. 5C,D). There were only three higher-order inconsistencies concerning *Cyanophora paradoxa*, for which the genome version deposited in GenBank fully lacked expert annotations (Price et al. 2012); *Guillardia theta*, which belongs to the cryptophytes, for which the phylogenetic position is controversial (Strassert et al. 2021); and *Nannochloropsis 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 1675 GPRs, a number within the same range as the other species from the data set. 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 14 reactions present in *C. okamuranus* and *S. japonica* but absent in other brown algae (see Supplemental Table S8). Among those 14 reactions, 12 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

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

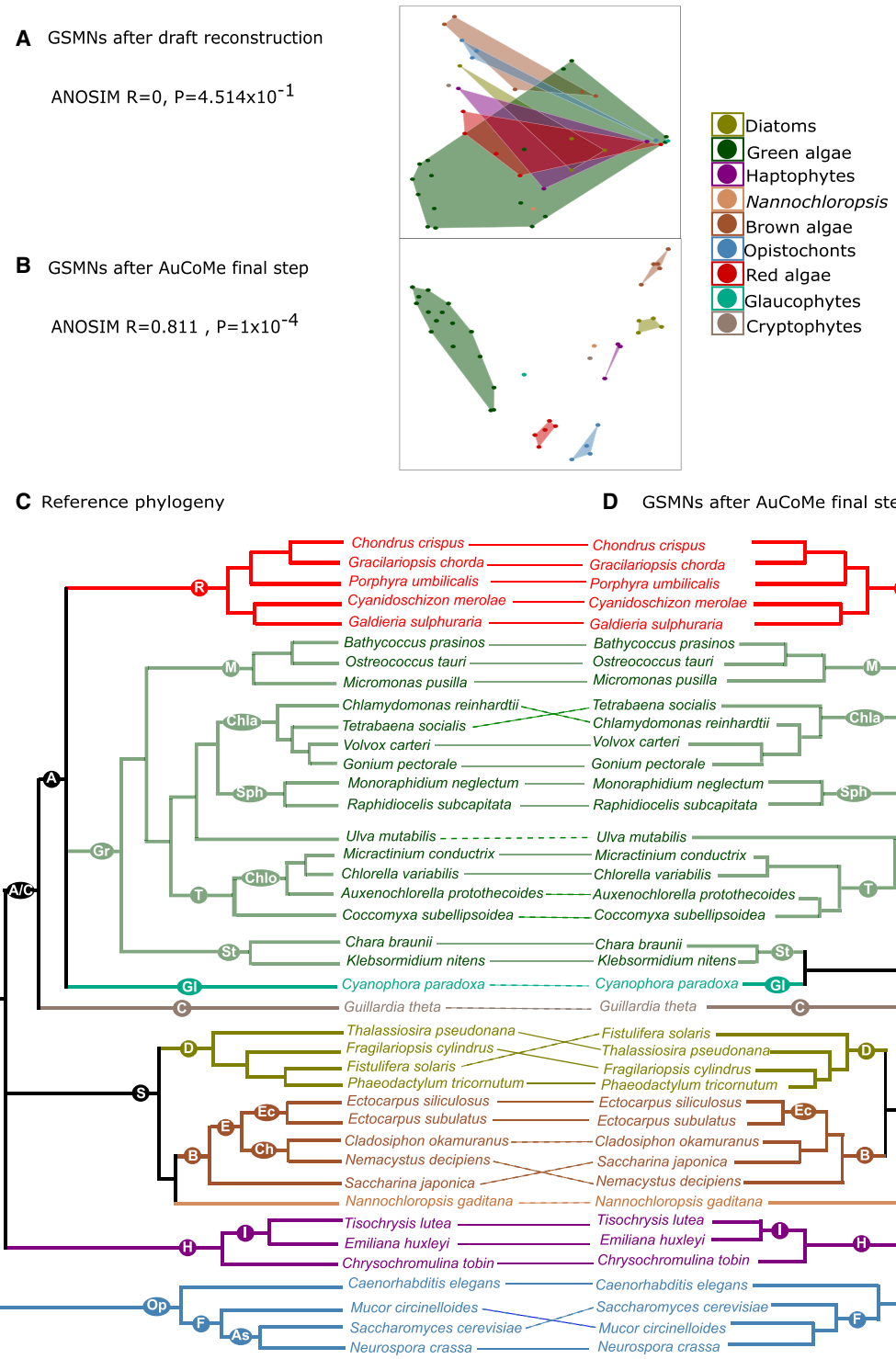


Figure 5. AuCoMe as a tool to improve taxonomic consistency of GSMNs. (*A*,*B*) MDS plots for GSMNs calculated with the AuCoMe draft reconstruction step (*A*) or after all AuCoMe steps (*B*). In both cases, ANOSIM values are indicated *below* (MDS and ANOSIM were computed using the vegan package [https://github.com/vegandevs/vegan] with R 4.1.2 [R Core Team 2023]). (*C*,*D*) Tanglegram evaluating the taxonomic consistency between reference phylogeny, compiled from Strassert et al. (2021) (*C*) with AuCoMe dendrograms based on metabolic distances using the pvclust package version 2.0.0 (Suzuki and Shimodaira 2006) with R 3.4.4 (R Core Team 2023) with the Jaccard distance (*D*). 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) glaucophytes, (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.

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

proteins present in the genome of *E. siliculosus* are predicted to be cytoplasmic or 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. 5D), we noticed that it grouped with the small unicellular alga *G. theta*, which belongs to the cryptophytes, usually grouping with the archeplastids (Fig. 5C) 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). We focused on reactions that distinguished at least two of these groups, namely, 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 *G. theta* with the three groups, focusing on pathways that were >50% complete. The metabolic networks showed, for instance, that *G. theta* (1) possesses, like haptophytes in our data set, parts of the mitochondrial L-carnitine shuttle pathway, (2) comprises, like the Stramenopiles, the complete pathway of glycine betaine synthesis, and (3) can synthesize, like terrestrial plants, 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 (Frioux et al. 2018; Machado 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 as 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 *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 enabling an unbiased metabolic comparison between species.

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

Running AuCoMe on the bacterial data set highlighted the impact of a single highly annotated genome on metabolic inference. This data set included a single well-annotated reference genome of the *E. coli* K–12 MG1655 strain, which caused a number of reactions initially propagated by orthology from the *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. Although the relevance of the filter was shown on the algal data set by avoiding the propagation of annotations related to photosynthesis to nonphotosynthetic 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 *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 Figure S4, which confirms that AuCoMe GSMNs have, by design, no reaction lacking gene association.

A typical use for GSMNs is their simulation, generally with flux-based approaches. As AuCoMe performs a 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 eukaryotes and their strong dependency on their environment make 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 from comparison of metabolic networks across species

Evolution

Our examples of the Calvin cycle and phycobiliprotein synthesis show 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 dendrogram calculated from the final AuCoMe reconstruction is 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 *G. theta*. This species has several potentially plesiomorphic metabolic traits in common with other marine lineages, which 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, 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 shown above, phylogenetic signals in metabolism are stronger than the adaptive signals we can expect, this approach would also benefit from the development or adaptation of statistical models that could help detect signals of adaptation in an overall noisy data set. Such models exist, for instance, to detect selective signatures in the evolution of the 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

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

from the phylogenetic signal will be reduced owing to the short evolutionary time since the separation. With such data sets, 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 data set 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 grixazone biosynthesis, that is, 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 they 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 data set* includes the 29 bacterial *E. coli* and *Shigella* strains studied previously (Vieira et al. 2011), downloaded from public databases (see Supplemental Table S1).

The *fungal data set* includes 74 fungal genomes which were selected according to the method of Wang et al. (2009) as representative of the fungal diversity, together with three 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 data set* contains 36 algal genomes selected to represent a wide diversity of photosynthetic eukaryotes and downloaded from public databases. The data set includes 16 Viridiplantae (green algae), five Phaeophyceae (brown algae), five Rhodophyceae (red algae), four diatoms, three haptophytes, one cryptophyte (*G. theta*), one Eustigmatophyceae (*N. gaditana*), and one Glaucophyceae (*C. paradoxa*). The genomes of *C. elegans* (Witting et al. 2018), *Mucor 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 data sets was curated manually in order to make it compatible with Pathway Tools v23.5. Curated genomes are available at Zenodo (https://doi.org/10.5281/zenodo.7851053).

AuCoMe, a method to reconstruct GSMNs homogenized across related species

AuCoMe is a Python package implementing a pipeline whose steps are described in Figure 1. The method aims at producing homogenized 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 (Aite et al. 2018; Hucka 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 (Belcour et al. 2020a), AuCoMe launches parallel processes of the PathoLogic algorithm of Pathway Tools (Karp et al. 2019). Pathway Tools creates Pathway/Genome Databases (PGDBs) for all genomes. The resulting PGDBs are converted into PADMET and SBML files (Hucka et al. 2003, 2018) using the PADMet package (Aite et al. 2018). During this conversion, 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 spontaneous reactions. For example, in Figure 1A, the draft reconstruction step generates six GPRs in total for the three considered genomes.

The `aucome orthology` command runs the orthology propagation step, which complements the previous GSMNs with GPR associations whose genes are predicted to be orthologs to genes from GPR relations of other GSMNs of the data set (Fig. 1B). To that purpose, the pipeline relies on OrthoFinder (Emms and Kelly 2015, 2019) 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 GPR

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

associations between the orthologs (see below). Nonrobust GPR associations are not integrated in the final GSMNs. In the example shown in Figure 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 `aucome structural` command runs the structural verification step to identify GPRs associated with missing structural annotations of the input genomes. This pipeline step complements GSMNs with GPR associations from other GSMNs according to protein-against-genome alignment criteria. This enables the identification of reactions that are associated with gene sequences absent from the initial structural annotations of the input genomes. A pairwise comparison of the reactions in the GSMNs produced during the previous step is performed (Fig. 1C). In this comparison, if a reaction is missing in an organism, a structural verification will be performed. 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 $<1 \times 10^{-20}$) 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 Figure 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 Figure 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 panmetabolic network after the orthology propagation was defined as illustrated in Algorithm 1 and detailed in the following. We denote by $org(g)$ the organism of a gene g. For every pair of genes $g1$, $g2$ of two different organisms, we denote $orth(g1, g2) = 1$ if the genes are predicted to be orthologs. We denote by $association(r, g) = 1$ a GPR association between a reaction r and a gene g that is predicted by the AuCoMe algorithm. When the gene association is predicted by the draft reconstruction step, we denote $annot_type(r, g) = 1$ (and zero otherwise). When the gene association is predicted according to orthology criteria, we denote $ortho_type(r, g) = 1$ (and zero otherwise).

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

The robustness assessment of a GPR between r and g propagated by orthology (L7, Algorithm 1) distinguishes two scenarios. In the first scenario, g belongs to an orthology cluster that is supported by at least two annotations. Formally, this means that there exist two genes, $g1$ to $g2$, both orthologs to g, such that $annot_type(r, g1) = 1$ and $annot_type(r, g2) = 1$. The presence of these genes leads us to consider g robustly associated with r (L8–L9, Algorithm 1).

In the second scenario, the GPR association between r and g was propagated from a unique gene $g1$ with $annot_type(r, g1) = 1$ in the orthology cluster (L11, Algorithm 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, $robust(r, g) = 1$ if $N_prop(r, g1) \leq \lceil robust_func(N_org(r) - 1) \times (N_org(r) - 1) \rceil$ (L12–L13, Algorithm 1). The robustness function $robust_func^{(t)}(x) = \min\left(1, \frac{1}{x}\max\left(\lceil tx \rceil, \lceil \frac{5}{x} \rceil\right)\right)$ was chosen such that it is one for low values of N_org and then decreases to a threshold value (by default $t = 0.05$) for large values of N_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 Supplemental Fig. S15.

Algorithm 1. Robustness criterion algorithm

```
1:  for r in panmetabolism do
2:  N_org(r) ← # {org(g), ∃g1 s.t. association(r, g) = 1} ▷ Number of
    organisms with GPR 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) ← # {org(g1), ∃g1 s.t. org(g1) ≠ org(g),
      orth(g, g1) = 1, association(r, g1) = 1} ▷ Number of organisms to
      which the GPR has been propagated
6:  end for
7:  for all genes g s.t. annot_type(r; g) = 0 and orth_type(r; g) = 1 do
    ▷ Restrict the family of gene candidates to be associated with a
    new reaction
8:    if ∃g1, g2 s.t. orth(g, g1) = orth(g, g2) = 1 and
      annot_type(r, g1) = annot_type(r, g2) = 1 then ▷ At least two
      annotations support the GPR relation
9:      robust(r, g) = 1
10:   else ▷ Prevent the propagation of an isolated annotation to too
      many organisms
11:     g1 ← unique gene s.t. orth(g,g1) = 1 and
        annot_type(r; g1) = 1
12:     if N_prop(r, g1) ≤ robust_func(N_org(r) − 1) × (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
```

Validation of AuCoMe predictions

A first experiment was performed on the bacterial data set, for which we reconstructed the metabolic networks (29 bacteria containing strains of *E. 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"

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

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 four databases (KEGG, EcoCyc, ModelSEED, and BiGG). For more information on the reference EC catalog, see the Supplemental File (section "Methods").

A second comparison was made on the eukaryotes and especially the fungal data set (using five organisms: *Laccaria bicolor*, *N. crassa*, *Rhizopus oryzae*, *S. 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 *S. 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, 100 random GPR associations were randomly selected and examined across the metabolic networks generated by AuCoMe for the algal data set. Among them were 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 with 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 analyze 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 data sets. 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 nondegraded *E. coli* K–12 MG1655 GSMN. Two degradation types were simulated: (1) a degradation of the functional annotations of the genes, in which all the annotations like GO terms, EC numbers, gene names, etc., associated with a reaction were removed, and (2) a degradation of the structural annotation of the genes, in which 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 Supplemental 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 data sets, plus the data set with nondegraded *E. coli* K–12 MG1655 genome, which was called data set 0. Their characteristics are detailed in Supplemental Table S4. For each *E. coli* K–12 MG1655 replicate in a data set, AuCoMe produced a GSMN, which was compared with 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 the Supplemental File (section "Methods").

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

A data set 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 performed 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 performed using the 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).

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 GitHub (https://github.com/AuReMe/aucome) and as Supplemental Code.

Running AuCoMe on the data sets 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), and 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 GitHub (https://github.com/AuReMe/aucome/blob/master/recipes/Dockerfile). A version of AuCoMe, PADMet source code, and the scripts used to run some figures is available at Zenodo (https://doi.org/10.5281/zenodo.7752449) and as Supplemental Files S1–S14.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank the GenOuest bioinformatics core facility (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

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

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: A.B. was responsible for conceptualization, data curation, methodology, formal analysis, software, validation, visualization, writing the original draft, and reviewing and editing. J.G. was responsible for data curation, formal analysis, resources, software, validation, visualization, writing the original draft, and reviewing and editing. M.A. was responsible for conceptualization, data curation, methodology, and software. L.D. was responsible for formal analysis, validation, writing the original draft, and reviewing and editing. J.C. was responsible for formal analysis, validation, and reviewing and editing. C.F. was responsible for methodology, software, visualization, writing the original draft, and reviewing and editing. C.L. was responsible for funding acquisition, writing the original draft, and reviewing and editing. S.M.D. was responsible for conceptualization, data curation, formal analysis, funding acquisition, methodology, validation, writing the original draft, and reviewing and editing. S.B. was responsible for conceptualization, methodology, writing the original draft, and reviewing and editing. G.V.M. was responsible for data curation, formal analysis, methodology, supervision, validation, visualization, writing the original draft, and reviewing and editing. A.S. was responsible for conceptualization, formal analysis, funding acquisition, methodology, supervision, writing the original draft, and reviewing and editing.

References

Aite M, Chevallier M, Frioux C, Trottier C, Got J, Cortés MP, Mendoza SN, Carrier G, Dameron O, Guillaudeux N, et al. 2018. Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. *PLoS Comput Biol* **14:** e1006146. doi:10.1371/journal.pcbi.1006146

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410. doi:10.1016/S0022-2836(05)80360-2

Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, et al. 2018. KBase: the United States department of energy systems biology knowledgebase. *Nat Biotechnol* **36:** 566–569. doi:10.1038/nbt.4163

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9:** 75. doi:10.1186/1471-2164-9-75

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

Bauer E, Laczny CC, Magnusdottir S, Wilmes P, Thiele I. 2015. Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome* **3:** 55. doi:10.1186/s40168-015-0121-6

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

Belcour A, Frioux C, Aite M, Bretaudeau A, Hildebrand F, Siegel A. 2020a. Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species. *eLife* **9:** e61968. doi:10.7554/eLife.61968

Belcour A, Girard J, Aite M, Delage L, Trottier C, Marteau C, Leroux C, Dittami SM, Sauleau P, Corre E, et al. 2020b. Inferring biochemical reactions and metabolite structures to understand metabolic pathway drift. *iScience* **23:** 100849. doi:10.1016/j.isci.2020.100849

Bernard MS, Strittmatter M, Murúa P, Heesch S, Cho GY, Leblanc C, Peters AF. 2019. Diversity, biogeography and host specificity of kelp endophytes with a focus on the genera *Laminarionema* and *Laminariocolax* (Ectocarpales, Phaeophyceae). *Eur J Phycol* **54:** 39–51. doi:10.1080/09670262.2018.1502816

Bernstein DB, Sulheim S, Almaas E, Segrè D. 2021. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biol* **22:** 64. doi:10.1186/s13059-021-02289-z

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

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

Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, et al. 2015. *RASTtk:* a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep* **5:** 8365. doi:10.1038/srep08365

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

Burgunter-Delamare B, KleinJan H, Frioux C, Fremy E, Wagner M, Corre E, Le Salver A, Leroux C, Leblanc C, Boyen C, et al. 2020. Metabolic complementarity between a brown alga and associated cultivable bacteria provide indications of beneficial interactions. *Front Mar Sci* **7:** 85. doi:10.3389/fmars.2020.00085

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

Burki F, Roger AJ, Brown MW, Simpson AG. 2020. The new tree of eukaryotes. *Trends Ecol Evol (Amst)* **35:** 43–55. doi:10.1016/j.tree.2019.08.008

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

Capela J, Lagoa D, Rodrigues R, Cunha E, Cruz F, Barbosa A, Bastos J, Lima D, Ferreira EC, Rocha M, et al. 2022. merlin, an improved framework for the reconstruction of high-quality genome-scale metabolic models. *Nucleic Acids Res* **50:** 6052–6066. doi:10.1093/nar/gkac459

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972–1973. doi:10.1093/bioinformatics/btp348

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

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

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

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

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

Correia K, Mahadevan R. 2020. Pan-genome-scale network reconstruction: Harnessing phylogenomics increases the quantity and quality of metabolic models. *Biotechnol J* **15:** 1900519. doi:10.1002/biot.201900519

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

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

Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C. 2013. Automated genome annotation and metabolic model reconstruction in the SEED and model SEED. *Methods Mol Biol* **985:** 17–45. doi:10.1007/978-1-62703-299-5_2

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

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

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

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc Natl Acad Sci* **115:** E409–E417. doi:10.1073/pnas.1707515115

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16:** 157. doi:10.1186/s13059-015-0721-2

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20:** 238. doi:10.1186/s13059-019-1832-y

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

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

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

Giannakou K, Cotterrell M, Delneri D. 2020. Genomic adaptation of *Saccharomyces* species to industrial environments. *Front Genet* **11**. doi:10.3389/fgene.2020.00916

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

Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. 2019. Current status and applications of genome-scale metabolic models. *Genome Biol* **20:** 121. doi:10.1186/s13059-019-1730-3

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59:** 307–321. doi:10.1093/sysbio/syq010

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* **28:** 977–982. doi:10.1038/nbt.1672

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

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

Hucka M, Bergmann FT, Dräger A, Hoops S, Keating SM, Le Novère N, Myers CJ, Olivier BG, Sahle S, Schaff JC, et al. 2018. The systems biology markup language (SBML): language specification for level 3 version 2 core. *J Integr Bioinform* **15:** 20170081. doi:10.1515/jib-2017-0081

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

Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26:** 1669–1670. doi:10.1093/bioinformatics/btq243

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28:** 27–30. doi:10.1093/nar/28.1.27

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45:** D353–D361. doi:10.1093/nar/gkw1092

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

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

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

Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S. 2002. The EcoCyc database. *Nucleic Acids Res* **30:** 56–58. doi:10.1093/nar/30.1.56

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

Karp PD, Weaver D, Latendresse M. 2018b. How accurate is automated gap filling of metabolic models? *BMC Syst Biol* **12:** 73. doi:10.1186/s12918-018-0593-7

Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, Ong WK, Subhraveti P, Caspi R, Fulcher C, et al. 2019. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinformatics* **22:** 109–126. doi:10.1093/bib/bbz104

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30:** 772–780. doi:10.1093/molbev/mst010

Keseler IM, Gama-Castro S, Mackie A, Billington R, Bonavides-Martínez C, Caspi R, Kothari A, Krummenacker M, Midford PE, Muñiz-Rascado L, et al. 2021. The EcoCyc database in 2021. *Front Microbiol* **12:** 2098. doi:10.3389/fmicb.2021.711077

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

Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, et al. 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res* **44:** D73–D80. doi:10.1093/nar/gkv1226

Latendresse M, Karp PD. 2018. Evaluation of reaction gap-filling accuracy by randomization. *BMC Bioinformatics* **19:** 53. doi:10.1186/s12859-018-2050-4

Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol* **32:** 2798–2800. doi:10.1093/molbev/msv150

Lefort V, Longueville JE, Gascuel O. 2017. SMS: smart model selection in PhyML. *Mol Biol Evol* **34:** 2422–2424. doi:10.1093/molbev/msx149

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

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

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

Liaud MF, Brandt U, Scherzinger M, Cerff R. 1997. Evolutionary origin of cryptomonad microalgae: two novel chloroplast/cytosol-specific GAPDH genes as potential markers of ancestral endosymbiont and host cell components. *J Mol Evol* **44:** S28–S37. doi:10.1007/PL00000050

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

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

Manandhar B, Wagle A, Seong SH, Paudel P, Kim HR, Jung HA, Choi JS. 2019. Phlorotannins with potential anti-tyrosinase and antioxidant activity isolated from the marine seaweed *Ecklonia stolonifera*. *Antioxidants (Basel)* **8:** 240. doi:10.3390/antiox8080240

Marcellin-Gros R, Piganeau G, Stien D. 2020. Metabolomic insights into marine phytoplankton diversity. *Mar Drugs* **18:** 78. doi:10.3390/md18020078

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

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

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, et al. 2014. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res* **42:** D206–D214. doi:10.1093/nar/gkt1226

Pitkänen E, Jouhten P, Hou J, Syed MF, Blomberg P, Kludas J, Oja M, Holm L, Penttilä M, Rousu J, et al. 2014. Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. *PLoS Comput Biol* **10:** e1003465. doi:10.1371/journal.pcbi.1003465

Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber APM, Schwacke R, Gross J, Blouin NA, Lane C, et al. 2012. *Cyanophora paradoxa* genome

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press

elucidates origin of photosynthesis in algae and plants. *Science* **335:** 843–847. doi:10.1126/science.1213561
Prigent S, Nielsen JC, Frisvad JC, Nielsen J. 2018. Reconstruction of 24 *Penicillium* genome-scale metabolic models shows diversity based on their secondary metabolism. *Biotechnol Bioeng* **115:** 2604–2612. doi:10.1002/bit.26739
Psomopoulos FE, van Helden J, Médigue C, Chasapi A, Ouzounis CA. 2020. Ancestral state reconstruction of metabolic pathways across pangenome ensembles. *Microb Genom* **6:** mgen000429. doi:10.1099/mgen.0.000429
R Core Team. 2023. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
Rockwell NC, Lagarias JC. 2017. Ferredoxin-dependent bilin reductases in eukaryotic algae: ubiquity and diversity. *J Plant Physiol* **217:** 57–67. doi:10.1016/j.jplph.2017.05.022
Ryu JY, Kim HU, Lee SY. 2019. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc Natl Acad Sci* **116:** 13996–14001. doi:10.1073/pnas.1821905116
Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2019. GenBank. *Nucleic Acids Res* **47:** D94–D99. doi:10.1093/nar/gky989
Schulz C, 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, Beber M, et al. 2021. The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res* **49:** D575–D588. doi:10.1093/nar/gkaa746
Shapiro BJ, Alm EJ. 2008. Comparing patterns of natural selection across species using selective signatures. *PLoS Genet* **4:** e23. doi:10.1371/journal.pgen.0040023
Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31. doi:10.1186/1471-2105-6-31
Stamboulian M, Guerrero RF, Hahn MW, Radivojac P. 2020. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. *Bioinformatics* **36:** i219–i226. doi:10.1093/bioinformatics/btaa468
Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35:** 1026–1028. doi:10.1038/nbt.3988
Strassert JFH, Irisarri I, Williams TA, Burki F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat Commun* **12:** 1879. doi:10.1038/s41467-021-22044
Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22:** 1540–1542. doi:10.1093/bioinformatics/btl117
Suzuki H, Furusho Y, Higashi T, Ohnishi Y, Horinouchi S. 2006. A novel *o*-aminophenol oxidase responsible for formation of the phenoxazinone chromophore of grixazone. *J Biol Chem* **281:** 824–833. doi:10.1074/jbc.M505806200
Tamura Y, Takenaka S, Sugiyama S, Nakayama R. 1998. Occurrence of anserine as an antioxidative dipeptide in a red alga, *Porphyra yezoensis*. *Biosci Biotechnol Biochem* **62:** 561–563. doi:10.1271/bbb.62.561
Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Mol Biol Evol* **38:** 3022–3027. doi:10.1093/molbev/msab120
Thiele I, Palsson BØ. 2010. Reconstruction annotation jamborees: a community approach to systems biology. *Mol Syst Biol* **6:** 361. doi:10.1038/msb.2010.15
Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, Bazzani S, Charusanti P, Chen FC, Fleming RMT, Hsiung CA, et al. 2011. A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella* Typhimurium LT2. *BMC Syst Biol* **5:** 8. doi:10.1186/1752-0509-5-8
Vieira G, Sabarly V, Bourguignon PY, Durot M, Le Fèvre F, Mornico D, Vallenet D, Bouvet O, Denamur E, Schachter V, et al. 2011. Core and panmetabolism in *Escherichia coli*. *J Bacteriol* **193:** 1461–1472. doi:10.1128/JB.01192-10
Vongsangnak W, Klanchui A, Tawornsamretkit I, Tatiyaborwornchai W, Laoteng K, Meechai A. 2016. Genome-scale metabolic modeling of *Mucor circinelloides* and comparative analysis with other oleaginous species. *Gene* **583:** 121–129. doi:10.1016/j.gene.2016.02.028]
Wang H, Xu Z, Gao L, Hao B. 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol* **9:** 195. doi:10.1186/1471-2148-9-195
Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, Nielsen J, Kerkhoven EJ. 2018. RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*. *PLoS Comput Biol* **14:** e1006541. doi:10.1371/journal.pcbi.1006541
Witting M, Hastings J, Rodriguez N, Joshi CJ, Hattwell JPN, Ebert PR, van Weeghel M, Gao AW, Wakelam MJO, Houtkooper RH, et al. 2018. Modeling meets metabolomics: the WormJam consensus model as basis for metabolic studies in the model organism *Caenorhabditis elegans*. *Front Mol Biosci* **5:** 96. doi:10.3389/fmolb.2018.00096
Xu S, Wang J, Guo Z, He Z, Shi S. 2020. Genomic convergence in the adaptation to extreme environments. *Plant Commun* **1:** 100117. doi:10.1016/j.xplc.2020.100117
Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13:** 329–342. doi:10.1038/nrg3174
Young AD, Gillung JP. 2020. Phylogenomics: principles, opportunities and pitfalls of big-data phylogenetics. *Syst Entomol* **45:** 225–247. doi:10.1111/syen.12406
Zimmermann J, Kaleta C, Waschina S. 2021. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol* **22:** 81. doi:10.1186/s13059-021-02295-1

Received June 22, 2022; accepted in revised form May 23, 2023.

Downloaded from genome.cshlp.org on August 7, 2023 - Published by Cold Spring Harbor Laboratory Press



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

Arnaud Belcour, Jeanne Got, Méziane Aite, et al.

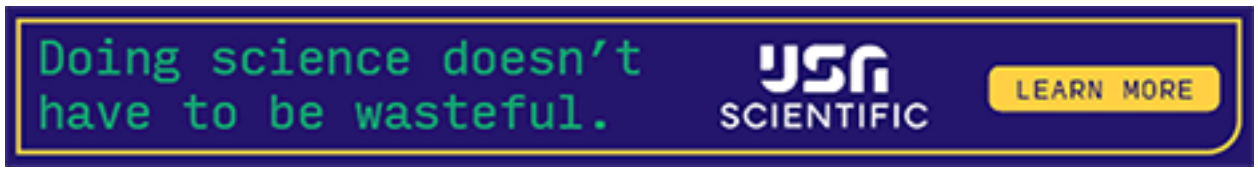
Genome Res. published online July 19, 2023
Access the most recent version at doi:10.1101/gr.277056.122

Supplemental Material http://genome.cshlp.org/content/suppl/2023/07/14/gr.277056.122.DC1

P<P Published online July 19, 2023 in advance of the print journal.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see https://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

© 2023 Belcour et al.; Published by Cold Spring Harbor Laboratory Press