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

Arnaud Belcour [1], Jeanne Got [1], 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,†] and Anne Siegel [1]

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

Table of Contents

2.3 **Comparison with gapseq, ModelSEED and CarveMe on a bacterial dataset**

- **Supplemental_Fig_S4** Methods comparison on the 29 bacterial species dataset.
- **Supplemental_Fig_S5** Distribution of EC in reference EC catalog for *E. coli* K-12 MG1655.
- **Supplemental_Fig_S6** Comparison between the metabolic networks inferred for *E. coli* K–12 MG1655 by 4 tools (AuCoMe, CarveMe, gapseq and ModelSEED) and the reference set of EC numbers from 4 databases (EcoCyc, KEGG, BiGG and ModelSEED).

2.4 **Comparison with gapseq and ModelSEED and on a fungal dataset**

- **Supplemental_Table_S5** Comparison between the predictions made by AuCoMe, gapseq and ModelSEED on 5 fungi.
- **Supplemental_Fig_S7** Metabolic pathways comparison between AuCoMe and gapseq for five fungal species.
- **Supplemental_Fig_S8** Number of pathways common or specific to YeastCyc and AuCoMe with their completeness ratio predicted by AuCoMe.
- **Supplemental_Fig_S9** Comparison of the metabolic pathways in YeastCyc, predicted by gapseq or AuCoMe.

2.5 **Validation of filtering steps and GPR associations**

- **Supplemental_Table_S6** Manual validation of 50 randomly chosen reactions found in any of the species.
- **Supplemental_Table_S7** Manual validation of 50 reactions absent from a species and randomly chosen.

2.6 **Validation of EC numbers with deep-learning approaches**

- **Supplemental_Fig_S10** Average number (among the fungal and the algal datasets) of EC numbers for GPR associations predicted by the orthology step of AuCoMe, and average number of EC numbers predicted by the DeepEC tool for the corresponding gene families.

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

- **Supplemental_Fig_S11** Prediction of pigment pathways in brown algae.

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

- **Supplemental_Fig_S12** Presence and absence of reactions in the pan-metabolism of the algal dataset.
- **Supplemental_Fig_S13** Phylogenetic distribution of brown algal o-Aminophenol Oxidases.
- **Supplemental_Table_S8** Reactions common to *Saccharina japonica* and *Cladosiphon okamuranus* but not found in other brown algae.
- **Supplemental_Table_S9** Additional homologs in *E. siliculosus* found by BLASTP searches for sequences inferred to be present only in *C. okamuranus* and *S. japonica*.
- **Supplemental_Table_S10** Additional o-aminophenol oxidases from *E. siliculosus* and their homologs in other Stramenopiles
- **Supplemental_Table_S11** Reactions distinguishing the cryptophyte, haptophyte, stramenopile, and archeplastida groups.
- **Supplemental_Table_S12** Shared metabolic pathways as well as the absence of pathways between chryptophytes, haptophytes, stramenopiles, and archaeplastida.

3. **Methods**

3.1 **Robustness criteria applied to a toy example**

- **Supplemental_Fig_S14** Robustness function used to identify robust gene-reaction association according to the number of organisms present in the graph of genes associated with a reaction.
- **Supplemental_Fig_S15** Application of the robustness criteria to gene-reaction associations predicted by the draft reconstruction and the orthology propagation steps of AuCoMe.

3.2 **Comparison to EcoCyc**

3.3 **Degradation of the *E. coli* K–12 MG1655 genome**

- **Algorithm 1** Degradation of K12MG1655 genome

3.4 **Availability of version 23.5 of the Pathway Tools software**

4. **Content of the Supplemental Files archive**

1 Datasets

The characteristics of the bacterial, fungal, and algal datasets are shown in tables S1, S2, S3. The characteristics of degraded genome of *Escherichia coli* str. K-12 substr. MG1655 strain and of the 32 in silico bacterial datasets are shown in Table S4.

2 Results

2.1 Running times of the AuCoMe pipeline

The bacterial dataset experimentations were run on a cluster with 10 CPUs and 20 GB of memory. The measured runtimes were the following: 35 mn for draft reconstruction, 3h 50mn for orthology propagation, 3 h for structural verification and 4 mn for spontaneous reaction completion.

The fungal dataset experimentations were run in a Go-Docker environment with 180 GB of RAM and 40 CPUs (`http://www.genouest.org/godocker/`), i.e., a computer cluster environment with a batch scheduling that manages Docker jobs. The annotation-based GSMN reconstruction step ran in 31mn 44s, the multiple orthology propagation step took 3h 26mn 38s, the structural step ran for 20h 52mn51s. Finally, the spontaneous-reaction step took 10mn 11s.

The algal dataset experimentations were run on the same Go-Docker environment with 180 GB of RAM and 40 CPUs. The annotation-based GSMNs reconstruction step ran in 2h 04mn 21s, the multiple orthology propagation step took 2h 20mn 35s, the structural step run during 40h 52mn 30s. Finally, the spontaneous-reaction step took 8mn 1s.

Strain	Assembly	Genome Assembly used	Link
Escherichia coli str. K-12 substr. MG1655	GCF_000005845.2	ASM584v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000005845.2
Escherichia coli str. K-12 substr. W3110	GCF_000010245.2	ASM1024v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000010245.2
Escherichia coli IAI1	GCF_000026265.1	ASM2626v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000026265.1
Escherichia coli 55989	GCF_000026245.1	ASM2624v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000026245.1
Shigella boydii Sb227	GCF_000012025.1	CP000036.1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000012025.1
Shigella sonnei Ss046	GCF_000092525.1	CP000038.1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000092525.1
Shigella flexneri 2a str. 301	GCF_000006925.2	AE005674.2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000006925.2
Shigella flexneri 2a str. 2457T	GCF_000007405.1	AE014073.1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000007405.1
Shigella flexneri 5 str. 8401	GCF_000013585.1	CP000266.1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000013585.1
Escherichia coli IAI39	GCF_000026345.1	ASM2634v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000026345.1
Shigella dysenteriae Sd197	GCF_000012005.1	CP000034.1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000012005.1
Escherichia coli O157:H7 str. EDL933	GCA_000006665.1	ASM666v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000006665.1
Escherichia coli O157:H7 str. Sakai	GCF_000008865.2	ASM886v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000008865.2
Escherichia coli UMN026	GCF_000026325.1	ASM2632v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000026325.1
Escherichia coli UTI89	GCF_000013265.1	ASM1326v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000013265.1
Escherichia coli APEC O1	GCF_000014845.1	ASM1484v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000014845.1
Escherichia coli S88	GCF_000026285.1	ASM2628v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000026285.1
Escherichia coli CFT073	GCF_000007445.1	ASM744v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000007445.1
Escherichia coli ED1a	GCF_000026305.1	ASM2630v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000026305.1
Escherichia coli 536	GCF_000013305.1	CP000247.1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000013305.1
Escherichia coli ATCC 8739	GCF_003591595.1	ASM359159v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_003591595.1
Escherichia coli O139:H28 str. E24377A	GCF_000017745.1	ASM1774v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000017745.1
Escherichia coli SE11	GCF_000010385.1	ASM1038v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000010385.1
Escherichia coli LF82	GCF_000284495.1	ASM28449v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000284495.1
Escherichia coli O127:H6 str. E2348/69	GCA_000026545.1	FM180568.1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000026545.1
Escherichia coli O157:H7 str. EC4115	GCF_000021125.1	ASM2112v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000021125.1
Escherichia coli HS	GCF_000017765.1	ASM1776v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000017765.1
Escherichia coli 042	GCF_000027125.1	ASM2712v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000027125.1
Escherichia coli SMS-3-5	GCF_000019645.1	ASM1964v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000019645.1

Table S1: **Description of the 29 Bacterial genomes** with link to the public available genomes in the NCBI database.

Strain	Assembly	Genome Assembly used	Link
Aspergillus clavatus NRRL 1	GCA_000002715.1	ASM271v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002715.2
Aspergillus flavus NRRL3357	GCA_000006275.2	JCVI-afl1-v2.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000006275.2
Aspergillus fumigatus	GCA_012656185.1	ASM1265618v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_012656185.1
Aspergillus nidulans FGSC A4	GCA_000011425.1	ASM1142v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000011425.1

Description of the 77 Fungal genomes (continued on next page)

Strain	Assembly	Genome Assembly used	Link
Aspergillus niger CBS 513.88	GCA_000002855.2	ASM285v2	https://www.ncbi.nlm.nih.gov/assembly/GCA_000002855.2
Aspergillus oryzae 3.042	GCA_000269785.2	AspOry3042	https://www.ncbi.nlm.nih.gov/assembly/GCA_000269785.2
Aspergillus terreus NIH2624	GCA_000149615.1	ASM14961v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149615.1
Batrachochytrium dendrobatidis JAM81	GCA_000203795.1	v1.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000203795.1
Batrachochytrium dendrobatidis JEL423	GCA_000149865.1	BD_JEL423	https://www.ncbi.nlm.nih.gov/assembly/GCA_000149865.1
Botrytis cinerea B05.10	GCA_000143535.4	ASM14353v4	https://www.ncbi.nlm.nih.gov/assembly/GCF_000143535.2
Caenorhabditis elegans	GCA_000002985.3	WBcel235	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002985.6
Candida albicans SC5314	GCA_000182965.3	ASM18296v3	https://www.ncbi.nlm.nih.gov/assembly/GCF_000182965.3
Candida albicans WO-1	GCA_000149445.2	ASM14944v2	https://www.ncbi.nlm.nih.gov/assembly/GCA_000149445.2
Candida glabrata CBS138	GCA_000002545.2	ASM254v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002545.3
Meyerozyma guilliermondii ATCC 6260	GCA_000149425.1	ASM14942v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149425.1
Clavispora lusitaniae P1	GCA_009498055.1	ASM949805v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_009498055.1
Candida parapsilosis CDC 317	GCA_000182765.2	ASM18276v2	https://www.ncbi.nlm.nih.gov/assembly/GCA_000182765.2
Candida tropicalis MYA-3404	GCA_000006335.3	ASM633v3	https://www.ncbi.nlm.nih.gov/assembly/GCF_000006335.3
Bipolaris maydis C5	GCA_000338975.1	CocheC5_3	https://www.ncbi.nlm.nih.gov/assembly/GCA_000338975.1
Chaetomium globosum CBS 148.51	GCA_000143365.1	ASM14336v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000143365.1
Coccidioides immitis H538.4	GCA_000149815.1	ASM14981v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000149815.1
Coccidioides immitis RMSCC 2394	GCA_000149895.1	ASM14989v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000149895.1
Coccidioides immitis RMSCC 3703	GCA_000150085.1	ASM15008v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000150085.1
Coccidioides immitis RS	GCA_000149335.2	ASM14933v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149335.2
Coccidioides posadasii RMSCC 3488	GCA_000150055.1	ASM15005v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000150055.1
Coccidioides posadasii Silveira	GCA_000170175.2	CPS2	https://www.ncbi.nlm.nih.gov/assembly/GCA_000170175.2
Coprinopsis cinerea okayama7#130	GCA_000182895.1	CC3	https://www.ncbi.nlm.nih.gov/assembly/GCF_000182895.1
Cryptococcusgattii Ru294	GCA_000836355.1	Cryp_gatt_Ru294_V1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000836355.1
Cryptococcus gattii WM276	GCA_000185945.1	ASM18594v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000185945.1
Cryptococcus neoformans var. grubii H99	GCA_000149245.3	CNA3	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149245.1
Cryptococcus neoformans var. neoformans JEC21	GCA_000091045.1	ASM9104v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000091045.1
Debaryomyces hansenii CBS767	GCA_000006445.2	ASM644v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000006445.2
Drosophila melanogaster	GCA_000001215.4	Release 6 plus ISO1 MT	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001215.4
Encephalitozoon cuniculi GB-M1	GCA_000091225.1	ASM9122v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000091225.1
Eremothecium gossypii ATCC 10895	GCA_000091025.4	ASM9102v4	https://www.ncbi.nlm.nih.gov/assembly/GCF_000091025.4

Description of the 77 Fungal genomes (continued on next page)

Strain	Assembly	Genome Assembly used	Link
Fusarium graminearum PH-1	GCA_000240135.3	ASM24013v3	https://www.ncbi.nlm.nih.gov/assembly/GCF_000240135.3
Fusarium oxysporum f. sp. lycopersici 4287	GCA_000149955.2	ASM14995v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149955.1
Histoplasma capsulatum NAm1	GCA_000149585.1	ASM14958v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149585.1
Fusarium verticillioides 7600	GCA_000149555.1	ASM14955v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149555.1
Kluyveromyces lactis NRRL Y-1140	GCA_000002515.1	ASM251v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002515.2
Laccaria bicolor S238N-H82	GCA_000143565.1	V1.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000143565.1
Lodderomyces elongisporus NRRL YB-4239	GCA_000149685.1	ASM14968v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149685.1
Pyricularia grisea strain NI907	GCA_004355905.1	ASM435590v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_004355905.1
Malassezia globosa CBS 7966	GCA_000181695.1	ASM18169v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000181695.1
Monosiga brevicollis MX1	GCA_000002865.1	V1.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002865.3
Mycosphaerella fijiensis CIRAD86	GCA_000340215.1	Mycfi2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000340215.1
Mycosphaerella graminicola IPO323	GCA_000219625.1	MYCGR v2.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000219625.1
Nectria haematococca mpVI 77-13-4	GCA_000151355.1	v2.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000151355.1
Neosartorya fischeri NRRL 181	GCA_000149645.3	ASM14964v3	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149645.2
Neurospora crassa OR74A	GCA_000182925.2	NC12	https://www.ncbi.nlm.nih.gov/assembly/GCF_000182925.2
Paracoccidioides brasiliensis Pb03	GCA_000150475.2	Paracocci_br_ Pb03_V2	https://www.ncbi.nlm.nih.gov/assembly/GCA_000150475.2
Paracoccidioides brasiliensis Pb18	GCA_000150735.2	Paracocci_br_ Pb18_V2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000150735.1
Paracoccidioides brasiliensis Pb300	GCA_001713645.1	ASM171364v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_001713645.1
Phycomyces blakesleeanus NRRL 1555(-)	GCA_001638985.2	Phybl2	https://www.ncbi.nlm.nih.gov/assembly/GCF_001638985.1
Pichia stipitis CBS 6054	GCA_000209165.1	ASM20916v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000209165.1
Podospora anserina S mat+	GCA_000226545.1	ASM22654v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000226545.1
Postia placenta MAD-698-R-SB12	GCA_002117355.1	PosplRSB12_1	https://www.ncbi.nlm.nih.gov/assembly/GCF_002117355.1
Puccinia graminis f. sp. tritici CRL 75-36-700-3	GCA_000149925.1	ASM14992v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149925.1
Pyrenophora tritici-repentis Pt-1C-BFP	GCA_000149985.1	ASM14998v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149985.1
Rhizopus delemar RA 99-880	GCA_000149305.1	RO3	https://www.ncbi.nlm.nih.gov/assembly/GCA_000149305.1
Naumovozyma castellii CBS 4309	GCA_000237345.1	ASM23734v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000237345.1
Saccharomyces cerevisiae GLBRCY22-3	GCA_001634645.1	GLBRCY22-3	https://www.ncbi.nlm.nih.gov/assembly/GCA_001634645.1
Saccharomyces cerevisiae S288C	GCA_000146045.2	R64	https://www.ncbi.nlm.nih.gov/assembly/GCF_000146045.2
Saccharomyces cerevisiae YJM789	GCA_000181435.1	ASM18143v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000181435.1
Saccharomyces kudriavzevii IFO 1802	GCA_000167075.2	Saccharomyces_kudriavzevii_strain_ IFO1802_v1.0	https://www.ncbi.nlm.nih.gov/assembly/GCA_000167075.2

Description of the 77 Fungal genomes (continued on next page)

Strain	Assembly	Genome Assembly used	Link
Schizosaccharomyces japonicus yFS275	GCA_000149845.2	SJ5	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149845.2
Schizosaccharomyces octosporus yFS286	GCA_000150505.2	SO6	https://www.ncbi.nlm.nih.gov/assembly/GCF_000150505.1
Schizosaccharomyces pombe 972h-	GCA_000002945.2	ASM294v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002945.1
Sclerotinia sclerotiorum 1980 UF-70	GCA_000146945.2	ASM14694v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000146945.2
Parastagonospora nodorum SN15	GCA_000146915.2	ASM14691v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000146915.1
Trichoderma atroviride IMI 206040	GCA_000171015.2	TRIAT v2.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000171015.1
Trichoderma reesei QM6a	GCA_000167675.2	v2.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000167675.1
Trichoderma virens Gv29-8	GCA_000170995.2	TRIVI v2.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000170995.1
Uncinocarpus reesii 1704	GCA_000003515.2	ASM351v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000003515.1
Ustilago maydis 521	GCA_000328475.2	Umaydis521_2.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000328475.2
Verticillium dahliae VdLs.17	GCA_000150675.2	ASM15067v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000150675.1
Yarrowia lipolytica CLIB89(W29)	GCA_001761485.1	ASM176148v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_001761485.1

Table S2: **Description of the 77 genomes of the fungal dataset** with link to the public available genomes in the NCBI database.

Strain	Assembly	Genome Assembly used	Link
Auxenochlorella protothecoides 0710	GCA_000733215.1	ASM73321v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000733215.1
Bathycoccus prasinos	GCA_002220235.1	ASM222023v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_002220235.1
Caenorhabditis elegans	GCA_000002985.3	WBcel235	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002985.6
Chara braunii	GCA_003427395.1	Cbr_1.0	https://www.ncbi.nlm.nih.gov/assembly/GCA_003427395.1
Chlamydomonas reinhardtii CC-503 cw92 mt+	GCA_000002595.2	v3.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002595.1
Chlorella variabilis	GCA_000147415.1	v 1.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000147415.1
Chondrus crispus Stackhouse	GCA_000350225.2	ASM35022v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000350225.1

Description of the 40 Algal genomes (continued on next page)

Strain	**Assembly**	**Genome Assembly used**	**Link**
Chrysochsromulina sp. CCMP291	GCA_001275005.1	Ctobinv2	https://www.ncbi.nlm.nih.gov/assembly/GCA_001275005.1
Cladosiphon okamuranus		v 0.4	https://marinegenomics.oist.jp/algae/viewer/download?project_id=53
Coccomyxa subellipsoidea C-169	GCA_000258705.1	Coccomyxa subellipsoidae v2.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000258705.1
Cyanidioschyzon merolae 10D	GCA_000091205.1	ASM9120v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000091205.1
Cyanophora paradoxa			http://cyanophora.rutgers.edu/cyanophora_v2018 (*)
Ectocarpus siliculosus			https://bioinformatics.psb.ugent.be/gdb/ectocarpusV2
Ectocarpus subulatus	LR740778-LR742460		http://application.sb-roscoff.fr/blast/subulatus/download.html
Emiliania huxleyi CCMP1516	GCA_000372725.1	Emiliana huxleyi CCMP1516 main genome assembly v1.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000372725.1
Fistulifera solaris JPCC DA0580	GCA_002217885.1	Fsol_1.0	https://www.ncbi.nlm.nih.gov/assembly/GCA_002217885.1
Fragilariopsis cylindrus CCMP1102	GCA_001750085.1	Fracy1	https://www.ncbi.nlm.nih.gov/assembly/GCA_001750085.1
Galdieria sulphuraria	GCA_000341285.1	ASM34128v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000341285.1
Gonium pectorale	GCA_001584585.1	ASM158458v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_001584585.1
Gracilarioppsis chorda	GCA_003194525.1	GraCho1.0	https://www.ncbi.nlm.nih.gov/assembly/GCA_003194525.1
Guillardia theta CCMP2712	GCA_000315625.1	Guith1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000315625.1
Klebsormidium nitens	GCA_000708835.1	ASM70883v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_000708835.1
Micractinium conductrix SAG 241.80	GCA_002245815.2	ASM224581v2	https://www.ncbi.nlm.nih.gov/assembly/GCA_002245815.2
Micromonas pusilla CCMP1545	GCA_000151265.1	Micromonas pusilla CCMP1545 v2.0	https://www.ncbi.nlm.nih.gov/assembly/GCF_000151265.2

Description of the 40 Algal genomes (continued on next page)

Strain	**Assembly**	**Genome Assembly used**	**Link**
Nannochloropsis gaditana B-31	GCA_000569095.1	NagaB31_1.0	https://www.ncbi.nlm.nih.gov/assembly/GCA_000569095.1
Monoraphidium neglectum SAG 48.87	GCA_000611645.1	mono_v1	https://www.ncbi.nlm.nih.gov/assembly/GCF_000611645.1
Mucor circinelloides f. lusitanicus CBS 277.49		v2	https://mycocosm.jgi.doe.gov/Mucci2/Mucci2.home.html
Nemacystus decipiens		v1.0	https://marinegenomics.oist.jp/ito_mozuku_v1/viewer/download?project_id=68
Neurospora crassa OR74A	GCA_000182925.2	NC12	https://www.ncbi.nlm.nih.gov/assembly/GCF_000182925.2
Ostreococcus tauri	GCA_000214015.2	version 140606	https://www.ncbi.nlm.nih.gov/assembly/GCF_000214015.3
Phaeodactylum tricornutum CCAP 1055/1	GCA_000150955.2	ASM15095v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000150955.2
Porphyra umbilicalis	GCA_002049455.2	P_umbilicalis_v1	https://www.ncbi.nlm.nih.gov/assembly/GCA_002049455.2
Raphidocelis subcapitata NIES-35	GCA_003203535.1	Rsub_1.0	https://www.ncbi.nlm.nih.gov/assembly/GCA_003203535.1
Saccharina japonica	GCA_008828725.1	ASM882872v1	http://124.16.129.28:8080/saccharina (*)
Saccharomyces cerevisiae S288C	GCA_000146045.2	R64	https://www.ncbi.nlm.nih.gov/assembly/GCF_000146045.2
Tetrabaena socialis NIES-571	GCA_002891735.1	TetSoc1	https://www.ncbi.nlm.nih.gov/assembly/GCA_002891735.1
Thalassiosira pseudonana CCMP1335	GCA_000149405.2	ASM14940v2	https://www.ncbi.nlm.nih.gov/assembly/GCF_000149405.2
Tisochrysis lutea	Tiso V2		https://www.seanoe.org/data/00411/52231
Ulva mutabilis	GCA_900538255.1	Ulvmu_WT_fa	https://bioinformatics.psb.ugent.be/gdb/ulva
Volvox carteri		v2.1	https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v2/organism/view/Volvox+carteri

Table S3: **Description of the 40 genomes of the algal dataset** with link to the public available genomes. (*) Data retrieved in December 2019. Please note that the site no longer accessible as of Jan 26th 2023. You can access an archived version of the genome used in this paper via our Zenodo deposit (https://zenodo.org/record/7752449#.ZBhOpiOZN-E).

Dataset	Ratio genes not degraded	Ratio functional degradation	Ratio structural degradation	Number conserved genes	Number of genes impacted by the functional degradation	Number of genes impacted by the structural degradation
Run_00	1	0	0	2267	0	0
Run_01	0.9	0.1	0	2066	201	0
Run_02	0.8	0.2	0	1798	469	0
Run_03	0.7	0.3	0	1559	708	0
Run_04	0.6	0.4	0	1371	896	0
Run_05	0.5	0.5	0	1112	1155	0
Run_06	0.4	0.6	0	929	1338	0
Run_07	0.3	0.7	0	675	1592	0
Run_08	0.2	0.8	0	460	1807	0
Run_09	0.1	0.9	0	226	2041	0
Run_10	0	1	0	0	2267	0
Run_11	0.9	0.05	0.05	2070	112	85
Run_12	0.8	0.1	0.1	1801	231	235
Run_13	0.7	0.15	0.15	1585	344	338
Run_14	0.6	0.2	0.2	1350	476	441
Run_15	0.5	0.25	0.25	1124	578	565
Run_16	0.4	0.3	0.3	921	672	674
Run_17	0.34	0.33	0.33	745	781	741
Run_18	0.3	0.35	0.35	713	738	816
Run_19	0.2	0.4	0.4	434	933	900
Run_20	0.1	0.45	0.45	232	1058	977
Run_21	0	0.5	0.5	0	1127	1140
Run_22	0.9	0	0.1	2049	0	218
Run_23	0.8	0	0.2	1803	0	464
Run_24	0.7	0	0.3	1582	0	685

Description of the 32 in silico bacterial datasets (continued on next page)

Dataset	**Ratio genes not degraded**	**Ratio functional degradation**	**Ratio structural degradation**	**Number conserved genes**	**Number of genes impacted by the functional degradation**	**Number of genes impacted by the structural degradation**
Run_25	0.6	0	0.4	1348	0	919
Run_26	0.5	0	0.5	1142	0	1125
Run_27	0.4	0	0.6	892	0	1375
Run_28	0.3	0	0.7	671	0	1596
Run_29	0.2	0	0.8	436	0	1831
Run_30	0.1	0	0.9	236	0	2031
Run_31	0	0	1	0	0	2267

Table S4: **Description of the 32 in silico bacterial datasets**. The tables shows the ratio and number of genes (associated with reactions) impacted by the functional and/or structural degradation the E. coli K12-MG1655 genome. The total number of genes (2267) corresponds to the number of genes identified as being involved in the metabolism in the non-degraded dataset 0, so it is inferior to the total number of genes in the genome. Each row corresponds to a dataset with a specific degradation: either only functional degradation (datasets 1 to 10) or only structural degradation (datasets 11 to 21) or both (datasets 22 to 31).

2.2 AuCoMe homogenizes the content of metabolic network collections

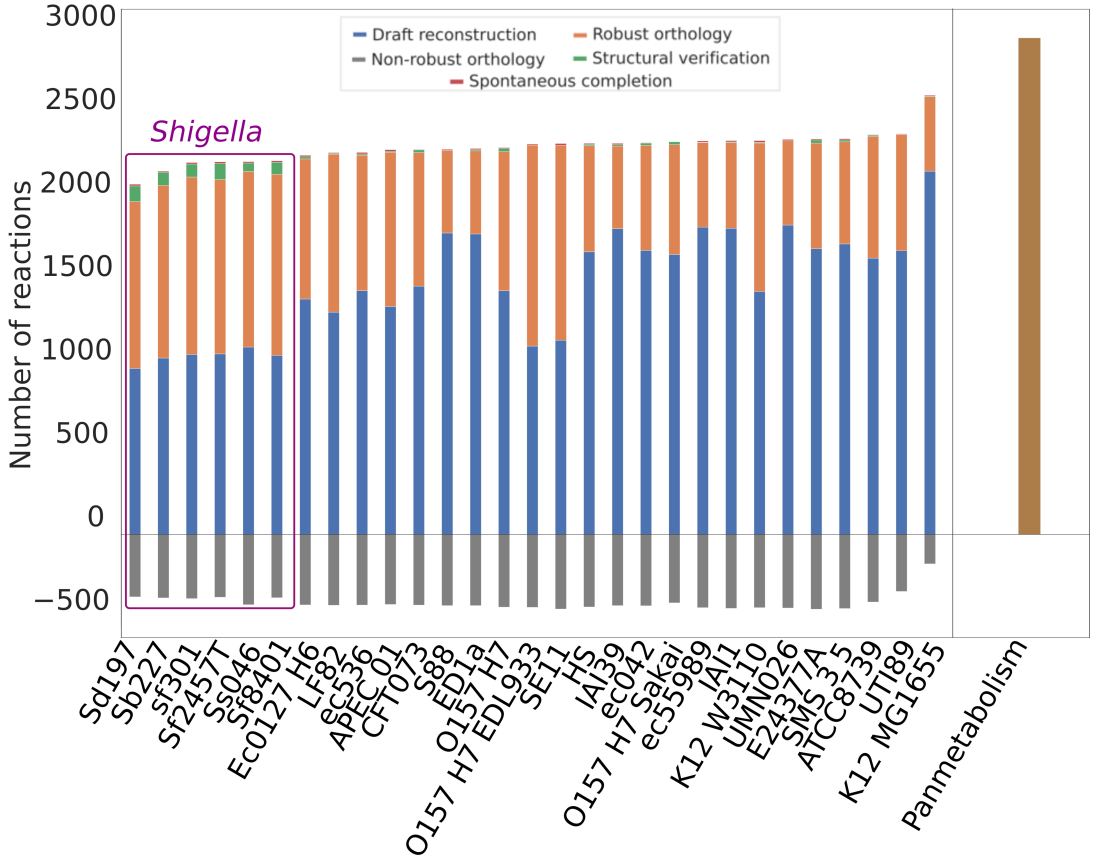

Figure S1: **Application of the AuCoMe pipeline to the bacterial dataset of genomes.** The number of reactions are identified for each species at each step of the AuCoMe pipeline: reactions recovered by the draft reconstruction step (blue), unreliable reactions predicted by orthology propagation and removed by the filter (gray), robust reactions predicted by orthology propagation that passed the filter (orange), additional reactions predicted by the structural verification step (green), and spontaneous completion (red). The final metabolic networks encompass all these reactions except the non-reliable ones. The panmetabolism of this dataset (all the reactions occurring in any of the organisms after the final step of AuCoMe) is presented in brown.

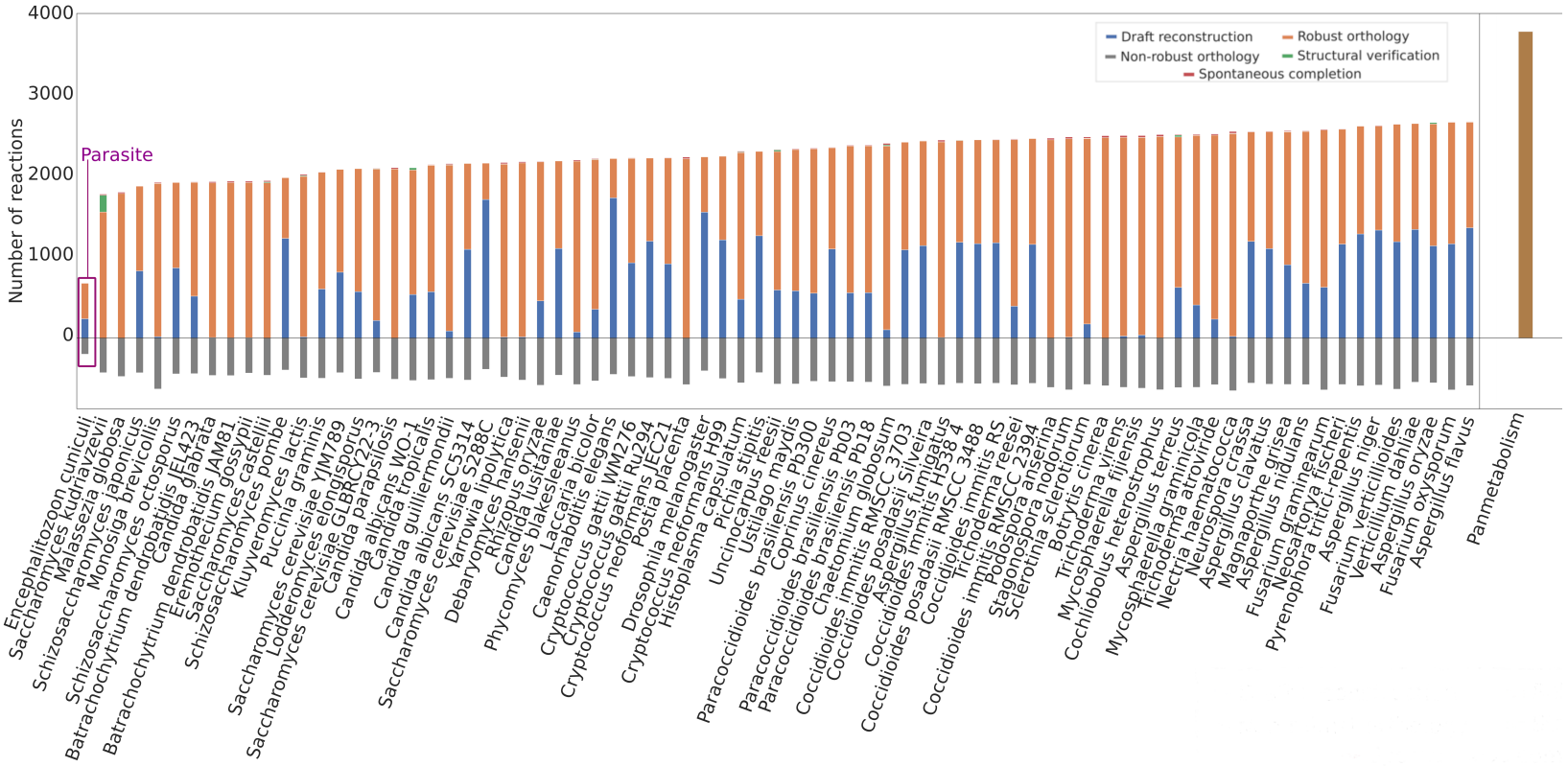

Figure S2: **Application of the AuCoMe pipeline to the fungal dataset of genomes.** The number of reactions are identified for each species at each step of the AuCoMe pipeline: reactions recovered by the draft reconstruction step (blue), unreliable reactions predicted by orthology propagation and removed by the filter (gray), robust reactions predicted by orthology propagation that passed the filter (orange), additional reactions predicted by the structural verification step (green), and spontaneous completion (red). The final metabolic networks encompass all these reactions except the non-reliable ones. The panmetabolism of this dataset is presented in brown.

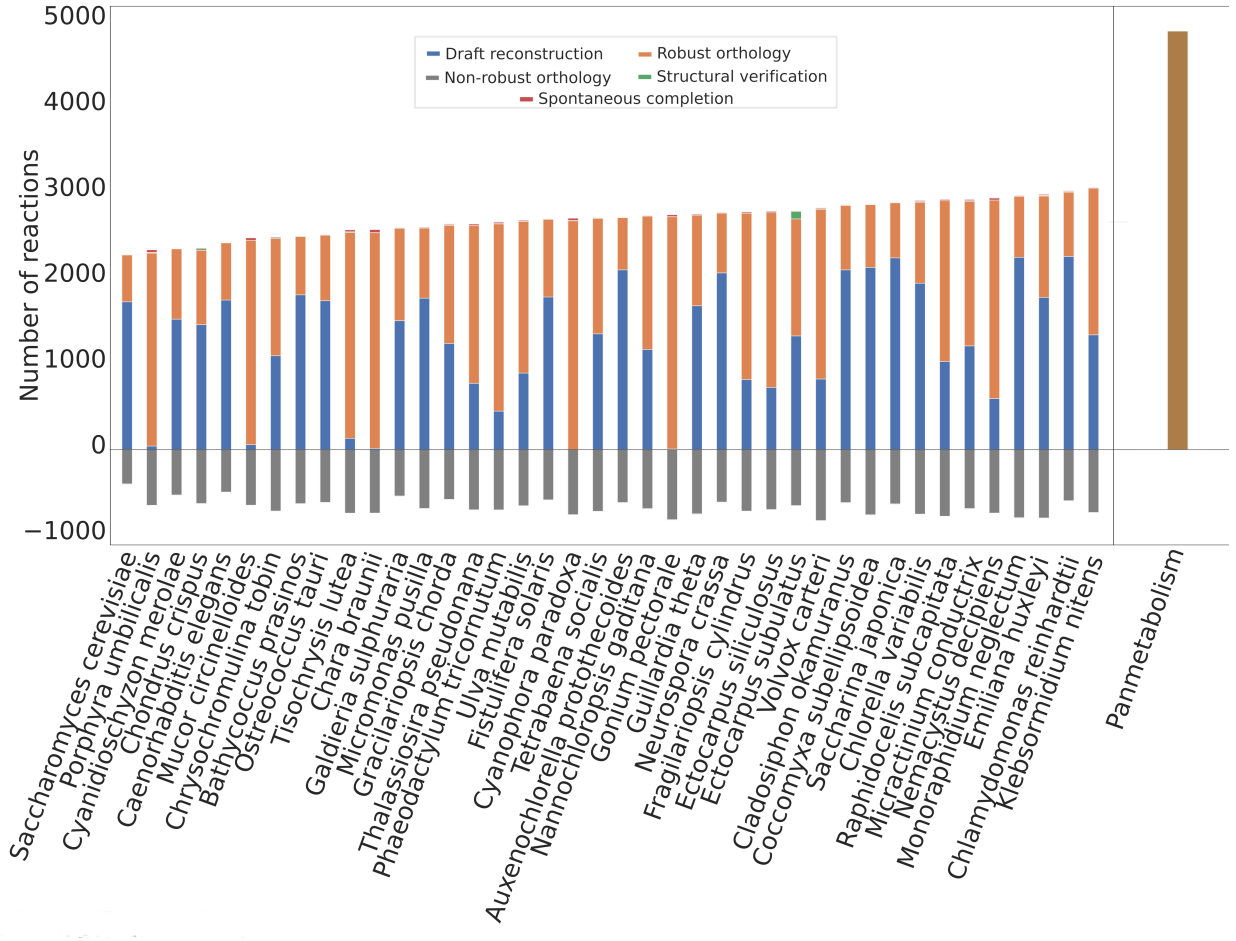

Figure S3: **Application of the AuCoMe pipeline to the algal dataset of genomes.** The number of reactions are identified for each species at each step of the AuCoMe pipeline: reactions recovered by the draft reconstruction step (blue), unreliable reactions predicted by orthology propagation and removed by the filter (gray), robust reactions predicted by orthology propagation that passed the filter (orange), additional reactions predicted by the structural verification step (green), and spontaneous completion (red). The final metabolic networks encompass all these reactions except the non-reliable ones. The panmetabolism of this dataset is presented in brown.

2.3 Comparison with gapseq, ModelSEED and CarveMe on a bacterial dataset

To compare the metabolic networks created by AuCoMe with metabolic networks from other methods we performed a first experiment on the bacterial dataset.

We made the following key observations regarding the pipelines (Figure S4 A-F): (i) reactions without genes: AuCoMe has by design no reaction without gene association; (ii) reactions without gene: CarveMe has the most reactions associated with no gene, estimated using gap filling; (iii) all reactions: gapseq has a higher total number of reactions than AuCoMe, some come from gap filling; (iv) unique ECs: AuCoMe has the highest number of unique ECs as it propagates annotations from different genomes, potentially merging annotations from multiple sources (multiple annotation methods, potential manual annotations).

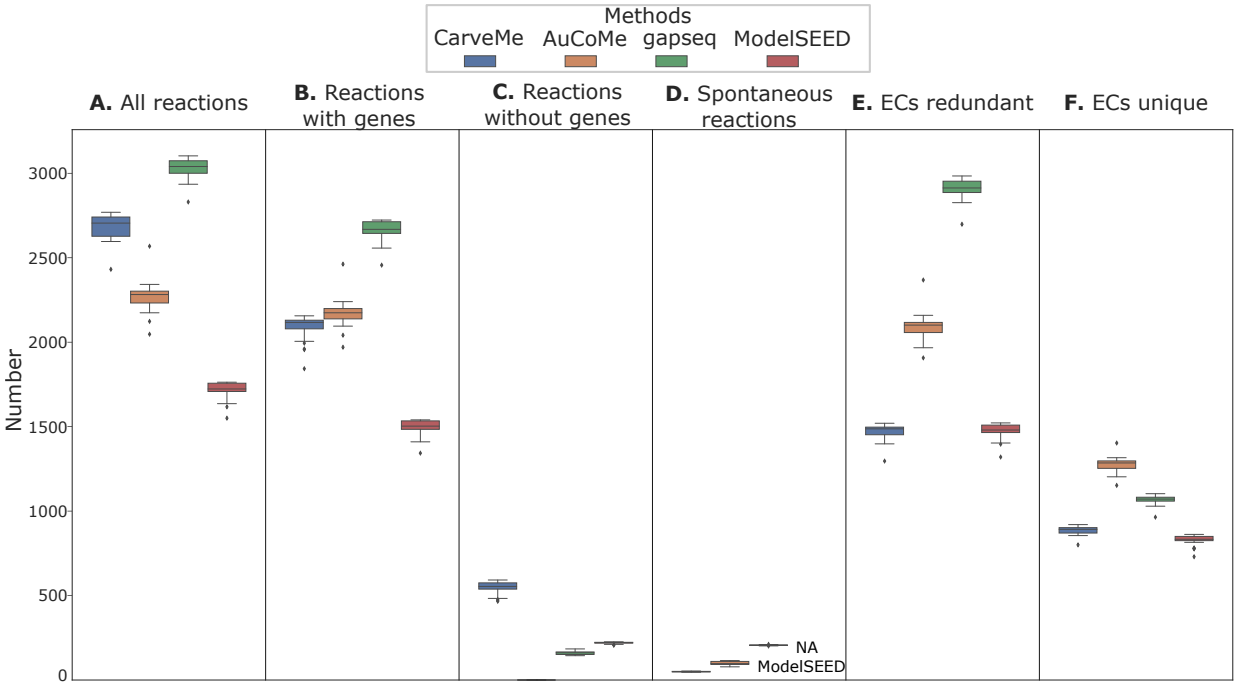

Figure S4: **Methods comparison on the 29 bacterial species dataset.** Distribution of the number of reactions, reactions with genes, reactions without genes, spontaneous reactions, EC numbers redundant and EC numbers unique, according to the methods used.

To evaluate the quality of the ECs predicted by the different tools, we created a reference set of EC numbers associated with *E. coli* K–12 MG1655 , that we called *reference EC catalog*. This reference EC catalog contained ECs from 4 databases (KEGG, EcoCyc, ModelSEED and BiGG). For KEGG, we retrieved the ECs with a REST query for the organism code 'eco'. For EcoCyc, we converted the PGDB of EcoCyc

version 23.5 in padmet format to retrieve the EC number associated with the reactions. For ModelSEED, we reconstructed a metabolic network for the corresponding strain on the ModelSEED website and then we mapped the reactions of ModelSEED to ECs (found in the file 'Unique_ModelSEED_Reaction_ECs.txt' in the GitHub repository of ModelSEED). For BiGG, we extracted ECs from the json files of the models iAF1260, iAF1260b, iJO1366, iJR904 and iML1515. Then we merged all these ECs in one set to create the reference EC catalog containing 1,868 ECs.

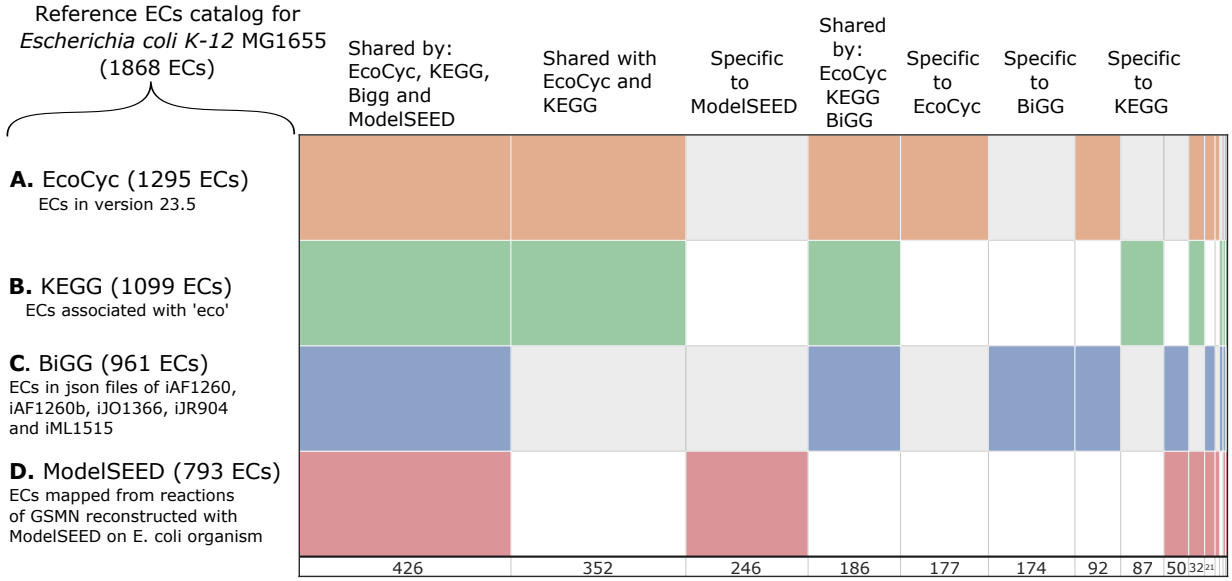

Figure S5: **Distribution of EC in reference EC catalog for *E. coli* K–12 MG1655 .** It has been created from the ECs of four databases containing data on *E. coli* K–12 MG1655 : EcoCyc, BiGG, KEGG and ModelSEED. Each row corresponds to one database and the column corresponds to set of ECs. Each colored block indicated that the corresponding database contains the set of corresponding ECs. The size of the block is proportional to the number of ECs in the block.

The distribution of the ECs contained in the reference EC catalog is represented in Figure S5 A-D. This figure was generated using supervenn [3]. Each block corresponds to a set of EC numbers shared by one or more databases. EcoCyc and KEGG contained the highest number of EC numbers and shared most of them. However, each database also had a set of ECs that were unique to it. Using this reference EC catalog, we compared the number of true positives, false positives and false negatives between this reference EC catalog and the networks reconstructed by AuCoMe, CarveMe, gapseq and ModelSEED in Figure S6. We considered a true positive to be an EC present in both the reference EC catalog and the metabolic network, a false positive to be an EC present in the reference EC catalog but not in the metabolic network and a false negative to be an EC not present in both the reference EC catalog and the metabolic network. We did

not consider true negatives as informative in this case, as it should correspond to EC number not present in both the metabolic network and the reference EC catalog.

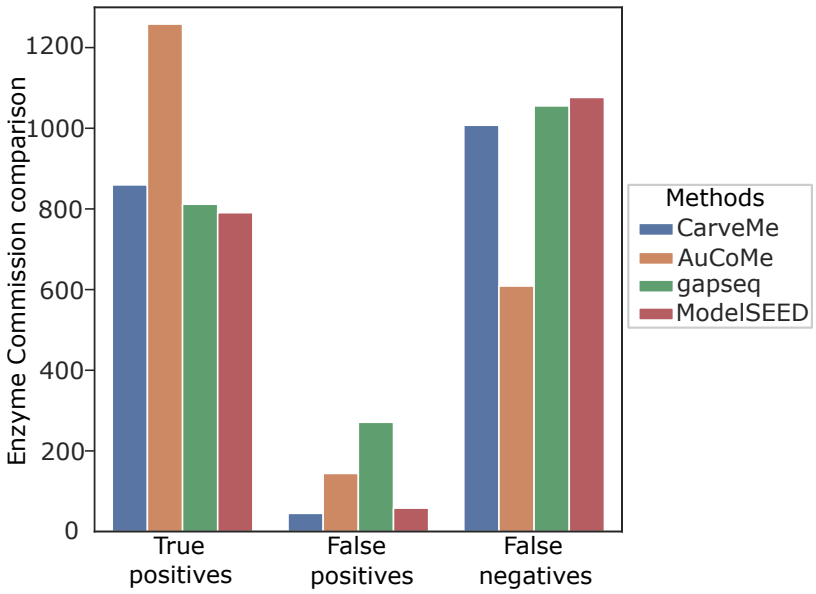

Figure S6: **Comparison between the metabolic networks inferred for *E. coli* K–12 MG1655 .** The comparison was performed using 4 tools (AuCoMe, CarveMe, gapseq and ModelSEED) and the reference set of EC numbers from 4 databases (EcoCyc, KEGG, BiGG and ModelSEED).

In the Figure S6, we can see that AuCoMe predicted the highest number of true positives. It also predicted the lowest number of false negatives. CarveMe and ModelSEED predicted the lowest number of false positives.

2.4 Comparison with gapseq and ModelSEED and on a fungal dataset

The second comparison was made on 5 organisms of the fungal dataset. The results can be seen in Table S5.

AuCoMe predicts the highest number of pathways, reactions and EC numbers (except for *S. cerevisiae*). For pathways with a completion rate greater than 70%, AuCoMe and gapseq predicted similar numbers of reactions for *Saccharomyces cerevisiae S288C*, respectively 405 and 409. But for other organisms, AuCoMe predicted a constant number of metabolic pathways (approximately 400) whereas gapseq was more heterogeneous (between 24 and 409). For metabolic pathways with a completion ratio inferior to 70%, AuCoMe predicted a high number of pathways (between 656 and 815) whereas gapseq predicted around ten pathways. This could be explained by the fact that AuCoMe does not filter the metabolic pathways and proposes all possible metabolic pathways containing at least one reaction whereas gapseq filters metabolic pathways according to

A. Pathways and reactions

	Pathways					**Reactions**		
	AuCoMe		**gapseq find**		**ModelSEED**	**AuCoMe**	**gapseq**	**ModelSEED**
Organism	>= 70%	<70%	>= 70%	<70%				
Schizosaccharomyces pombe	370	658	306	14	405	1,995	1,232	1,062
Saccharomyces cerevisiae S288C	405	656	409	11	394	2,176	1,726	982
Rhizopus oryzae	407	754	90	10	405	2,196	235	1,062
Neurospora crassa	456	815	273	17	458	2,567	1,007	1,059
Laccaria bicolor	398	757	24	1	433	2,232	35	1,234

B. EC and runtimes

	Enzyme Commission (unique EC)			**Runtimes (in seconds)**		
Organism	**AuCoMe final**	**gapseq find**	**ModelSEED**	**AuCoMe**	**gapseq find**	**ModelSEED**
Schizosaccharomyces pombe	1,897 (1,051)	1,531 (549)	70 (67)	90,082s (for the 77 fungal genomes)	45,598s	4,200s
Saccharomyces cerevisiae S288C	2,073 (1,129)	2,075 (741)	69 (64)		45,565s	6,060s
Rhizopus oryzae	2,114 (1,186)	249 (159)	70 (67)		49,806s	4,380s
Neurospora crassa	2,446 (1,360)	1,275 (482)	91 (90)		31,252s	11,160s
Laccaria bicolor	2,151 (1,195)	32 (26)	76 (73)		76,126s	11,460s

Table S5: **Comparison between the predictions made by AuCoMe, gapseq and ModelSEED on 5 fungi.** For AuCoMe, the 5 metabolic networks were obtained from the fungal dataset containing 77 fungal genomes. For gapseq, we used the find module to identify the metabolic pathways associated with the fungal genome sequences. ModelSEED networks were reconstructed using the module 'Build Fungal Model' in KBase. **A.** Number of metabolic pathways (with a separation between pathways complete at more than 70% or pathway's ratio lower than 70%) and reactions for AuCoMe, gapseq find module and ModelSEED for 5 fungi from the fungal dataset. **B.** Count of EC numbers and runtimes for gapseq, AuCoMe and ModelSEED. The runtime for the AuCoMe final step corresponds to the runtime for the 77 fungi of the fungal dataset.

their completion ratios.

Then we compared the MetaCyc metabolic pathways predicted by AuCoMe and gapseq. We also compared the completion of the metabolic pathways completion formats for gapseq and AuCoMe. To visualise it, Fig. S7 shows the distribution of metabolic pathways according to their completion (from 0 percent to 100 percent). In each bar of the histogram we coloured the number of pathways, if they were only found in gapseq (blue), only found in AuCoMe (orange), or in both at the same completion rate (green). We also indicated four other colours in the case where the metabolic pathways were found by the two methods but not in the same completion rate. The highest number of pathways corresponded to the completion rate between 90 and 100%. In this range, half of the pathways were found by both tools except for *R. oryzae*. AuCoMe was the most constant as it predicted similar number of pathways with the same distribution of completion rates compared to gapseq. For example, the case of *R. oryzae* where gapseq predicted around 100 metabolic pathways. In this case, AuCoMe seemed more relevant to predict metabolic pathways for non-model organisms.

To assess the quality of these predictions, we performed a comparison on the metabolic network of *Saccharomyces cerevisiae S288C*. We used the metabolic pathways contained in the manually curated database YeastCyc as a reference. There were 342 metabolic pathways in YeastCyc version 26.5. The results of this comparison were shown in Figure S8. In total, 236 metabolic pathways of YeastCyc were predicted by AuCoMe, most of them were complete between 91% to 100%. Most of the metabolic pathways predicted by AuCoMe with low completeness ratio were not tagged as present in YeastCyc.

We compared the performance of AuCoMe for pathways with a completeness ratio equal or higher than 70% with gapseq. The results are shown in Figure S9. For the true positives the 205 + 20 in Figure S9 corresponds to the sum of the intervals 70-79, 80-89, 90-99 and 100 for the 'Common to YeastCyc and AuCoMe' in Figure S8. Similarly, for the false positives 89+90 in Figure S9 corresponds to the intervals 70-79, 80-89, 90-99 and 100 for the 'Only in AuCoMe' in Figure S8. But there is no correspondence between the false negatives in these figures. As the Figure S9 was made with a subset of pathways for AuCoMe (the ones with a completeness rate superior to 70%), some pathways were detected as false negatives whereas they were found with lower completeness rate in S8. Compared to YeastCyc, AuCoMe and gapseq displayed similar results. There were similar numbers of false positives (93 for gapseq against 89 for AuCoMe) and true positives (236 for gapseq and 225 for AuCoMe).

In conclusion, AuCoMe predicts ECs that are consistent with the known ECs as shown in the comparison

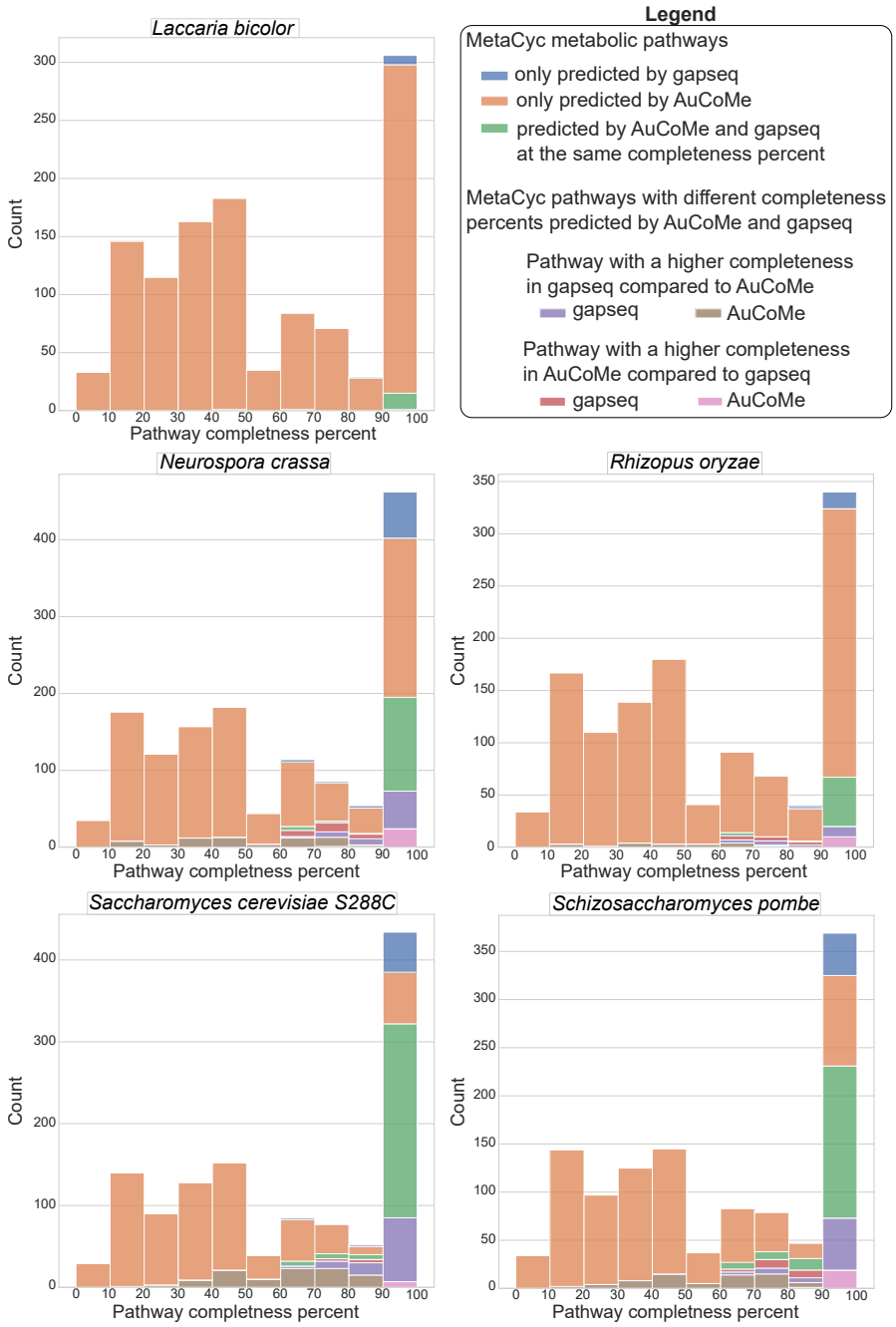

Figure S7: **Metabolic pathways comparison between AuCoMe and gapseq for five fungal species.** Abscissa coordinates indicate the completion rate of the MetaCyc pathways associated with predicted reactions. Color bars indicate which method predicted each MetaCyc pathways, blue: pathways predicted by gapseq only; orange: pathways predicted by AuCoMe only; green: pathways predicted with both AuCoMe and gapseq within the same range of completion. Other colors indicate MetaCyc pathways predicted by both AuCoMe and gapseq, but within different ranges of completion. Pathway predicted with a higher completion rate in gapseq compared to AuCoMe (blue: gapseq pathways; brown: AuCoMe pathways). Pathway predicted with a higher completion rate in AuCoMe compared to gapseq (red: gapseq pathways; purple: AuCoMe pathways).

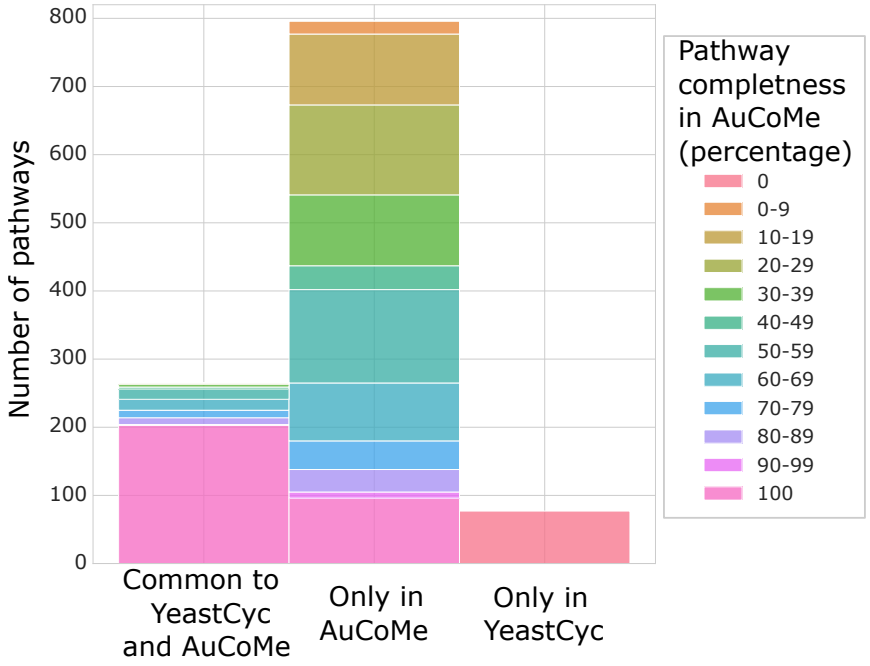

Figure S8: **Number of pathways common or specific to YeastCyc and AuCoMe with their completeness ~~ratio~~percentage predicted by AuCoMe.**

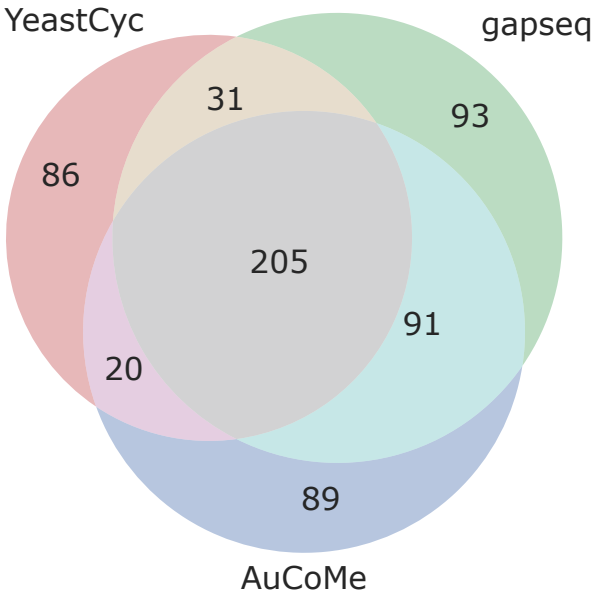

Figure S9: **Comparison of the metabolic pathways in YeastCyc, predicted by gapseq or AuCoMe.** Only pathways with the completeness ratio superior or equal to 70% were counted.

with the reference EC catalog of *E. coli* K–12 MG1655 . Also, AuCoMe performs similarly to gapseq in predicting metabolic pathways with completeness rate superior to 70% for the model organism *Saccharomyces cerevisiae S288C*. It predicts numerous metabolic pathways with a lower completeness ratio, that are candidates that could be looked into as a small part of them were true positives in the comparison with YeastCyc. The advantage of AuCoMe is that it keeps similar prediction performance with less known organisms compared to gapseq (as seen in Table S5).

2.5 Validation of filtering steps and GPR associations

100 random GPR associations were randomly selected and examined across the metabolic networks generated by AuCoMe for the algal dataset. Among them, 50 reactions that were predicted to be present and 50 reactions that were predicted to be absent from metabolic networks. Regarding the former, their first associated gene was manually annotated based on reciprocal BLAST searches against UniProt [1] and the presence of conserved domains and the result of this manual annotation was compared to the predicted metabolic reaction. For absent reactions, we searched for characterized proteins known to catalyze the reaction in question, and then performed reciprocal BLASTP searches with the corresponding algal proteome. Results are shown in Tables S6 and S7.

Reaction		**Organism**	**Genes associated**	**First gene verified**	**Manual annotation**	**Evaluation**	**Comment**
RXN-7568	carboxy-1,4-naphthoquinone	*Auxenochlorella protothecoides*	1	F751_5387	carboxy-1,4-naphthoquinone phytyltransferase	OK	
RXN-19353	RBR-type E3 ubiquitin transferase	*Bathycoccus prasinos*	1	Bathy13g01040	Probable E3 ubiquitin-protein ligase ARI7	OK	
3.6.4.7-RXN	peroxisome-assembly ATPase	*Caenorhabditis elegans*	1	NC_003279_8_01728	AFG1-like ATPase	OK	
2KETO-3METHYLVALERATE-RXN		*Chara braunii*	1	CBR_g46	Dihydrolipoamide S-acetyltransferase component 5 of pyruvate dehydrogenase	FALSE	
RXN-8001	histidinol dehydrogenase	*Chara braunii*	3	CBR_g8928	histidinol dehydrogenase	OK	

Manual validation of 50 randomly chosen reactions found in any of the species (continued on next page)

Reaction		**Organism**	**Genes associated**	**First gene verified**	**Manual annotation**	**Evaluation**	**Comment**
RXN-9537	3-hydroxyacyl-[acyl-carrier-protein] dehydratase	*Chlamydomonas reinhardtii*	1	CHLRE-DRAFT_102512	3-hydroxyacyl-[acyl-carrier-protein] dehydratase	OK	
STEROL-GLUCOSYL-TRANSFERASE-RXN	3β-hydroxy sterol glucosyltransferase	*Chlamydomonas reinhardtii*	3	CHLRE-DRAFT_154976	3β-hydroxy sterol glucosyltransferase	OK	
CATAL-RXN	Catalase	*Chlorella variabilis*	1	CHLNC-DRAFT_57817	Unknown/ glutamine amidotransferase	OK	
RXN-17608	Amidase	*Chlorella variabilis*	7	CHLNC-DRAFT_141670	Fatty acid amide hydrolase	OK	
1.2.1.25-RXN	Branched-chain alpha-keto acid dehydrogenase system	*Chrysochromulina tobin*	4	Ctob_011740	Branched-chain alpha-keto acid dehydrogenase system	OK	
RXN0-308	Cysteine desulfurase	*Chrysochromulina tobin*	5	Ctob_004932	Cysteine desulfurase	OK	
TRYPTOPHAN-AMINOTRANSFERASE-RXN	L-tryptophan aminotransferase	*Chrysochromulina tobin*	1	Ctob_005574	Unknown protein	FALSE	Probably rather 2.6.1.7
RXN-18532	long-chain fatty acid adenylyltransferase FadD28: fadD28	*Cladosiphon okamuranus*	1	C_okamuranus_06292	Long-chain-fatty-acid–AMP ligase	OK	
DCYSDESULF-RXN	D-cysteine desulfhydrase	*Cladosiphon okamuranus*	1	C_okamuranus_02325	Unknown protein	FALSE	
2-AMINOADIPATE-AMINOTRANSFERASE-RXN	kynurenine/alpha-aminoadipate aminotransferase	*Coccomyxa subellipsoidea*	1	COCSU-DRAFT_44586	kynurenine/alpha-aminoadipate aminotransferase	OK	
TIGLYLCOA-HYDROXY-RXN	enoyl-CoA hydratase	*Ectocarpus siliculosus*	9	E_siliculosus_12401	Unknown protein	OK	
3-ISOPROPYLMALISOM-RXN	isopropylmalate dehydratase	*Ectocarpus subulatus*	5	E_siliculosus_10344	isopropylmalate dehydratase	OK	

Manual validation of 50 randomly chosen reactions found in any of the species (continued on next page)

Reaction		**Organism**	**Genes associated**	**First gene verified**	**Manual annotation**	**Evaluation**	**Comment**
RXN-14270	Adenine phosphoribosyltransferase	*Ectocarpus subulatus*	1	E_siliculosus_18104	Unknown protein	OK	
RXN-19832	sterol C-14 demethylase: CYP51	*Emiliana huxleyi*	2	EMIHU-DRAFT_260386	sterol C-14 demethylase: CYP51	OK	
ASPARTATE-SEMIALDEHYDE-DEHYDROGENASE-RXN	aspartate semialdehyde dehydrogenase	*Fistulifera solaris*	1	FisN_14Lu089	aspartate semialdehyde dehydrogenase	OK	
L-LACTATE-DEHY-DROGENASE-CYTOCHROME-RXN	L-lactate dehydrogenase CYB2	*Fistulifera solaris*	2	FisN_5Hh051	Nitrate reductase	FALSE	Probably rather EC 1.1.3.15
RXN-12198	nucleoside diphosphate phosphatase	*Galdieria sulphuraria*	3	Gasu_48820	Soluble calcium-activated nucleotidase	OK	
5.99.1.2-RXN	DNA topoisomerase	*Gracilariopsis chorda*	15	BWQ96_09878	DNA topoisomerase (ATP-hydrolyzing)	OK	
RXN-12242	ζ-carotene desaturase	*Micractinium conductrix*	2	C2E20_0611	ζ-carotene desaturase	OK	
3.4.11.13-RXN	(clostridial) aminopeptidase	*Micromonas pusilla*	2	MICPUC-DRAFT_43781	Aspartyl aminopeptidase	OK	Instance of generic reaction
HISTALDEHYD-RXN	histidinol dehydrogenase	*Monoraphidium neglectum*	5	MNEG_8118	Aldehyde dehydrogenase EC 1.2.1.3	OK	
RXN-10700	3-ketoacyl-CoA thiolase	*Monoraphidium neglectum*	3	MNEG_15298	3-ketoacyl-CoA thiolase	OK	
RXN-18957	β-carotene ketolase/β-ring carotenoid 4-ketolase	*Monoraphidium neglectum*	1	MNEG_0808	β-carotene ketolase/β-ring carotenoid 4-ketolase	OK	
RXN0-1141	lipoate-protein ligase A	*Monoraphidium neglectum*	6	MNEG_4603	lipoate-protein ligase A	OK	
RXN-20894	Beta-lactamase 3.5.2.6	*Mucor circinelloides*	3	MUCCI-DRAFT_41384	Ribonuclease Z 3.1.26.11	OK	

Manual validation of 50 randomly chosen reactions found in any of the species (continued on next page)

Reaction		**Organism**	**Genes associated**	**First gene verified**	**Manual annotation**	**Evaluation**	**Comment**
RXN-9531	Beta-ketoacyl-[acyl-carrier-protein] synthase	*Nannochloropsis gaditana*	2	Naga_100004g8	L-aminoadipate-semialdehyde dehydrogenase-phosphopantetheinyl transferase 2.7.8.7	OK	
3.2.1.39-RXN	glucan endo-1,3-β-D-glucosidase	*Nemacystus decipiens*	6	g10939	glucan endo-1,3-β-D-glucosidase	OK	
L-GLN-FRUCT-6-P-AMINO-TRANS-RXN	Glutamine–fructose-6-phosphate transaminase	*Nemacystus decipiens*	3	g9659	Unkown protein	OK	
RXN-7253	Inositol-phosphate phosphatase	*Neurospora crassa*	5	NCU06022	Inositol-phosphate phosphatase 3.1.3.25	OK	
RXN66-316	sterol C4-methyl monooxygenase	*Neurospora crassa*	1	NCU06402	Methylsterol monooxygenase	OK	
RXN-11835	23S rRNA pseudouridine(2604)	*Ostreococcus tauri*	5	OT_ostta02g05280	40S ribosomal protein S4	OK	
ACYLCOA-SYN-RXN	long-chain acyl-CoA synthetase	*Phaeodactylum tricornutum*	5	PHATR-DRAFT_20143	long-chain acyl-CoA synthetase	OK	
RXN66-569	Phenylalanine 4-monooxygenase	*Porphyra umbilicalis*	1	KV918761_1_10819	Phenylalanine 4-monooxygenase	OK	
TRANS-RXN-366	ABC Transporter	*Porphyra umbilicalis*	20	KV918761_1_08330	ABC Transporter	OK	
3.2.2.24-RXN	ADP-ribosyl-[dinitrogen reductase]	*Raphidocelis subcapitata*	3	Rsub_01063	Unknown protein	FALSE	
DTDPKIN-RXN	nucleoside-diphosphate kinase	*Raphidocelis subcapitata*	6	Rsub_08794	Unknown protein	OK	
PGLYCDEHY-DROG-RXN	3-phosphoglycerate dehydrogenase	*Raphidocelis subcapitata*	3	Rsub_05207	3-phosphoglycerate dehydrogenase	OK	
ACETYLORN-TRANSAM-RXN	Acetylornithine transaminase	*Saccharina japonica*	3	SJ03992	TRAP transporter fused permease	FALSE	

Manual validation of 50 randomly chosen reactions found in any of the species (continued on next page)

Reaction		Organism	Genes associated	First gene verified	Manual annotation	Evaluation	Comment
RXN-14260	4-alpha-glucanotransferase	*Saccharina japonica*	1	SJ00859	Unknown protein	FALSE	
RXN-9530	Enoyl-[acyl-carrier-protein] reductase	*Saccharina japonica*	2	SJ00320	Enoyl-[acyl-carrier-protein] reductase	OK	
RXN-20610	[Histone H3]-lysine(4) N-trimethyltransferase	*Tetrabaena socialis*	22	TSOC_014450	Histone-lysine N-methyltransferase	OK	Incomplete sequence
RXN66-550	Electron-transferring-flavoprotein dehydrogenase	*Tetrabaena socialis*	1	TSOC_000826	Electron transfer flavoprotein-ubiquinone oxidoreductase	OK	
RXN-13398	Chlorophyllide a oxygenase	*Thalassiosira pseudonana*	2	THAPS-DRAFT_bd783	Chlorophyllide a oxygenase	OK	
LYSINE-N-ACETYLTRANSFERASE-RXN	Lysine N-acetyltransferase	*Volvox carteri*	1	VC00001G02580	Histone acetyltransferase	OK	
RXN-18356		*Saccharina japonica*	0	no gene - spontaneous			

Table S6: **Manual validation of 50 randomly chosen reactions found in any of the species.**

Reaction	Organism	UniProt identifiant	Evalue best UniProt hit	Hit description	Evaluation
2.4.1.214-RXN	*Auxenochlorella protothecoides*	Q9FX97	NO HIT		OK
RXN-16217	*Auxenochlorella protothecoides*	Q6RX91	NO HIT		OK
RXN-13159	*Bathycoccus prasinos*	B1VTI5	NO HIT		OK
DEOXYINOPHOSPHOR-RXN	*Bathycoccus prasinos*	P0ABP8	NO HIT		OK

Manual validation of 50 reactions absent from a species and randomly chosen (continued on next page)

Reaction	**Organism**	**UniProt identifiant**	**Evalue best UniProt hit**	**Hit description**	**Evaluation**
RXN-9805	*Clamydomonas reinhartii*	Q9C788	E-value 5×10^{-37}	>PNW87838.1 hypothetical protein CHLRE_01g003850v5 [Chlamydomonas reinhardtii]. Not best reciprocal hit (other CYP450).	OK
3.1.13.5-RXN	*Clamydomonas reinhartii*	P09155	NO HIT		OK
RXN-2	*Caenorhabditis elegans*	G2IQS7	NO HIT		OK
RXN-7660	*Caenorhabditis elegans*	Q9CA67	NO HIT		OK
RXN-12112	*Chara braunii*	P23295	NO HIT		OK
RXN0-6566	*Chara braunii*	P46850	NO HIT		OK
RXN-20552	*Chlorella variabilis*	Q8S8N6	NO HIT		OK
CARBOXY-CISCIS-MUCONATE-CYCLASE-RXN	*Chondrus crispus*	P38677	NO HIT		OK
1.2.7.4-RXN	*Chondrus crispus*	Q8TXF7	NO HIT		OK
RXN-8580	*Cladosiphon okamuranus*	P39849	NO HIT		OK
RXN66-623	*Cyanidioschyzon merolae*	P19097	NO HIT		OK
MANNONDE-HYDRAT-RXN	*Cyanophora paradoxa*	P24215	NO HIT		OK
ALLENE-OXIDE-CYCLASE-RXN	*Ectocarpus siliculosus*	Q9LS02	NO HIT		OK
LYSINE-N-ACETYLTRANSFERASE-RXN	*Fistulifera solaris*	P41929	NO HIT		OK
RXN-17811	*Fistulifera solaris*	Q8BGW1	NO HIT		OK
RXN-12423	*Fistulifera solaris*	Q9RGX8	NO HIT		OK
RXN-14796	*Fragilariopsis cylindrus*	P13711	E-value 1×10^{-59}	>OEU12659.1 acyl-CoA oxidase [Fragilariopsis cylindrus CCMP1102].	FALSE

Manual validation of 50 reactions absent from a species and randomly chosen (continued on next page)

Reaction	Organism	UniProt identifiant	Evalue best UniProt hit	Hit description	Evaluation
1.4.1.21-RXN	*Fragilariopsis cylindrus*	A6ND91	NO HIT		OK
RXN-8023	*Galdieria sulphuraria*	Q9JP98	NO HIT		OK
RXN-13680	*Galdieria sulphuraria*	P45381	NO HIT		OK
PROPION-LACT-RXN	*Gonium pectorale*	Q9L3F7	NO HIT		OK
RXN-15311	*Gonium pectorale*	D0E8I4	NO HIT		OK
2.4.1.47-RXN	*Gracilariopsis chorda*	Q16880	NO HIT		OK
15-OXOPROS-TAGLANDIN-13-REDUCTA-SE-RXN	*Guillardia theta*	Q8VDQ1	E-value 5×10^{-43}	>XP_005824337.1 hypothetical protein GUITHDRAFT_154983 [Guillardia theta CCMP2712].	FALSE
RXN-13374	*Klebsormidium nitens*	Q9HAY6	E-value 3×10^{-25}	>GAQ81854.1 carotenoid oxygenase [Klebsormidium nitens]. Best reciprocal hit, but more likely to be a Carotenoid 9,10(9',10')-cleavage dioxygenase not found in humans.	OK
RXN66-354	*Klebsormidium nitens*	P14060	E-value 1×10^{-24}	>GAQ83800.1 sterol-4alpha-carboxylate 3-dehydrogenase (decarboxylating) [Klebsormidium nitens]. Not best reciprocal hit (rather a Sterol-4-alpha-carboxylate 3-dehydrogenase).	OK
RXN-18388	*Micromonas pusilla*	P77718	NO HIT		OK
RXN-10720	*Micromonas pusilla*	Q16773	E-value 1×10^{-48}	>XP_003058689.1 aminotransferase class I and II [Micromonas pusilla CCMP1545], second BRH, probable Kynurenine–oxoglutarate transaminase.	FALSE
CHOLOYLGLY-CINE-HYDRO-LASE-RXN	*Nannochloropsis gaditana*	Q9KK62	NO HIT		OK
RXN-14553	*Nemacystus decipiens*	Q70IY1	NO HIT		OK
HEPARITIN-SULFOTRANS-FERASE-RXN	*Neurospora crassa*	P5284	NO HIT		OK
RXN3DJ-35528	*Phaeodactylum tricornutum*	O88816	NO HIT		OK

Manual validation of 50 reactions absent from a species and randomly chosen (continued on next page)

Reaction	**Organism**	**UniProt identifiant**	**Evalue best UniProt hit**	**Hit description**	**Evaluation**
RXN-11562	*Phaeodactylum tricornutum*	O60243	NO HIT		OK
RXN-5901	*Saccharina japonica*	D3UAK7	NO HIT		OK
RXN-9909	*Saccharina japonica*	P39849	NO HIT		OK
3.4.13.19-RXN	*Tetrabaena socialis*	P16444	NO HIT		OK
RXN-11058	*Thalassiosira pseudonana*	P50226	E-value 1×10^{-25}	>XP_002292441.1 predicted protein [Thalassiosira pseudonana CCMP1335]. This seems to be a valid sulfotransferase with complete Sulfotransfer_3 domain.	FALSE
RXN-19006	*Ulva mutabilis*	Q9LFR0	NO HIT		OK
1.1.1.265-RXN	*Volvox carteri*	Q12068	E-value 1×10^{-23}	>XP_002956636.1 heme peroxidase-related protein [Volvox carteri f. nagariensis] Not BRH. Hit rather phenylacetaldehyde dehydrogenase or similar.	OK
PDXJ-RXN	*Volvox carteri*	P0A794	NO HIT		OK
RXN-14850	*Chondrus crispus*	spontaneous			spontaneous
RXN-15140	*Chara braunii*	spontaneous			spontaneous
RXN-20576	*Bathycoccus prasinos*	spontaneous			spontaneous
RXN-20079	*Gonium pectorale*	spontaneous			spontaneous
RXN-18894	*Micractinium conductrix*	spontaneous			spontaneous
RXN-18431	*Neurospora crassa*	spontaneous			spontaneous

Table S7: **Manual validation of 50 reactions absent from a species and randomly chosen.**

2.6 Validation of EC numbers with deep-learning approaches

An important feature of the AuCoMe approach is to complement annotation-based GSMNs (draft reconstruction) with reactions supported by orthology assessment. However, to avoid the propagation of erroneous gene

annotations leading to misleading gene-reaction associations, the method introduces a robustness criterion to filter gene-reaction associations predicted by orthologies. As shown in the Fig. 4 of the paper, this robustness criterion is crucial to separate algae from fungi in terms of photosynthesis capabilities. To study the interest of this robustness criterion, we compared non-robust GPR associations with robust GPR associations according to the predictions of EC numbers with the DeepEC tool [10], which was used as an independent validation of the predictions made according to orthology propagation. As plotted in Fig. S10, in the fungal dataset, the annotation and the orthology steps predicted on average 1,230 EC numbers for each GSMN after applying the robustness criteria, among which, 404 (32%) were consistent with the predictions of DeepEC. In the algal dataset, 1,435 EC numbers were predicted in average for each GSMN, among which, 312 (22%) were consistent with the predictions of DeepEC. When running DeepEC on the non-robust GPRs, the percentage of consistent predictions was only 4.1 and 1.4 %, respectively.

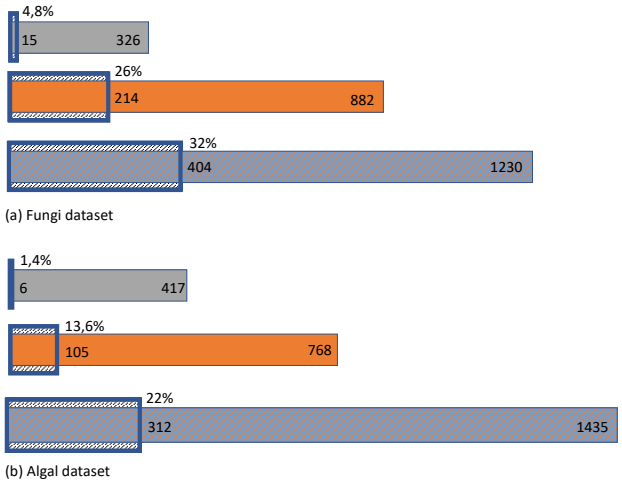

Figure S10: **Average number (among the fungal and the algal datasets) of EC numbers for GPR associations predicted by the orthology step of AuCoMe, and average number of EC numbers predicted by the DeepEC tool for the corresponding gene families.** Grey: EC number of reactions associated only with non-robust GPRs from the orthology step. Orange: EC numbers of robust GPR associations predicted only at the orthology step. Blue/gray degraded: EC number of reactions associated with robust GPR associations predicted either at the annotation or at the orthology step.

2.7 Exploration of Calvin cycle and pigment pathways in algae

The genomes can be clustered into three groups with respect to the Calvin Cycle. The first group, corresponding to archeplastids and containing the green algal (Viridiplantae), red algal (Rhodophyta), and *Cyanophora*

paradoxa genomes, was predicted to contain 12 reactions of the pathway. The second group consisted of the four outgroup organisms. The last group contained brown algae, diatoms, haptophytes, *Guillardia theta* and *Nannochloropsis gaditana* that lack the 1.2.1.13-RXN reaction associated with the enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH). The GAPDH reactions have been propagated to green algae, red algae and *C. paradoxa* because they share a common ortholog associated with this reaction, the plastid gene *GapA*, whereas this gene has not been found in diatoms and brown algae. Here a duplicated cytosolic GAPDH can functionally replace this gene [6]. The presence of this gene was validated using sequence alignment with a known cytosolic GAPDH and the reaction was manually added in yellow.

None of the 40 GSMNs contained the reaction SEDOBISALDOL-RXN. In the MetaCyc v23.5 database, which was used for the experiment, this reaction was associated with the EC number 4.1.2.X, whereas it was associated with the EC number 4.1.2.13 in the KEGG [4] and Brenda [11] databases. This prevented the draft reconstruction tool (Pathway Tools v23.5) from associating genes annotated with EC 4.1.2.13 to this reaction. We therefore manually added the corresponding GPR associations, indicated in yellow in Fig. 4 of the main text.

We also used the results of AuCoMe to explore pigment pathways of five brown algae, *Cladosiphon okamuranus*, *Ectocarpus siliculosus*, *E. subulatus*, *Nemacystus decipiens*, and *Saccharina japonica*. These pathways correspond to synthesize heme groups and to catalyze three alternative pigment biosynthesis pathways phycoerythrobilin (PWY-5915), phycocyanobilin (PWY-5917), and phytochromobilin (PWY-7170) (Fig. S11). These pathways consist of four, three, and three reactions, respectively, and all have protoheme as a starting point. They also share the same starting reaction (RXN-17523) that transforms heme into biliverdin in the presence of a reduced ferredoxin (Fig. S11).

The draft reconstruction and the orthology propagation steps contributed the most to the predicted reactions in the pigment pathways. The spontaneous step (red in Fig. S11) retrieved two reactions. RXN-13968 was not found in any of the analyzed GSMNs, hence we added it manually to all the five brown algae. RXN-13968 [5] was not considered spontaneous but had no associated enzymes in the MetaCyc database and hence was not found in any of the analyzed algal networks. We manually added this reaction to the GSMNs of all five brown algae.

Two of the studied pathways generate phycobilins with the enzymes encoded by the genes *pebA*, *pebB*, *pcyA* and *HY2* (Supplemental Fig. S11). In red algae, cryptophytes, and cyanobacteria, phycobilins bind to light-harvesting proteins called phycobiliproteins. Brown algae are thought to have lost phycobiliproteins

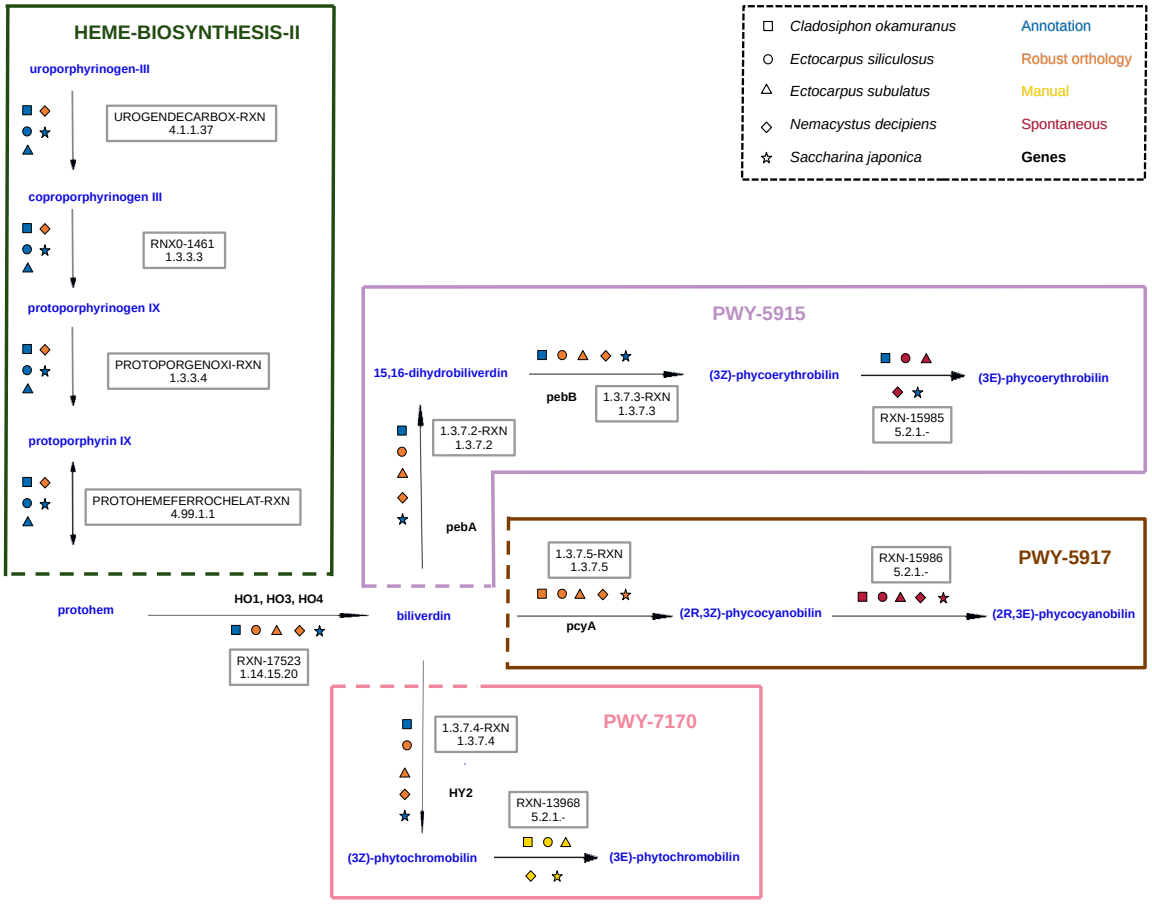

Figure S11: **Prediction of pigment pathways in brown algae.** Links between four MetaCyc pigment pathways are presented: heme b biosynthesis (HEME-BIOSYNTHESIS-II, dark green frame), phycoerythrobilin biosynthesis (PWY-5915, mauve frame), phycocyanobilin biosynthesis (PWY-5917, brown frame), and phytochromobilin biosynthesis (PWY-7170, pink frame). PWY-5915, PWY-5917, and PWY-7170 are three alternative biosynthetic pathways of bilin proteins. Here we detailed every AuCoMe step employed to complete the GSMNs of five brown algae, from the GPR associations found with draft reconstruction (blue), via GPR associations specifically found by orthology propagation, to GPR associations that passed the filter (orange). In the end, AuCoMe also added missing spontaneous reactions to GSMNs (see Materials & Methods). Furthermore, we chose to manually add a few reactions (gold). We focused on five brown algae: *C. okamuranus* (square), *E. siliculosus* (circle), *E. subulatus* (triangle), *N. decipiens* (diamond), and *S. japonica* (star). Genes are noted in bold black.

during evolution [2]. However they have, in many cases, retained *pebA*, *pebB*, and *HY2* [9, 8]. According to AuCoMe, the *Ectocarpus siliculosus pebA* and *pebB* genes were also associated with the reactions generally attributed to the *HY2* and *pcyA* genes. This likely constitutes an artifact since these proteins share nearly 30% of sequence identity with *pebA* and *pebB*. These genes, although inherited from a red alga, have diversified and may now be used to make photoreceptors rather than photosynthetic pigments [7].

2.8 AuCoMe GSMNs are consistent with species phylogeny

In order to get a global view of the metabolic differences underlying the metabolic dendrogram calculated after the AuCoMe step (see right part of Fig. 5B in the main text), we plotted the presence-absence of metabolic reactions in a supervenn diagram (Fig. S12).

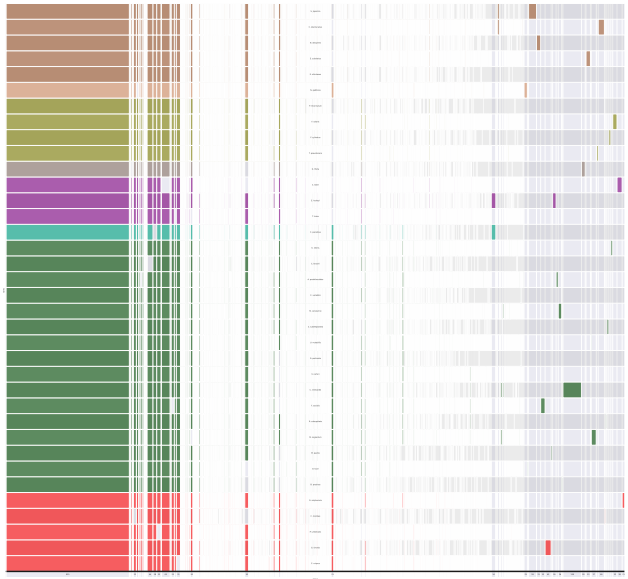

Figure S12: **Presence and absence of reactions in the pan-metabolism of the algal dataset**. Each row corresponds to a genome, ordered by phylogenetic group. Brown: brown algae. Yellow: Eustigmatophytes. Purple: Cyanidophora. Green: green algae. Red: red algae. Each column corresponds to a reaction of the pan-metabolism. A grey cell means than the corresponding reaction is not predicted to belong to the GSMN of the considered organism.

We focused on a cluster of fourteen reactions common to the brown algae *Cladosiphon okamuranus* and *Saccharina japonica*, but absent in other brown algal GSMNs (see Supplemental Table S8). Four of these reactions correspond to putative enzymatic functions (glucuronate reductase, o-aminophenol oxidase, 11-oxo-β-amyrin 30-oxidase and keratan sulfotransferase), whereas two were predicted to occur spontaneously (RXN-13160 and RXN-17356). There was no obvious connection between these reactions as they are assigned

to metabolic pathways involving different molecules. The presence of the four enzymatic reactions in *C. okamuranus* and *S. japonica* was assigned based on annotations, but orthology propagation in the AuCoMe pipeline identified only a subset of the potential orthologs (see Supplemental Table S9). Complementary analysis by BLASTP searches, however, showed that potential homologs of those four proteins were present in other brown algae, some stramenopile microalgae, and *Symbiodinium* dinoflagellates (see Supplemental Fig. S13). These results suggest that the corresponding enzymes potentially exhibit other substrate or activity specificities in stramenopiles. The o-aminophenol oxidase family proteins present in the genome of *Ectocarpus siliculosus* are predicted to be cytoplasmic, extracellular, or to target the membrane (see Supplemental Table S10), suggesting different roles depending on their subcellular localization. In this case, AuCoMe, with the support of more focused analyses, led to the identification of numerous candidate o-aminophenol oxidases in stramenopiles, whose biochemical functions and biological roles remain to be discovered.

Then we also looked at the basis for the unstable phylogenetic position of the cryptophyte *Guillardia theta*, with a detailed exploration of reactions distinguishing the cryptophyte, haptophyte, stramenopile, and archeplastida groups (Supplemental Table S11), and also shared metabolic pathways between those groups (Supplemental Table S12).

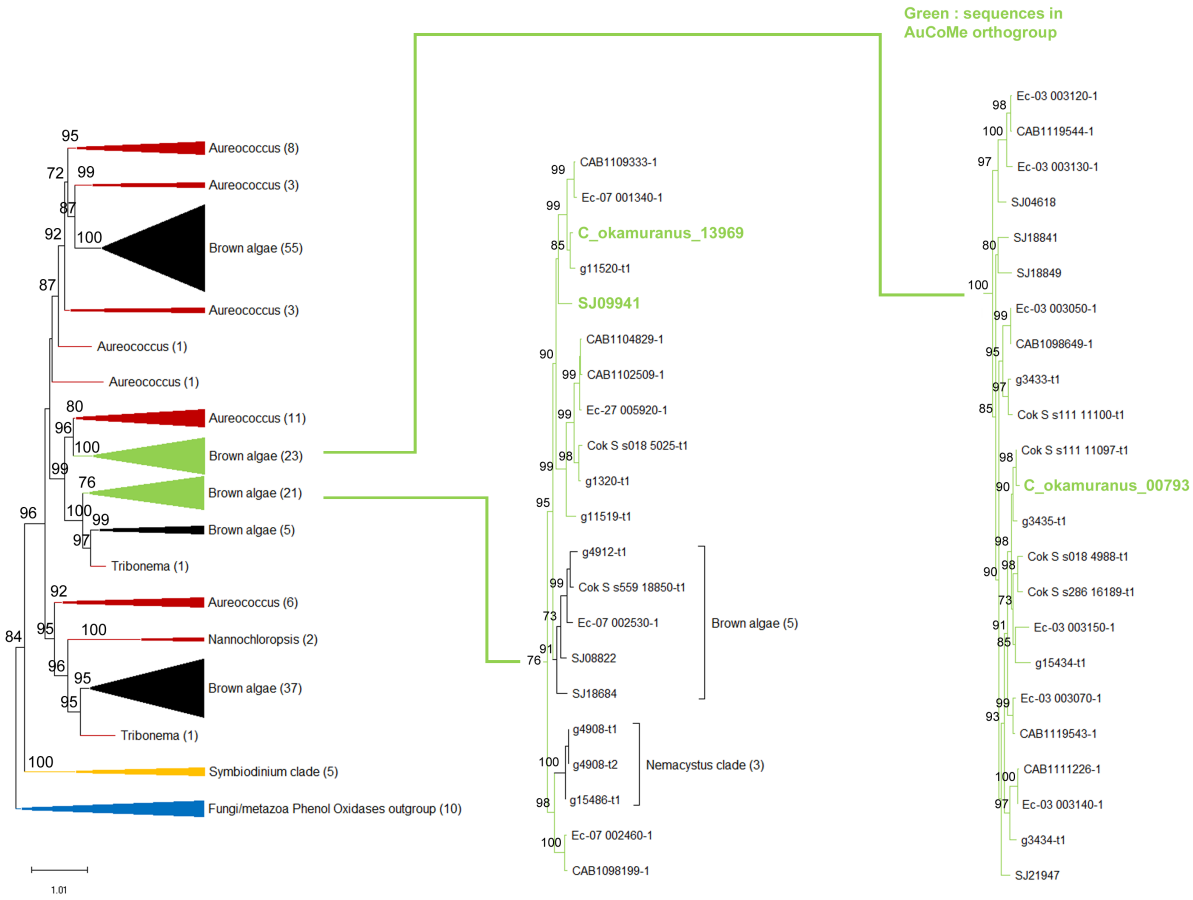

Figure S13: **Phylogenetic distribution of brown algal o-Aminophenol Oxidases**. The brown algae are represented by *Ectocarpus siliculosus* (38 sequences) *Ectocarpus sp.* CCAP 1310/34 (21 sequences), *Cladosiphon okamuranus* (21 sequences), *Saccharina japonica* (21 sequences) and *Nemacystus decipiens* (40 sequences) are in black color. Other stramenopile microalgae, *Aureococcus anophagefferens* (33 sequences), *Tribonema minus* (2 sequences), and *Nannochloropsis oceanica* (2 sequences) are shown in brown color. The Alveolate (5 sequences) and Fungi-Metazoa (10 sequences) groups are displayed in yellow and in blue colors, respectively. The number of sequences grouped in each clade is given in parentheses. Bootstrap values of 100 replicates are indicated above the branches and the distance scale is defined below the dendrogram. An expanded view is provided for the two separate clades containing phenoloxidases identified as the initial orthogroup by AuCoMe (in green).

REACTION NAME	**EC NUMBER**	**PATHWAYS**	**ENZYME HOMOLOG**	**REACTION(S)**	***Cladosiphon okamuranus* GENE ASSOCIATION**	***Saccharina japonica* GENE ASSOCIATION**
GLUCURONATE-REDUCTASE-RXN	1.1.1.19	D-glucuronate degradation I, L-ascorbate biosynthesis IV (animals, D-glucuronate pathway), L-ascorbate biosynthesis VI (plants, myo-inositol pathway)	*Mus musculus*: alcohol dehydrogenase [$NADP^+$] Inferred from experiment: Akr1a4	L-gulonate + $NADP^+$ ⇐ aldehydo-D-glucuronate + NADPH + H^+	Found at scaffold_218 position 23140-23316	SJ05644
KERATAN-SULFOTRANSFERASE-RXN	2.8.2.21	None (keratan sulfate biosynthesis)	*Homo sapiens*: carbohydrate sulfotransferase Inferred from experiment: CHST1	n 3'-phosphoadenylyl-sulfate + a keratan ⇒ n adenosine 3',5'-bisphosphate + a keratan sulfate + n H^+	C_okamuranus_09447	SJ14732
RXN-13160	None	None	This reaction can occur spontaneously – no enzyme is required.	4 1,2-Benzoquinone monoimine + O2 ⇔ 2 Isophenoxazine + 2 H_2O	None	None
RXN-17356	no EC number assigned	aurachin A, B, C and D biosynthesis	This reaction can occur spontaneously – no enzyme is required.	aurachin C epoxide ⇒ 4-hydroxy-2-methyl-4-[(2E,6E)-3,7,11-trimethyldodeca-2,6,10-trien-1-yl]quinolin-3(4H)-one 1-oxide	None	None
RXN-13494	1.14.14.115	glycyrrhetinate biosynthesis		glycyrrhetaldehyde + a reduced [NADPH-hemoprotein reductase] + oxygen ⇒ glycyrrhetinate + an oxidized [NADPH-hemoprotein reductase] + H_2O + H^+	C_okamuranus_03849 C_okamuranus_03853 C_okamuranus_08689 C_okamuranus_03848	SJ11349 SJ09087 SJ11348 SJ11355

Reactions common to *Saccharina japonica* and *Cladosiphon okamuranus* but not found in other brown algae (continued on next page)

REACTION NAME	**EC NUMBER**	**PATHWAYS**	**ENZYME HOMOLOG**	**REACTION(S)**	***Cladosiphon okamuranus* GENE ASSOCIATION**	***Saccharina japonica* GENE ASSOCIATION**
RXN-13506		None	Thecc.07G140900 Inferred by computational analysis: Thecc.07G140900 Thecc.07G140800 Inferred by computational analysis: Thecc.07G140800 Thecc.08G023700 Inferred by computational analysis: Thecc.08G023700 Thecc.09G079300 Inferred by computational analysis: Thecc.09G079300 Thecc.04G203900 Inferred by computational analysis: Thecc.0 4G203900 Thecc.04G249700 Inferred by computational analysis: Thecc.04G249700	11-oxo-β-amyrin + a reduced [NADPH-hemoprotein reductase] + oxygen ⇒ 30-hydroxy-11-oxo-β-amyrin + an oxidized [NADPH-hemoprotein reductase] + H_2O, 30-hydroxy-11-oxo-β-amyrin + a reduced [NADPH-hemoprotein reductase] + oxygen ⇒ glycyrrhetaldehyde + an oxidized [NADPH-hemoprotein reductase] + 2 H_2O,		
	1.14.14.115		Thecc.10G138900 Inferred by computational analysis: Thecc.10G138900 Thecc.02G185500 Inferred by computational analysis: Thecc.02G185500 Thecc.02G324100 Inferred by computational analysis: Thecc.02G324100 Thecc.02G323900 Inferred by computational analysis: Thecc.02G323900 Thecc.02G324600 Inferred by computational analysis: Thecc.02G324600	glycyrrhetaldehyde + a reduced [NADPH-hemoprotein reductase] + oxygen ⇒ glycyrrhetinate + an oxidized [NADPH-hemoprotein reductase] + H_2O + H^+	C_okamuranus_03849 C_okamuranus_03853 C_okamuranus_08689 C_okamuranus_03848	SJ11349 SJ11355 SJ09087 SJ11348
RXN-13492		glycyrrhetinate biosynthesis	*Glycyrrhiza uralensis*: 11-oxo-β-amyrin 30-oxidase Inferred from experiment: CYP72A154	11-oxo-β-amyrin + 3 a reduced [NADPH-hemoprotein reductase] + 3 oxygen ⇒ glycyrrhetinate + 3 an oxidized [NADPH-hemoprotein reductase] + 4 H_2O + H^+		

Reactions common to *Saccharina japonica* and *Cladosiphon okamuranus* but not found in other brown algae (continued on next page)

REACTION NAME	**EC NUMBER**	**PATHWAYS**	**ENZYME HOMOLOG**	**REACTION(S)**	***Cladosiphon okamuranus* GENE ASSOCIATION**	***Saccharina japonica* GENE ASSOCIATION**
RXN-13493	1.14.14.115	glycyrrhetinate biosynthesis	*Glycyrrhiza uralensis*: 11-oxo-β-amyrin 30-oxidase Inferred from experiment: CYP72A154	30-hydroxy-11-oxo-β-amyrin + 3 a reduced [NADPH-hemoprotein reductase] + 3 oxygen ⇒ glycyrrhetaldehyde + 3 an oxidized [NADPH-hemoprotein reductase] + 4 H2O + H^+	C_okamuranus_03849 C_okamuranus_03853 C_okamuranus_08689 C_okamuranus_03848	SJ11349 SJ11355 SJ09087 SJ11348
RXN-13868	1.10.3.15	grixazone biosynthesis	*Streptomyces griseus*: grixazone synthase Inferred from experiment: griF	2 O_2 + Acetylcysteine + 2 3-amino-4-hydroxybenzaldehyde ⇒ 3 H_2O + Formate + H^+ + 1-[[2-(acetylamino)-2-carboxyethyl]thio]-2-amino-3-oxo-8-formyl-3H-phenoxiazine		
RXN-15414		grixazone biosynthesis	*Streptomyces griseus*: grixazone synthase Inferred from experiment: griF	2 3-amino-4-hydroxybenzoate + N-acetyl-L-cysteine + 2 oxygen + H^+ ⇒ grixazone B + CO_2 + 4 H_2O grixazone B = 1-[[2-(acetylamino)-2-carboxyethyl]thio]-2-amino-3-oxo-8-carboxyl-3H-phenoxiazine		
RXN-17077	1.10.3.4	None	*Streptomyces antibioticus*: phenoxazinone synthase Inferred from experiment: phsA	4 3-hydroxy-4-methyl-anthranilate + 3 oxygen ⇒ 2 actinocin + 6 H_2O	C_okamuranus_13969 C_okamuranus_00793	SJ09941
O-AMINOPHENOL-OXIDASE-RXN		None	*Streptomyces griseus*: grixazone synthase Inferred from experiment: griF	4 2-aminophenol + 2 oxygen ⇒ 4 1,2-benzoquinone monoimine + 4 H_2O		
RXN-17067		actinomycin D biosynthesis	*Streptomyces antibioticus*: phenoxazinone synthase Inferred from experiment: phsA	4 3-hydroxy-4-methyl-anthranilate pentapeptide lactone + 3 oxygen ⇒ 2 actinomycin D + 6 H_2O		
RXN-13159	no EC number assigned	None	*Streptomyces griseus*: grixazone synthase Inferred from experiment: griF	4 2-aminophenol + 2 oxygen ⇒ 4 1,2-benzoquinone monoimine + 4 H_2O		

Table S8: **Reactions common to *Saccharina japonica* and *Cladosiphon okamuranus* but not found in other brown algae.** When several reaction variants are associated with a given set of sequences, all are indicated. Reaction classes are separated by two lines.

Enzyme type	**Pathway**	**Enzyme Type Reaction (ETR)**	***S. japonica* GPR association**	**Best *S. japonica* ETR BLASTP hit**	***E. siliculosus* homologs**
GLUCURONATE-REDUCTASE	D-glucuronate degradation, L-ascorbate biosynthesis	Akr1a4	SJ05644 (39%/95%)	SJ05644	Ec-03_003640.1 (79% id; 99% cover)
O-AMINOPHENOL-OXIDASE grixazone synthasephenoxazinone synthase	grixazone biosynthesisactinomycin D biosynthesis	griFphsA	SJ09941 (29%/17%) None (no identity)	SJ10714 (26%/35%) None (no identity)	Ec-07_001340.1 (51% id ;91% cover) Ec-27_005920.1 (53% id; 77% cover) Ec-07_002530.1 (51% id; 100% cover)
CYTOCHROME P450	glycyrrhetinate biosynthesis	CYP72A154	SJ11349 (24%/61%)	SJ07228 (26%/72%)	Ec-12_000880.1 (48%id; 57% cover) Ec-20_004230.1 (46% id; 50% cover)
			SJ09087 (28%/69%)		Ec-20_004230.1 (53% id; 94% cover) Ec-12_000880.1 (51% id; 100% cover)
			SJ11348 (24%/48%)		Ec-12_000880.1 (47% id; 67% cover) Ec-20_004230.1 (44% id; 64% cover)
			SJ11355 (22%/93%)		Ec-12_000880.1 (47% id; 86% cover) Ec-20_004230.1 (45% id; 83% cover)
KERATAN-SULFOTRANSFERASE	None (keratan sulfate biosynthesis)	CHST1	SJ14732 (28%/33%)	SJ17294 (23%/56%)	No orthologue in Esil proteins

Table S9: **Additional homologs in *E. siliculosus* found by BLASTP searches for sequences inferred to be present only in *C. okamuranus* and *S. japonica*.**

					Best homologs in stramenopile species (Amino Acid identity %)				
Sequences producing significant alignments with SJ09941	**AA number**	**Topology**	**Putative localization (by HECTAR)**	**Conserved domains**	**UniProt *Ectocarpus sp.* CCAP 1310/34**	***Cladosiphon okamuranus***	***Nemacystus decipiens***	***Saccharina japonica***	**NCBI *Aureococcus anophagefferens***
Ec-00_008280.1 Tyrosinase copper-binding domain (557) ;mRNA;... 149 5×10^{-36}	556	soluble protein (no transmembrane)	Cytoplasm (no signal peptide)	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5KV26 (45%)	Cok_S_s096_10419.t1 (59%)	g5891.t1 (59%)	SJ01153 (48%)	XP_009039249.1 (31%)
Ec-01_002800.1 Tyrosinase copper-binding domain (695) ;mRNA;... 184 1×10^{-46}	694	N-t Signal peptide SPI extracellular domain	Extracellular	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5K485 (70%)	Cok_S_s485_18441.t1 (66%)	g7402.t1 (70%)	SJ10707 (54%)	XP_009037689.1 (33%)

Additional o-aminophenol oxidases from *E. siliculosus* and their stramenopile homologs (continued on next page)

					Best homologs in stramenopile species (Amino Acid identity %)				
Sequences producing significant alignments with SJ09941	**AA number**	**Topology**	**Putative localization (by HECTAR)**	**Conserved domains**	**UniProt *Ectocarpus sp.* CCAP 1310/34**	***Cladosiphon okamuranus***	***Nemacystus decipiens***	***Saccharina japonica***	**NCBI *Aureococcus anophagefferens***
Ec-01_004540.1 Uncharacterised domain, di-copper centre (714... 573) 1×10^{-163}	713	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5L614 (92%)	Cok_S_s057_8386.t1 (73%)	g14528.t1 (75%)	SJ19488 (69%)	XP_009034914.1 (31%)
Ec-03_003050.1 Uncharacterised domain, di-copper centre (690... 319) 4×10^{-87}	689	N-t Signal peptide SPI extracellular domain	Extracellular	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JHA2 (95%)	Cok_S_s111_11097.t1 (70%)	g3433.t1 (66%)	SJ18841 (63%)	XP_009034914.1 (39%)
Ec-03_003070.1 Uncharacterised domain, di-copper centre (678... 286) 3×10^{-77}	677	Possible N-t SPI extracellular domain	Extracellular	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5L3M0 (87%)	Cok_S_s111_11097.t1 (75%)	g3435.t1 (69%)	SJ18841 (58%)	XP_009034914.1 (39%)
Ec-03_003120.1 Uncharacterised domain, di-copper centre (663... 288) 8×10^{-78}	662	N-t Signal peptide SPI extracellular domain	Extracellular	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5LKB4 (94%)	Cok_S_s111_11097.t1 (58%)	g3433.t1 (57%)	SJ04618 (70%)	XP_009034914.1
Ec-03_003130.1 Uncharacterised domain, di-copper centre (681... 293) 2×10^{-79}	680	Possible N-t SPI extracellular domain	Extracellular	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5LKB4 (78%)	Cok_S_s018_4988.t1 (60%)	g3433.t1 (58%)	SJ04618 (71%)	XP_009036795.1

Additional o-aminophenol oxidases from *E. siliculosus* and their stramenopile homologs (continued on next page)

					Best homologs in stramenopile species (Amino Acid identity %)				
Sequences producing significant alignments with SJ09941	**AA number**	**Topology**	**Putative localization (by HECTAR)**	**Conserved domains**	**UniProt *Ectocarpus sp.* CCAP 1310/34**	***Cladosiphon okamuranus***	***Nemacystus decipiens***	***Saccharina japonica***	**NCBI *Aureococcus anophagefferens***
Ec-03_003140.1 Tyrosinase (675) ;mRNA; f:3632027-3643304 280 3×10^{-75}	674	N-t Signal peptide SPI extracellular domain	Extracellular	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre IPR000601 PKD domain	A0A6H5KFJ6 (88%)	Cok_S_s111_11097.t1 (71%)	g3434.t1 (74%)	SJ18841 (62%)	XP_009034914.1
Ec-03_003150.1 Uncharacterised domain, di-copper centre (651... 285 9×10^{-77}	650	N-t Signal peptide SPI extracellular domain	Extracellular	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5L3M0 (62%)	Cok_S_s111_11097.t1 (68%)	g3435.t1 (65%)	SJ18841 (60%)	XP_009034914.1
Ec-03_004260.1 Uncharacterised domain, di-copper centre (683... 177) 2×10^{-44}	682	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5K485 (66%)	Cok_S_s485_18441.t1 (64%)	g7402.t2 (73%) g7402.t1 (73%)	SJ10707 (55%)	None
Ec-04_000730.1 Tyrosinase copper-binding domain (769) ;mRNA;... 126 5×10^{-29}	768	extracellular domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5KPW4 (90%)	Cok_S_s298_16368.t1 (96%)	g3842.t1 (51%)	SJ01153 (41%)	None

Additional o-aminophenol oxidases from *E. siliculosus* and their stramenopile homologs (continued on next page)

					Best homologs in stramenopile species (Amino Acid identity %)				
Sequences producing significant alignments with SJ09941	**AA number**	**Topology**	**Putative localization (by HECTAR)**	**Conserved domains**	**UniProt *Ectocarpus sp.* CCAP 1310/34**	***Cladosiphon okamuranus***	***Nemacystus decipiens***	***Saccharina japonica***	**NCBI *Aureococcus anophagefferens***
Ec-04_000780.1 Uncharacterised domain, di-copper centre (580... 135) 7×10^{-32}	579	soluble protein (no transmembrane)	Cytoplasm (no signal peptide)	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5KMG1 (93%)	Cok_S_s298_16368.t1 (71%)	g3842.t1 (58%)	SJ01153 (47%)	None
Ec-04_000790.1 Uncharacterised domain, di-copper centre (621... 151) 1×10^{-36}	620	soluble protein (no transmembrane)	Cytoplasm (no signal peptide)	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5KV26 (91%)	Cok_S_s298_16367.t1 (60%)	g3842.t1 (51%)	SJ01153 (58%)	None
Ec-05_001830.1 Uncharacterised domain, di-copper centre (631... 163) 3×10^{-40}	630	soluble protein (no transmembrane)	Cytoplasm (no signal peptide)	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5KV26 (44%)	Cok_S_s096_10419.t1 (57%)	g5891.t1 (59%)	SJ01153 (49%)	None
Ec-05_001900.1 Uncharacterised domain, di-copper centre (717... 154) 3×10^{-37}	716	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5KV26 (44%)	Cok_S_s096_10419.t1 (54%)	g5891.t1 (53%)	SJ01153 (47%)	XP_009039249.1
Ec-07_001340.1 Tyrosinase copper-binding domain (751) ;mRNA;... 674 0.0	750	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5KKI9 (86%)	Cok_S_s097_10482.t1 (57%)	g11520.t1 (60%)	SJ08822 (46%)	XP_009034555.1 XP_009032956.1

Additional o-aminophenol oxidases from *E. siliculosus* and their stramenopile homologs (continued on next page)

					Best homologs in stramenopile species (Amino Acid identity %)				
Sequences producing significant alignments with SJ09941	**AA number**	**Topology**	**Putative localization (by HECTAR)**	**Conserved domains**	**UniProt *Ectocarpus sp.* CCAP 1310/34**	***Cladosiphon okamuranus***	***Nemacystus decipiens***	***Saccharina japonica***	**NCBI *Aureococcus anophagefferens***
Ec-07_002460.1 Uncharacterised domain, di-copper centre (693... 542) 2×10^{-154}	692	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JFW1 (87%)	Cok_S_s559_18850.t1 (46%)	g4908.t2 (57%)	SJ08822 (44%)	XP_009034914.1 (34%)
Ec-07_002530.1 Tyrosinase copper-binding domain (616) ;mRNA;... 635 0.0	615	soluble protein (no transmembrane)	Cytoplasm (no signal peptide)	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5KKI9 (49%)	Cok_S_s559_18850.t1 (69%)	g11520.t1 (51%)	SJ08822 (61%)	XP_009034914.1 (30%)
Ec-07_003560.1 Uncharacterised domain, di-copper centre (114... 123) 4×10^{-28}	113	soluble protein (no transmembrane)	Cytoplasm (no signal peptide)	IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JTC5 (84%)	Cok_S_s559_18850.t1 (63%)	g1320.t1 (72%)	SJ08822 (58%)	XP_009038595.1 (42%)
Ec-07_005640.1 Tyrosinase copper-binding domain (789) ;mRNA;... 162 7×10^{-40}	788	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5K0B2 (94%)	Cok_S_s071_9172.t1 (72%)	g10835.t1 (70%)	SJ04880 (62%)	None
Ec-09_001110.1 Tyrosinase (776) ;mRNA; r:1342210-1351305 120 4×10^{-27}	775	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JVU3 (92%)	Cok_S_s298_16368.t1 (51%)	g2489.t1 (64%)	SJ01153 (42%)	None

Additional o-aminophenol oxidases from *E. siliculosus* and their stramenopile homologs (continued on next page)

					Best homologs in stramenopile species (Amino Acid identity %)				
Sequences producing significant alignments with SJ09941	**AA number**	**Topology**	**Putative localization (by HECTAR)**	**Conserved domains**	**UniProt *Ectocarpus sp.* CCAP 1310/34**	***Cladosiphon okamuranus***	***Nemacystus decipiens***	***Saccharina japonica***	**NCBI *Aureococcus anophagefferens***
Ec-20_004210.1 Uncharacterised domain, di-copper centre (754... 167) 2×10^{-41}	753	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (59%)	Cok_S_s173_13265.t1 (56%)	g13002.t1 (57%)	SJ19049 (53%)	XP_009037689.1 (33%)
Ec-20_004370.1 Tyrosinase copper-binding domain (786) ;mRNA;... 188 9×10^{-48}	785	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (57%)	Cok_S_s173_13265.t1 (63%)	g13001.t1 (57%)	SJ10707 (60%)	XP_009037689.1 (34%)
Ec-20_004380.1 Tyrosinase copper-binding domain (801) ;mRNA;... 189 4×10^{-48}	800	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (89%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (69%)	SJ10707 (62%)	XP_009037689.1 (35%)
Ec-20_004410.1 Tyrosinase copper-binding domain (802) ;mRNA;... 186 3×10^{-47}	801	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (90%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (70%)	SJ10707 (62%)	XP_009037689.1 (34%)
Ec-20_004410.2 Tyrosinase copper-binding domain (802) ;mRNA;... 187 1×10^{-47}	801	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (89%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (70%)	SJ10707 (62%)	XP_009037689.1 (34%)

Additional o-aminophenol oxidases from *E. siliculosus* and their stramenopile homologs (continued on next page)

					Best homologs in stramenopile species (Amino Acid identity %)				
Sequences producing significant alignments with SJ09941	**AA number**	**Topology**	**Putative localization (by HECTAR)**	**Conserved domains**	**UniProt *Ectocarpus sp.* CCAP 1310/34**	***Cladosiphon okamuranus***	***Nemacystus decipiens***	***Saccharina japonica***	**NCBI *Aureococcus anophagefferens***
Ec-20_004410.3 Tyrosinase copper-binding domain (802) ;mRNA;... 186 3×10^{-47}	801	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (89%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (70%)	SJ10707 (62%)	XP_009037689.1 (34%)
Ec-20_004410.4 Tyrosinase copper-binding domain (802) ;mRNA;... 187 2×10^{-47}	801	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (89%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (69%)	SJ10707 (62%)	XP_009037689.1 (34%)
Ec-20_004410.5 Tyrosinase copper-binding domain (802) ;mRNA;... 187 1×10^{-47}	801	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (89%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (70%)	SJ10707 (62%)	XP_009037689.1 (34%)
Ec-20_004410.6 Tyrosinase copper-binding domain (802) ;mRNA;... 190 2×10^{-48}	801	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (88%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (69%)	SJ10707 (61%)	XP_009037689.1 (34%)
Ec-20_004410.7 Tyrosinase copper-binding domain (802) ;mRNA;... 187 1×10^{-47}	801	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (89%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (70%)	SJ10707 (61%)	XP_009037689.1 (34%)

Additional o-aminophenol oxidases from *E. siliculosus* and their stramenopile homologs (continued on next page)

					Best homologs in stramenopile species (Amino Acid identity %)				
Sequences producing significant alignments with SJ09941	**AA number**	**Topology**	**Putative localization (by HECTAR)**	**Conserved domains**	**UniProt *Ectocarpus sp.* CCAP 1310/34**	***Cladosiphon okamuranus***	***Nemacystus decipiens***	***Saccharina japonica***	**NCBI *Aureococcus anophagefferens***
Ec-20_004410.8 Tyrosinase copper-binding domain (802) ;mRNA;... 187 1×10^{-47}	801	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (89%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (70%)	SJ10707 (62%)	XP_009037689.1 (34%)
Ec-20_004410.9 Tyrosinase copper-binding domain (802) ;mRNA;... 187 1×10^{-47}	801	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JC11 (89%)	Cok_S_s173_13265.t1 (66%)	g13002.t1 (70%)	SJ10707 (62%)	XP_009037689.1 (34%)
Ec-20_004420.1 Uncharacterised domain, di-copper centre (771... 187) 2×10^{-47}	770	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5J7R8 (94%)	Cok_S_s173_13266.t1 (70%)	g13001.t1 (68%)	SJ10707 (55%)	XP_009037689.1 (34%)
Ec-20_004450.1 Tyrosinase copper-binding domain (772) ;mRNA;... 181 9×10^{-46}	771	soluble protein (no transmembrane)	Cytoplasm (no signal peptide)	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5J676 (95%)	Cok_S_s027_6073.t1 (67%)	g10700.t1 (68%)	SJ10707 (59%)	XP_009037689.1 (33%)
Ec-27_001330.1 Tyrosinase copper-binding domain (746) ;mRNA;... 201 1×10^{-51}	745	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JRR4 (94%)	Cok_S_s485_18441.t1 (63%)	g6594.t1 (68%)	SJ10707 (51%)	XP_009037689.1 (32%)

Additional o-aminophenol oxidases from *E. siliculosus* and their stramenopile homologs (continued on next page)

					Best homologs in stramenopile species (Amino Acid identity %)				
Sequences producing significant alignments with SJ09941	**AA number**	**Topology**	**Putative localization (by HECTAR)**	**Conserved domains**	**UniProt *Ectocarpus sp.* CCAP 1310/34**	***Cladosiphon okamuranus***	***Nemacystus decipiens***	***Saccharina japonica***	**NCBI *Aureococcus anophagefferens***
Ec-27_001330.2 Tyrosinase copper-binding domain (760) ;mRNA;... 201 1×10^{-51}	759	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JRR4 (94%)	Cok_S_s485_18441.t1 (63%)	g6594.t1 (68%)	SJ10707 (51%)	XP_009037689.1 (32%)
Ec-27_005590.1 Tyrosinase copper-binding domain (730) ;mRNA;... 167 2×10^{-41}	729	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5K485 (68%)	Cok_S_s485_18441.t1 (66%)	g7402.t2 (70%)	SJ10707 (55%)	XP_009037689.1 (34%)
Ec-27_005920.1 putative scytonemin-related tyrosinase (814) ... 655 0.0	813	Cyto domain transmembrane extracellular domain	Plasma membrane	IPR002227 Tyrosinase copper-binding domain IPR008922 Uncharacterised domain, di-copper centre	A0A6H5JTC5 (93%) A0A6H5K0F0 (92%)	Cok_S_s018_5025.t1 (64%)	g1320.t1 (64%)	SJ08822 (48%)	None

Table S10: **Additional o-aminophenol oxidases from *E. siliculosus* and their homologs in other stramenopiles**

Reactions	Relative presence (species with reaction/total species in group)				Cryptophyte network resembles (1=yes, 0=no)		
	Crypto-phytes	Hapto-phytes	Strame-nopiles	Arche-plastids	Hapto-phytes	Stra-nopiles	Arche-plastids
RXN-1827	1.00	0.00	0.10	1.00	0	0	1
RXN-12278	1.00	0.00	0.10	1.00	0	0	1
RXN-12384	1.00	0.00	0.10	1.00	0	0	1
RXN-12279	1.00	0.00	0.10	1.00	0	0	1
RXN-15909	1.00	0.00	0.10	1.00	0	0	1
GLYCOPHOSPHORYL-RXN	1.00	0.00	0.00	1.00	0	0	1
RXN-9025	1.00	0.00	0.00	1.00	0	0	1
RXN-15292	1.00	0.00	0.00	0.95	0	0	1
RXN-15294	1.00	0.00	0.00	0.95	0	0	1
RXN-12203	1.00	0.00	0.00	0.95	0	0	1
RXN-15293	1.00	0.00	0.00	0.95	0	0	1
2.7.9.4-RXN	1.00	0.00	0.00	0.95	0	0	1
RXN-12201	1.00	0.00	0.00	0.95	0	0	1
RXN-20394	1.00	0.33	0.00	0.86	0	0	1
RXN3O-4042	1.00	0.33	0.00	0.86	0	0	1
RXN-7796	1.00	0.00	0.10	0.86	0	0	1
2.7.9.5-RXN	1.00	0.00	0.00	0.86	0	0	1
SUCROSE-PHOSPHATASE-RXN	1.00	0.00	0.00	0.86	0	0	1
RXN-12204	1.00	0.00	0.00	0.86	0	0	1
RXN-12202	1.00	0.00	0.00	0.86	0	0	1
CARNOSINE-SYNTHASE-RXN	1.00	0.00	0.20	0.82	0	0	1
RXN-11222	1.00	0.00	0.20	0.82	0	0	1
RXN-11858	1.00	0.00	0.10	0.82	0	0	1
RXN-11857	1.00	0.00	0.10	0.82	0	0	1
RXN-11859	1.00	0.00	0.10	0.82	0	0	1
SERINE-GLYOXYLATE-AMINOTRANS-FERASE-RXN	1.00	0.00	0.10	0.82	0	0	1
SERINE–GLYOXYLATE-AMINOTRANS-FERASE-RXN	1.00	0.00	0.10	0.82	0	0	1
RXN-17849	1.00	0.00	1.00	0.18	0	1	0
PPENTOMUT-RXN	1.00	0.00	0.90	0.50	0	1	0
D-PPENTOMUT-RXN	1.00	0.00	0.90	0.50	0	1	0
RXN-9679	1.00	0.00	0.90	0.32	0	1	0
RXN-13405	1.00	0.00	0.90	0.32	0	1	0
RXN-9680	1.00	0.00	0.90	0.32	0	1	0
RXN-10057	1.00	0.00	0.90	0.18	0	1	0
RXN-14171	1.00	0.00	0.90	0.18	0	1	0
CITRYLY-RXN	1.00	0.00	0.90	0.14	0	1	0
RXN0-6731	1.00	0.00	0.90	0.05	0	1	0

Reactions distinguishing the cryptophyte, haptophyte, stramenopile, and archeplastida groups (continued on next page)

Reactions	Relative presence (species with reaction/total species in group)				Cryptophyte network resembles (1=yes, 0=no)		
	Crypto-phytes	Hapto-phytes	Strame-nopiles	Arche-plastids	Hapto-phytes	Strame-nopiles	Arche-plastids
ALKYLGLYCERONE-PHOSPHATE-SYN-THASE-RXN	1.00	0.00	0.90	0.00	0	1	0
RXN-17130	1.00	0.00	1.00	1.00	0	1	1
PANTOATE-BETA-ALANINE-LIG-RXN	1.00	0.00	1.00	1.00	0	1	1
RXN-12487	1.00	0.00	1.00	1.00	0	1	1
L-AMINO-ACID-OXIDASE-RXN	1.00	0.00	1.00	1.00	0	1	1
ETHANOLAMINE-KINASE-RXN	1.00	0.00	1.00	0.95	0	1	1
RXN-11485	1.00	0.00	1.00	0.91	0	1	1
TRANS-PENTAPRENYLTRANS-FERASE-RXN	1.00	0.00	1.00	0.91	0	1	1
RXN-9106	1.00	0.00	1.00	0.91	0	1	1
DTDPDEHYDRHAMEPIM-RXN	1.00	0.00	1.00	0.82	0	1	1
NAPHTHOATE-SYN-RXN	1.00	0.00	1.00	0.77	0	1	1
GLYCINE-AMINOTRANSFERASE-RXN	1.00	0.00	1.00	0.77	0	1	1
RXN-12070	1.00	0.00	1.00	0.73	0	1	1
2.7.7.44-RXN	1.00	0.00	1.00	0.68	0	1	1
2.7.7.11-RXN	1.00	0.00	1.00	0.68	0	1	1
H2NEOPTERINP3PYROPHOSPHOHY-DRO-RXN	1.00	0.00	1.00	0.64	0	1	1
RXN-8141	1.00	0.00	0.90	0.95	0	1	1
CARDIOLIPSYN-RXN	1.00	0.00	0.90	0.95	0	1	1
CARBOXYCYCLOHEXADIENYL-DEHY-DRATASE-RXN	1.00	0.00	0.60	0.95	0	1	1
RXN-12322	1.00	0.00	0.60	0.95	0	1	1
RXN-13072	1.00	0.00	0.90	0.91	0	1	1
OMEGA-AMIDASE-RXN	1.00	0.00	0.90	0.91	0	1	1
RXN-20553	1.00	0.00	0.90	0.82	0	1	1
RXN-20550	1.00	0.00	0.90	0.82	0	1	1
RXN-20515	1.00	0.00	0.90	0.82	0	1	1
RXN-20518	1.00	0.00	0.90	0.82	0	1	1
RXN-12116	1.00	0.00	0.90	0.82	0	1	1
RXN-20516	1.00	0.00	0.90	0.82	0	1	1
RXN-20517	1.00	0.00	0.90	0.82	0	1	1
RXN-20502	1.00	0.00	0.90	0.82	0	1	1
RXN-17879	1.00	0.00	0.90	0.55	0	1	1
RXN-19954	1.00	0.00	0.70	0.86	0	1	1
RXN0-5468	1.00	0.00	0.70	0.86	0	1	1
RXN-19953	1.00	0.00	0.70	0.86	0	1	1
1.11.1.15-RXN	1.00	0.00	0.70	0.86	0	1	1

Reactions distinguishing the cryptophyte, haptophyte, stramenopile, and archeplastida groups (continued on next page)

Reactions	Relative presence (species with reaction/total species in group)				Cryptophyte network resembles (1=yes, 0=no)		
	Crypto-phytes	Hapto-phytes	Strame-nopiles	Arche-plastids	Hapto-phytes	Strame-nopiles	Arche-plastids
3.1.1.73-RXN	1.00	0.00	0.60	0.86	0	1	1
RXN-1321	1.00	0.00	0.60	0.82	0	1	1
2.4.1.123-RXN	1.00	1.00	0.0	0.45	1	0	0
2.4.2.12-RXN	1.00	1.00	0.00	0.36	1	0	0
RXN-16314	1.00	1.00	0.00	0.32	1	0	0
RXN-16313	1.00	1.00	0.00	0.32	1	0	0
2.4.1.228-RXN	1.00	1.00	0.00	0.32	1	0	0
RXN-16302	1.00	1.00	0.00	0.32	1	0	0
3.2.1.49-RXN	1.00	1.00	0.00	0.27	1	0	0
RXN66-577	1.00	1.00	0.40	0.00	1	0	0
SARCOSINE-DEHYDROGENASE-RXN	1.00	1.00	0.40	0.00	1	0	0
RXN-13680	1.00	1.00	0.40	0.00	1	0	0
ASPARTOACYLASE-RXN	1.00	1.00	0.40	0.00	1	0	0
ENDOGLYCOSYLCERAMIDASE-RXN	1.00	1.00	0.00	0.09	1	0	0
GLYCOGENSYN-RXN	1.00	1.00	0.00	1.00	1	0	1
RXN-18777	1.00	1.00	0.00	1.00	1	0	1
RXN-14371	1.00	1.00	0.00	1.00	1	0	1
GLYCOGEN-BRANCH-RXN	1.00	1.00	0.00	1.00	1	0	1
RXN-14372	1.00	1.00	0.00	1.00	1	0	1
RXN-7710	1.00	1.00	0.00	1.00	1	0	1
RXN-7669	1.00	1.00	0.00	1.00	1	0	1
3.4.23.25-RXN	1.00	1.00	0.00	0.73	1	0	1
CPM-KDOSYNTH-RXN	1.00	1.00	0.00	0.55	1	0	1
1.5.1.15-RXN	1.00	1.00	1.00	0.00	1	1	0
ALLANTOICASE-RXN	1.00	1.00	1.00	0.00	1	1	0
TRIMETHYLLYSINE-DIOXYGENASE-RXN	1.00	1.00	0.90	0.00	1	1	0
RXN-19672	1.00	1.00	0.60	0.00	1	1	0
RXN-18370	1.00	0.67	0.90	0.00	1	1	0
CARNITINE-O-ACETYLTRANSFERASE-RXN	1.00	0.67	0.90	0.00	1	1	0
RXN-18368	1.00	0.67	0.90	0.00	1	1	0
RXN-18371	1.00	0.67	0.90	0.00	1	1	0
CARNITINE-O-PALMITOYLTRANS-FERASE-RXN	1.00	0.67	0.90	0.00	1	1	0
RXN-18367	1.00	0.67	0.90	0.00	1	1	0
RXN-18372	1.00	0.67	0.90	0.00	1	1	0
RXN-9918	1.00	0.67	0.90	0.00	1	1	0
RXN-18373	1.00	0.67	0.90	0.00	1	1	0
RXN-10960	0.00	1.00	1.00	0.00	0	0	1
RXN-10959	0.00	1.00	1.00	0.00	0	0	1

Reactions distinguishing the cryptophyte, haptophyte, stramenopile, and archeplastida groups (continued on next page)

Reactions	Relative presence (species with reaction/total species in group)				Cryptophyte network resembles (1=yes, 0=no)		
	Crypto-phytes	Hapto-phytes	Strame-nopiles	Arche-plastids	Hapto-phytes	Strame-nopiles	Arche-plastids
3.1.3.66-RXN	0.00	1.00	1.00	0.00	0	0	1
2.7.11.15-RXN	0.00	1.00	0.90	0.00	0	0	1
2.7.11.14-RXN	0.00	0.67	0.90	0.00	0	0	1
2.7.11.16-RXN	0.00	0.67	0.90	0.00	0	0	1
RXN-12488	0.00	1.00	0.00	0.86	0	1	0
RXN-15011	0.00	1.00	0.00	0.73	0	1	0
GLUCURONOKINASE-RXN	0.00	1.00	0.00	0.73	0	1	0
PEPTIDYLGLYCINE-MONOOXY-GENASE-RXN	0.00	1.00	0.00	0.59	0	1	0
PEPTIDYLAMIDOGLYCOLATE-LYASE-RXN	0.00	1.00	0.00	0.59	0	1	0
RXN-16480	0.00	0.67	0.00	0.91	0	1	0
RXN-10856	0.00	0.67	0.00	0.91	0	1	0
RXN-13426	0.00	1.00	0.00	0.18	0	1	1
RXN-17112	0.00	1.00	0.00	0.18	0	1	1
RXN-16132	0.00	1.00	0.00	0.18	0	1	1
RXN-16065	0.00	1.00	0.00	0.18	0	1	1
1.14.19.3-RXN	0.00	1.00	0.00	0.18	0	1	1
RXN-16248	0.00	1.00	0.20	0.00	0	1	1
RXN-13788	0.00	1.00	0.20	0.00	0	1	1
RXN-16249	0.00	1.00	0.20	0.00	0	1	1
RXN-14574	0.00	1.00	0.20	0.00	0	1	1
3.2.1.40-RXN	0.00	1.00	0.20	0.00	0	1	1
QUERCITRINASE-RXN	0.00	1.00	0.20	0.00	0	1	1
RXN-16250	0.00	1.00	0.20	0.00	0	1	1
RXN-15289	0.00	1.00	0.10	0.00	0	1	1
BETA-GLUCURONID-RXN	0.00	1.00	0.10	0.00	0	1	1
RXN-15291	0.00	1.00	0.10	0.00	0	1	1
RXN-9990	0.00	1.00	0.00	0.05	0	1	1
RXN-17890	0.00	0.00	1.00	0.95	1	0	0
RXN-17888	0.00	0.00	1.00	0.95	1	0	0
RXN-17889	0.00	0.00	1.00	0.95	1	0	0
ARGINYLTRANSFERASE-RXN	0.00	0.00	1.00	0.95	1	0	0
RXN-17891	0.00	0.00	1.00	0.95	1	0	0
RXN-9279	0.00	0.00	1.00	0.86	1	0	0
R82-RXN	0.00	0.00	1.00	0.82	1	0	0
3.1.3.77-RXN	0.00	0.00	1.00	0.77	1	0	0
R83-RXN	0.00	0.00	1.00	0.77	1	0	0
R147-RXN	0.00	0.00	1.00	0.77	1	0	0
R145-RXN	0.00	0.00	1.00	0.73	1	0	0

Reactions distinguishing the cryptophyte, haptophyte, stramenopile, and archeplastida groups (continued on next page)

Reactions	Relative presence (species with reaction/total species in group)				Cryptophyte network resembles (1=yes, 0=no)		
	Crypto-phytes	Hapto-phytes	Strame-nopiles	Arche-plastids	Hapto-phytes	Strame-nopiles	Arche-plastids
4.99.1.3-RXN	0.00	0.00	1.00	0.73	1	0	0
RXN-8759	0.00	0.00	1.00	0.73	1	0	0
L-ASPARTATE-OXID-RXN	0.00	0.00	1.00	0.68	1	0	0
RXN-9772	0.00	0.00	1.00	0.68	1	0	0
PREPHENATE-ASP-TRANSAMINE-RXN	0.00	0.00	1.00	0.64	1	0	0
L-LACTATE-DEHYDROGENASE-CYTO-CHROME-RXN	0.00	0.00	1.00	0.64	1	0	0
PREPHENATE-TRANSAMINE-RXN	0.00	0.00	1.00	0.64	1	0	0
SQUALENE-MONOOXYGENASE-RXN	0.00	0.00	0.70	0.95	1	0	0
QUERCETIN-23-DIOXYGENASE-RXN	0.00	0.00	0.50	0.91	1	0	0
RXN-12730	0.00	0.00	0.90	0.77	1	0	0
PYRIMSYN1-RXN	0.00	0.00	0.90	0.77	1	0	0
QUINOLINATE-SYNTHE-MULTI-RXN	0.00	0.00	0.90	0.73	1	0	0
R146-RXN	0.00	0.00	0.90	0.68	1	0	0
GLYOXYLATE-OXIDASE-RXN	0.00	0.00	0.60	0.86	1	0	0
RXN-13722	0.00	0.00	0.50	0.86	1	0	0
HOMOACONITATE-HYDRATASE-RXN	0.00	0.00	0.50	0.86	1	0	0
1.6.5.4-RXN	0.00	0.00	0.60	0.82	1	0	0
RXN-12503	0.00	0.00	0.50	0.82	1	0	0
RXN-12504	0.00	0.00	0.50	0.82	1	0	0
RXN-19968	0.00	0.00	0.50	0.82	1	0	0
RXN-12502	0.00	0.00	0.50	0.82	1	0	0
RXN-11860	0.00	0.00	0.50	0.82	1	0	0
RXN-3522	0.00	0.00	0.50	0.82	1	0	0
RXN-11865	0.00	0.00	0.50	0.82	1	0	0
1.1.1.220-RXN	0.00	0.00	1.00	0.18	1	0	1
SEPIAPTERIN-REDUCTASE-RXN	0.00	0.00	1.00	0.18	1	0	1
RXN-8853	0.00	0.00	1.00	0.18	1	0	1
RXN-8854	0.00	0.00	1.00	0.18	1	0	1
RXN-13222	0.00	0.00	0.90	0.18	1	0	1
CHOLINESTERASE-RXN	0.00	0.33	0.90	0.00	1	0	1
RXN-17111	0.00	0.00	0.10	1.00	1	1	0
RXN-12997	0.00	0.00	0.10	1.00	1	1	0
RXN-13306	0.00	0.00	0.10	1.00	1	1	0
RXN-20335	0.00	0.00	0.10	1.00	1	1	0
RXN-16114	0.00	0.00	0.10	1.00	1	1	0
RXN-13309	0.00	0.00	0.10	1.00	1	1	0
RXN-14495	0.00	0.00	0.10	1.00	1	1	0
RXN-7711	0.00	0.00	0.10	1.00	1	1	0

Reactions distinguishing the cryptophyte, haptophyte, stramenopile, and archeplastida groups (continued on next page)

Reactions	Relative presence (species with reaction/total species in group)				Cryptophyte network resembles (1=yes, 0=no)		
	Crypto-phytes	Hapto-phytes	Strame-nopiles	Arche-plastids	Hapto-phytes	Strame-nopiles	Arche-plastids
RXN-12971	0.00	0.00	0.10	1.00	1	1	0
RXN-20339	0.00	0.00	0.10	1.00	1	1	0
RXN-13445	0.00	0.00	0.10	1.00	1	1	0
RXN-20347	0.00	0.00	0.10	1.00	1	1	0
RXN-20343	0.00	0.00	0.10	1.00	1	1	0
RXN-13308	0.00	0.00	0.10	1.00	1	1	0
RXN-20351	0.00	0.00	0.10	1.00	1	1	0
RXN-16022	0.00	0.00	0.10	1.00	1	1	0
RXN-13307	0.00	0.00	0.10	1.00	1	1	0
RXN-16131	0.00	0.00	0.10	1.00	1	1	0
RXN-16097	0.00	0.00	0.10	1.00	1	1	0
RXN-16156	0.00	0.00	0.10	1.00	1	1	0
RXN-14486	0.00	0.00	0.10	1.00	1	1	0
1.2.1.13-RXN	0.00	0.00	0.00	1.00	1	1	0
RXN-4301	0.00	0.00	0.00	1.00	1	1	0
RXN-12280	0.00	0.00	0.00	1.00	1	1	0
3.2.1.68-RXN	0.00	0.00	0.00	1.00	1	1	0
RXN-14380	0.00	0.00	0.00	1.00	1	1	0
RXN-5283	0.00	0.00	0.30	0.91	1	1	0
RXN-5284	0.00	0.00	0.30	0.91	1	1	0
RXN-13191	0.00	0.00	0.30	0.91	1	1	0
RXN-5282	0.00	0.00	0.30	0.91	1	1	0
ARALKYLAMINE-N-ACETYL-TRANS-FERASE-RXN	0.00	0.00	0.00	0.91	1	1	0
RXN3DJ-35528	0.00	0.00	0.00	0.91	1	1	0
3.1.4.2-RXN	0.00	0.00	0.00	0.91	1	1	0
UDP-NACMURALGLDAPLIG-RXN	0.00	0.00	0.20	0.82	1	1	0
RXN-19230	0.00	0.33	0.00	0.82	1	1	0
ISOCITRATE-DEHYDROGENASE-NAD+-RXN	0.00	0.33	0.00 0.00	0.82	1	1	0
URUR-RXN	0.00	0.00	0.00	0.82	1	1	0

Table S11: **Reactions distinguishing the cryptophyte, haptophyte, stramenopile, and archeplastida groups.**

Pathways >= 50% complete, represented by reactions shared between cryptophytes and stramenopiles

Pathway ID	Reactions in pathway	Reactions shared	Completeness of pathway	
HEXPPSYN-PWY	TRANS-PENTAPRENYLTRANS-FERASE-RXN	TRANS-PENTAPRENYLTRANS-FERASE-RXN	1	Quinone synthesies
PWY-5806	RXN-9106	RXN-9106	1	all-trans-decaprenyl diphosphate biosynthesis
PWY-6111	CARNITINE-O-PALMI-TOYLTRANSFERASE-RXN; TRANS-RXN-177;RXN-9918	CARNITINE-O-PALMI-TOYLTRANSFERASE-RXN; RXN-9918	0.67	mitochondrial L-carnitine shuttle
P541-PWY	GLYCINE-N-METHYL-TRANSFERASE-RXN;RXN-9680;RXN-9679	RXN-9680;RXN-9679	0.67	glycine betaine biosynthesis IV (from glycine)
PWY-6004	RXN-9680;GLYCINE-N-METHYLTRANSFERASE-RXN;RXN-9679	RXN-9680;RXN-9679	0.67	glycine betaine biosynthesis V
PWY-5697	ALLANTOINASE-RXN; ALLANTOICASE-RXN	ALLANTOICASE-RXN	0.5	allantoin degradation to ureidoglycolate I (urea producing)
PWY-6038	CITRYLY-RXN; CITTRANS-RXN	CITRYLY-RXN	0.5	citrate degradation

Pathways >= 50% complete, represented by reactions shared between cryptophytes and archeplastids

Pathway ID	Reactions in pathway	Reactions shared	Completeness of pathway	
GLYCOGEN-SYNTH-PWY	PHOSPHOGLUCMUT-RXN;GLUC1PADENYL-TRANS-RXN;GLYCOGEN-BRANCH-RXN;GLYCOGENSYN-RXN	GLYCOGEN-BRANCH-RXN;GLYCOGENSYN-RXN	0.5	glycogen biosynthesis
PWY-6724	RXN-12276;RXN-12203; RXN-12391;RXN-12384; RXN-12278;RXN-12280; RXN-12279;RXN-12277; RXN-12204	RXN-12203;RXN-12204; RXN-12279;RXN-12384; RXN-12278	0.56	starch degradation II
HEXPPSYN-PWY	TRANS-PENTAPRENYLTRANS-FERASE-RXN	TRANS-PENTAPRENYLTRANS-FERASE-RXN	1	Quinone synthesies

Shared pathways and the absence of pathways between several algal families (continued on next page)

Pathways >= 50% complete, represented by reactions shared between cryptophytes and archeplastids

Pathway ID	Reactions in pathway	Reactions shared	Completeness of pathway	
PWY66-420	CARNOSINE-SYN-THASE-RXN	CARNOSINE-SYN-THASE-RXN	1	carnosine biosynthesis
PWY-5806	RXN-9106	RXN-9106	1	all-trans-decaprenyl diphosphate biosynthesis
PWY66-421	RXN-11222	RXN-11222	1	homocarnosine biosynthesis

Pathways >= 50% complete, represented by reactions absent between cryptophytes and archeplastids

Pathway ID	Reactions in pathway	Reactions absent	Completeness of pathway	
PWY-7445	RXN-15291;RXN-15290;RXN-15289;RXN-15288	RXN-15291;RXN-15289	0.5	luteolin triglucuronide degradation
PWY-6000	1.14.19.3-RXN;RXN-9673	1.14.19.3-RXN	0.5	γ-linolenate biosynthesis II
PWY-5664	4.2.3.12-RXN;1.1.1.220-RXN;GTP-CYCLOHYDRO-I-RXN;RXN-8854	1.1.1.220-RXN;RXN-8854	0.5	erythro-tetrahydrobiopterin biosynthesis

Pathways >= 50% complete, represented by reactions absent between cryptophytes and haptophytes

Pathway ID	Reactions in pathway	Reactions absent	Completeness of pathway	
PWY-5389	R146-RXN	R146-RXN	1	3-methylthiopropanoate biosynthesis
PWY-4361	RXN-13072;RXN-15650; R147-RXN;5.3.1.23-RXN; R83-RXN;R145-RXN; R82-RXN	R147-RXN;R83-RXN; R145-RXN;R82-RXN	0.57	S-methyl-5-thio-α-D-ribose 1-phosphate degradation I
PWY-5670	SQUALENE-MONOOXYGENASE-RXN; RXN-13162	SQUALENE-MONOOXYGENASE-RXN	0.5	epoxysqualene biosynthesis
PWY-6890	PYRIMSYN3-RXN; PYRIMSYN1-RXN	PYRIMSYN1-RXN	0.5	4-amino-2-methyl-5-diphosphomethylpyrimidine biosynthesis I
PWY-5664	RXN-8854;GTP-CYCLOHYDRO-I-RXN; 4.2.3.12-RXN;1.1.1.220-RXN	RXN-8854;1.1.1.220-RXN	0.5	erythro-tetrahydrobiopterin biosynthesis

Shared pathways and the absence of pathways between several algal families (continued on next page)

Pathways >= 50% complete, represented by reactions shared between cryptophytes and haptophytes

Pathway ID	Reactions in pathway	Reactions shared	Completeness of pathway	
PWY-6111	CARNITINE-O-PALMITOYLTRANS-FERASE-RXN;TRANS-RXN-177;RXN-9918	CARNITINE-O-PALMITOYLTRANS-FERASE-RXN;RXN-9918	0.67	mitochondrial L-carnitine shuttle
PWY-5697	ALLANTOICASE-RXN; ALLANTOINASE-RXN	ALLANTOICASE-RXN	0.5	allantoin degradation to ureidoglycolate I (urea producing)
NAD-BIOSYN-THESIS-III	2.7.7.1-RXN;2.4.2.12-RXN	2.4.2.12-RXN	0.5	NAD biosynthesis III (from nicotinamide)
GLYCOGEN-SYNTH-PWY	PHOSPHOGLUCMUT-RXN;GLYCOGEN-BRANCH-RXN; GLUC1PADENYLTRANS-RXN;GLYCOGENSYN-RXN;	GLYCOGEN-SYN-RXN; GLYCOGEN-BRANCH-RXN	0.5	glycogen biosynthesis

Pathways >= 50% complete, represented by reactions absent between cryptophytes and stramenopiles

Pathway ID	Reactions in pathway	Reactions absent	Completeness of pathway	
PWY-6000	1.14.19.3-RXN;RXN-9673	1.14.19.3-RXN	0.5	γ-linolenate biosynthesis II
PWY-7445	RXN-15288;RXN-15291; RXN-15289;RXN-15290	RXN-15289;RXN-15291	0.5	luteolin triglucuronide degradation

Table S12: **Shared metabolic pathways as well as the absence of pathways between chryptophytes, haptophytes, stramenopiles, and archaeplastids.** Only pathways that consist of at least 50% group-specific reactions were considered.

3 Methods

3.1 Robustness criteria applied to a toy example

The robustness function $robust\_func^{(t)}(x) = \min(1, \frac{1}{x}\max(\lceil t\,x \rceil, \lceil \frac{5}{x} \rceil))$ was chosen such that it is 1 for low values of N_org and then decreases to a threshold value (by default $t = 0.05$) for large values of N_org. Is it plotted in Fig. S14. An example of the application of the robustness criteria is shown in Fig. S15. It contains $N\_org(r) = 18$ organisms with genes which are predicted to be associated with the reaction r. Five genes are predicted according to at least one genome annotation (green nodes, $annot\_type(r, g) = 1$) and 19 genes are predicted according to orthology relations only (blue nodes, $orth\_type(r, g) = 1$ and $annot\_type(r, g) = 0$). An edge between two nodes means that an orthology relation was detected between the two considered genes.

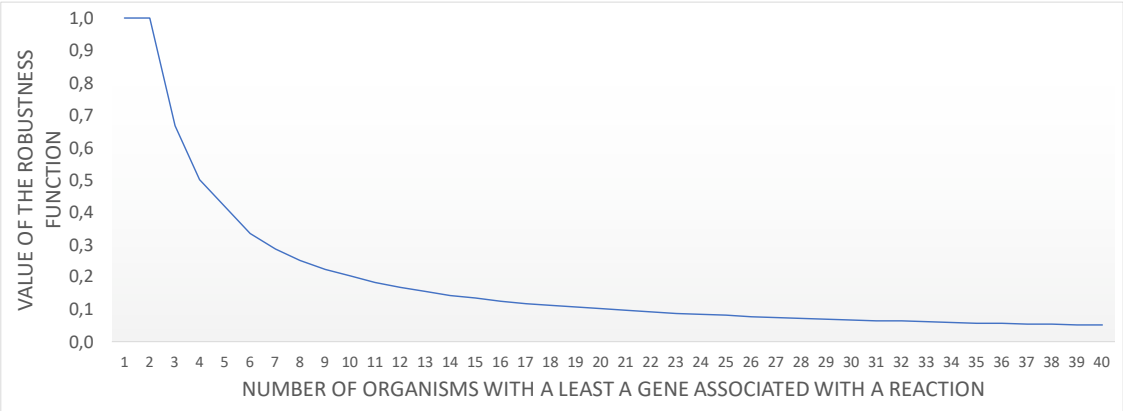

Figure S14: **Robustness function used to identify robust gene-reaction association according to the number of organisms present in the graph of genes associated with a reaction.**

According to the graph, the association between $gene1$ and r propagates to $N\_prop(r, gene1) = 5$ organisms, including two annotated ones ($gene2$ in $org2$ and $gene3$ in $org3$), and the organism $Org6$, which contains two genes ($gene6$ and $gene7$). Similarly, we note that $N\_prop(r, gene2) = 5$, $N\_prop(r, gene3) = 5$, $N\_prop(r, gene4) = 12$ and $N\_prop(r, gene5) = 2$. For genes 4 and 5, however, GPRs are not supported by an other orthology with an annotated gene.

Let g be a gene associated with r by AuCoMe. If $annot\_type(r, g) = 1$ (e.g., g is a green node), then the GPR is considered robust because it is supported by at least one annotation, and we define $robust(r, g) = 1$. This is the case for genes 1 to 5.

If $annot\_type(r, g) = 0$, we have to consider the different scenarios described in Materials and Methods.

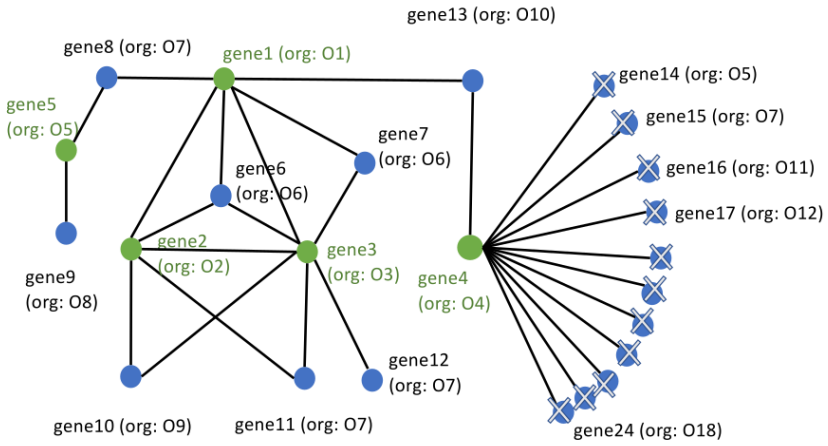

Figure S15: **Application of the robustness criteria to gene-reaction associations predicted by the draft reconstruction and the orthology propagation steps of AuCoMe.** Graph of robust genes associated with a reaction according to annotation criteria (green nodes) or orthology criteria (blue nodes).

If the gene g is part of an orthology cluster which is supported by least two annotations (for example genes 1, 2 and 3 in Fig. S15), the GPR associations are considered robust. Formally this means that there exists an orthologous gene $g2$ to $g1$ such that $annot\_type(r, g2)= 1$.

If, in addition, we know that if $orth\_type(r, g1) = 1$, then there exist a gene $g2$ orthologous to $g1$ (e.g. $orth(g1, g2) = 1$) such that $annot\_type(r, g2) = 1$, and we consider that the GPR between r and $g1$ is strongly supported by two independent assertions. This concerns the genes 1, 2 and 3. If $annot\_type(r, g) = 0$ but g is an ortholog of a strongly supported gene $g1$ (defined as $annot\_type(r, g1) = 1$ and $orth\_type(r, g1) = 1$), then we define $robust(r, g) = 1$. This correspond to all blue nodes which are connected to a green node that is itself connected to another green node. This is the case for genes 6, 7, 8, 10, 11, 12, and 13, which are orthologs of the strongly supported annotations of genes 1, 2 and 3.

In the second scenario the reaction is propagated from an isolated GPR (for example gene 4 in Fig. S15). We are in the case when $annot\_type(r, g) = 0$, $orth\_type(r, g) = 1$ and all genes orthologous to g supported by annotation are not orthologs of the other genes supported by annotation: this corresponds to all blue nodes which are themselves connected to green nodes. For example, gene 4 is orthologous to 12 genes, which are only predicted by orthology, and belong to 12 organisms among the 18 whose metabolism is predicted to contain the reaction r. As $N\_prop(r, gene4) = 12 > \lceil robust\_func(18) \times 18 \rceil = 2$, we consider that there is not enough evidence to set up a GPR association. Therefore, the associations between r and genes 14 to 24 are not conserved, on the contrary to the association between gene 13 and r which is supported by

additional evidences related to gene 1. By applying the same principle, the gene association between gene 9 and r is conserved as $N\_prop(r, gene5) = 2$.

3.2 Comparison to EcoCyc

We considered the reactions both present in a GSMN produced by AuCoMe and in the EcoCyc database as *True Positives (TPs). False Positives (FPs)* are reactions that are present in the GSMN produced by AuCoMe but not present in the EcoCyc metabolic network, and *False Negatives (FN)* are reactions present in the EcoCyc metabolic network but not present in the GSMN produced by AuCoMe. There were no True Negative reactions because each considered reaction either belongs to GSMNs produced by AuCoMe or the EcoCyc metabolic network.

The F-measure of each AuCoMe dataset was defined as $F = \frac{2PR}{P+R}$, where $P = \frac{TP}{TP+FP}$ is the precision (number of reactions inferred by AuCoMe and present in EcoCyc among all the reactions predicted by AuCoMe) and $R = \frac{TP}{TP+FN}$ is the recall (number of reactions inferred by AuCoMe and present in EcoCyc among all reactions in EcoCyc).

A maximal theoretical value of 0.79 was computed for the F-measure by considering as false negatives 1,019 reactions of EcoCyc that could not be found in the panmetabolism of bacterial dataset 0. This value was obtained by considering as false negatives 1,019 reactions from EcoCyc that were not found in the panmetabolism of all the strains in the bacterial dataset 0 as they could not be retrieved by AuCoMe. A first analysis was performed on the non-degraded dataset 0, and after running the AuCoMe process, the F-measure between the GSMN of the *E. coli* K–12 MG1655 genome in dataset 0 and EcoCyc was 0.67. Therefore we consider 0.67 as the reference F-measure for this experiment.

3.3 Degradation of the *E. coli K12-MG1655* genome

To create the 32 bacterial datasets used in the "Validation of the orthology propagation and structural verification steps" subsection, degradation of the genomes of the bacterial dataset were performed. The non-degraded genome (Run_00) and its corresponding metabolic network was used as input.

For each of the degraded genomes, we selected a ratio indicating the ratio of genes impacted by the structural degradation (*structural_annotation_degradation_ratio*, column *Ratiostructuraldegradation* in Table S4). And also another ratio indicating the ratio of genes impacted by the functional degradation

($functional\_annotation\_degradation\_ratio$, column $Ratiofunctionaldegradation$ in Table S4).

Then, the ratio of genes degraded by both degradation was computed:

$degradation\_ratio\_limit = functional\_annotation\_degradation\_ratio + structural\_annotation\_degradation\_ratio$

The degradation was then performed using the Algorithm 1. For each degraded run, the degradation ratios were used to identify the genes impacted. For each gene of the 2267 genes associated with reaction in the dataset Run_00 ($K12MG1655\_GPR$), a random number between 0 and 1 was selected and compared to the degradation ratio.

If the random number was inferior to the *structural_annotation_degradation_ratio*, then the gene was completely removed from the genome. If the random number was superior to this ratio but inferior to the *degradation_ratio_limit* then the functional annotation of the gene (GO Terms, EC numbers, gene name, etc) was removed. Finally, if the random number was superior to the *degradation_ratio_limit*, then the gene was not degraded.

Algorithm 1 Degradation of K12MG1655 genome

for *gene* in $K12MG1655\_GPR$ **do**
  $random\_number = random\_number\_generator()$ ▷ Find a random number between 0 and 1.
  if $random\_number < structural\_annotation\_degradation\_ratio$ **then**
    $remove\_structural\_annotation(gene)$ ▷ Completely remove the gene (location, annotation) from the genome.

  else if $structural\_annotation\_degradation\_ratio \leqslant random\_number < degradation\_ratio\_limit$ **then**
    $remove\_functional\_annotation(gene)$ ▷ Remove all GO terms, EC numbers, gene name, gene product associated with current gene.

  else if $random_number \geqslant degradation\_ratio\_limit$ **then** ▷ Gene not degraded.
  end if
end for

3.4 Availability of version 23.5 of the Pathway Tools software

Results were produced with Pathway Tools v23.5 `https://bioinformatics.ai.sri.com/ptools/`. For the sake of reproducibility of the results, the installation package will be sent on request gem-aureme@inria.fr to any academic user who holds a valid Pathway Tools licence (`https://biocyc.org/download-bundle.shtml`).

4 Content of the Supplemental Files archive

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

Additional file The associated archive contains analyses (all tabulated files used to create the figures and results of the paper), the datasets on which AuCoMe was run: the bacterial, fungal, and algal datasets, and the 32 synthetic datasets, which contain an *E. coli* K–12 MG1655 genome to which various degradations were applied, together with 28 other bacterial genomes. It contains a version of AuCoMe, PADMet source code, and the scripts used to run some figures. It is available at `https://zenodo.org/record/7752449#.ZBhOpiOZN-E`, and is also included in Supplemental Files S1-S14.

There are 14 zipped Supplemental Files.

- **Belcour_Supplemental_File_S1.zip** contains all AuCoMe code, and every scripts employed to generate the figures, supplemental figures and synthetical bacterial datasets.
  - **aucome_v0.5.1** this directory contains the code of AuCoMe used to run the three datasets.
  - **padmet_v5.0.1** this directory contains the code of PADMET used to run AuCoMe.
  - **scripts_analyses** this directory contains several scripts to generate the figures, supplemental figures and a script to degrade the *E. coli* K–12 MG1655 genome.
- **Belcour_Supplemental_File_S2.zip** enocpasses the metacyc_23.5.padmet file and the bacterial dataset repertory.
- **Belcour_Supplemental_File_S3.zip** gathers the 32 synthetic datasets, which contain an *E. coli* K–12 MG1655 genome to which various degradations were applied, together with 28 other bacterial genomes.
- **Belcour_Supplemental_File_S4.zip** contains a part of the fungal dataset.

- **Belcour_Supplemental_File_S5.zip** encompasses a part of the fungal dataset.
- **Belcour_Supplemental_File_S6.zip** gathers a part of the fungal dataset.
- **Belcour_Supplemental_File_S7.zip** includes a part of the fungal dataset.
- **Belcour_Supplemental_File_S8.zip** contains a part of the fungal dataset.
- **Belcour_Supplemental_File_S9.zip** encompasses a part of the algal dataset.
- **Belcour_Supplemental_File_S10.zip** gathers a part of the algal dataset.
- **Belcour_Supplemental_File_S11.zip** includes a part of the algal dataset.
- **Belcour_Supplemental_File_S12.zip** contains a part of the algal dataset.
- **Belcour_Supplemental_File_S13.zip** encompasses a part of the algal dataset.
- **Belcour_Supplemental_File_S14.zip** gathers three subdirectories: algae, bacteria, and fungi. It includes all files used to create the figures, supplemental figures, and results of the paper.

All these Supplemental Files are also available in a Zenodo archive (`https://zenodo.org/record/7752449#.ZBhOpiOZN-E`).

4.1 Content of the Belcour_Supplemental_File_S1.zip file

It encompasses three repertories.

4.1.1 Content of the aucome_v0.5.1 subdirectory

This directory contains a copy of the AuCoMe project on the GitHub site: `https://github.com/AuReMe/aucome` (downloaded the 15/11/2022). It is composed of two subdirectories and five files:

- **LICENCE** licence of the AuCoMe software.
- **README.rst** README of the AuCoMe software.
- **requirements.txt** contains the list of requires Python packages.
- **setup.cfg** contains metadata about AuCoMe package and is used with setup.py to distribute AuCoMe.

- **setup.py** contains various information relevant to the AuCoMe package including options and metadata. Then, it is used to distribute AuCoMe with PyPI. It is also used to create an entrypoint when installing it with pip.
- **recipes** this subdirectory contains two files:
  - **Dockerfile** contains instructions to run AuCoMe in a Docker environment.
  - **Singularity** contains instructions to run AuCoMe in a Singularity container.
- **aucome** this directory contains 11 Python files:
  - **__init__.py** indicates the directory as a Python module.
  - **__main__.py** contains the functions implementing the command-line interface of AuCoMe.
  - **analysis.py** contains the functions to analyse the AuCoMe results.
  - **check.py** contains the functions to check the input files.
  - **compare.py** contains the functions to compare the AuCoMe results between two distinct subgroups.
  - **orthology.py** contains the functions to propagate reaction through orthology.
  - **reconstruction.py** contains the functions to perform the reconstruction of draft GSMNs by using Pathway Tools in a parallel implementation.
  - **spontaneous.py** contains the functions to add spontaneous reactions to some GSMNs if it completes MetaCyc metabolic pathway.
  - **structural.py** contains the functions to check that no reactions are missing due to missing gene structures. A genomic search is performed for all reactions present in one organism but not in another.
  - **utils.py** contains a function to analyse the configuration file.
  - **workflow.py** contains functions to run all the steps of AuCoMe.

4.1.2 Content of the padmet_v5.0.1 subdirectory

This directory contains a copy of the PADMET project on the GitHub site: `https://github.com/AuReMe/padmet/` (downloaded the 15/11/2022). It is composed of two subdirectories and six files:

- **CHANGELOG.md** records of all notable changes made in the PADMET project.
- **docs** this repertory contains all the documentation files of PADMET package in the RST format.
- **LICENCE** licence of the PADMET package.
- **README.md** manual of the PADMET package.
- **requirements.txt** contains the list of requires Python packages.
- **setup.cfg** contains metadata about PADMET package and is used with setup.py to distribute PADMET.
- **setup.py** contains various information relevant to the PADMET package including options and metadata. Then, it is used to distribute PADMET with PyPI. It is also used to create an entrypoint when installing it with pip.
- **padmet** this repertory contains two files and two subdirectories:
  - **__init__.py** indicates the directory as a Python module.
  - **__main__.py** contains the functions implementing the command-line interface of PADMET.
  - **classes** contains 7 files.
    - **__init__.py** indicates the directory as a Python module.
    - **instantiation.py** contains a function to instantiate Padmet object.
    - **node.py** contains a class defining a Node object which is representing an element in a metabolic network (e.g: compound, reaction).
    - **padmetRef.py** contains a class defining a PadmetRef object which is representing a database of metabolic network.
    - **padmetSpec.py** contains a class defining a PadmetSpec object which is representing the metabolic network of a species/organism based on a reference database PadmetRef.
    - **policy.py** contains a class defining a Policy object that is defining the types of Relations and Nodes of a network.
    - **relation.py** contains a class defining a Relation object which is representing a link between two elements (Node) in a metabolic network.
  - **utils** contains 4 files and 3 subdirectories.

* **__init__.py** indicates the directory as a Python module.
* **gbr.py** implements a lexical analysis to handle genes relationship associated with a reaction, either a complex (with and relation between genes) or isozyme (with or relation between genes).
* **sbmlPlugin.py** contains functions to handle SBML element (ex: species or reaction), then it returns all the sections named notes in a dictionary.
* **utils.py** contains a function that checks paths of file.
* **connection** this subdirectory contains 22 files:
  - **__init__.py** indicates the directory as a Python module.
  - **biggAPI_to_padmet.py** allows to extract the BIGG database from the API to create a padmet. An Internet access is required.
  - **check_orthology_input.py** is written to check if the metabolic network and the proteome of the model organism use the same identifiers for genes (or at least more than a given cutoff), before running orthology based reconstruction.
  - **enhanced_meneco_output.py** extracts the results from Meneco gap-filling to add more information to the gap-filled reactions. Then it returns a PADMET file with more information for each reaction.
  - **extract_orthofinder.py** after running Orthofinder on n FASTA files, it reads the output file 'Orthogroups.tsv' to identify the orthologous genes. It is used by AuCoMe to extract the orthologous genes.
  - **extract_rxn_with_gene_assoc.py** from a given SBML file, it creates a SBML with only the reactions associated to a gene.
  - **gbk_to_faa.py** extracts protein sequence from a GenBank into a FASTA file with Biopython package.
  - **gene_to_targets.py** from a list of genes, it gets the products associated with the reactions linked to the genes. For example: G1 is linked to R1, R1 produces M1 and M2, this scripts outputs M1 and M2.
  - **get_metacyc_ontology.py** from the PadmetRef of MetaCyc, it extracts the MetaCyc ontology.

- **metexploreviz_export.py** converts a PADMET object representing a metabolic network into a JSON compatible with MetExplore.
- **modelSeed_to_padmet.py** from ModelSEED reactions and pathways files, it creates a PADMET file.
- **network_to_gnn.py** creates input for GNN (Graph Neural Networks) from PADMET or SBML.
- **padmet_to_asp.py** converts PADMET to Answer Set Programming.
- **padmet_to_matrix.py** creates a stoichiometry matrix from a PADMET file, in which the columns represent the reactions and rows represent metabolites.
- **padmet_to_padmet.py** allows to merge 1-n PADMET.
- **padmet_to_tsv.py** converts a PADMET representing a database (PadmetRef) and/or a PADMET representing a model (PadmetSpec) to TSV files.
- **pgdb_to_padmet.py** reads a PGDB folder (from BIOCYC/Pathway Tools) and creates a PADMET. It is used by AuCoMe to create PADMET files from PGDB in the annotation-based step.
- **sbmlGenerator.py** contains functions to generate SBML files from PADMET and TXT files using the libsbml package. It is used by AuCoMe to create SBML files at the annotation-based, orthology and final steps.
- **sbml_to_curation_form.py** extracts one or several reactions from a SBML file to the form used in AuReMe for curation.
- **sbml_to_padmet.py** converts a SBML file into a PADMET file (with or without a reference database).
- **sbml_to_sbml.py** creates a SBML file from another one. Use it to change the SBML level.
- **wikiGenerator.py** contains all necessary functions to generate wiki pages from a PADMET file and update a wiki online. It requires WikiManager module (with wikiMate, Vendor).

* **exploration** this subdirectory contains 15 files:
  - **__init__.py** indicates the directory as a Python module.

- **compare_padmet.py** compares 1-n PADMET files, and creates a folder with 4 output files (compounds.tsv, genes.tsv, pathways.tsv and reactions.tsv). It is used by AuCoMe to create these files to analyse the metabolic networks.
- **compare_sbml.py** compares 2 or 1-n SBML, then it creates two output files reactions.tsv and metabolites.tsv with the reactions/metabolites in each SBML files.
- **compare_sbml_padmet.py** compares reaction identifiers in SBML versus PADMET, then returns the number of reactions in both, and reaction identifiers not in SBML or not in PADMET.
- **convert_sbml_db.py** uses the MetaNetX database to check or convert a SBML. Flat files from MetaNetx are required to run this script. They can be found in the AuReMe workflow or from the MetaNetx website.
- **dendrogram_reactions_distance.py** uses the reactions.tsv file from compare_padmet.py to create a dendrogram using the R package pvclust. It has been used in the article to create the metabolic dendrogram.
- **flux_analysis.py** runs the flux balance analyse with cobra package on an already defined reaction. It needs to set in the SBML the value 'objective_coefficient' to 1.
- **get_pwy_from_rxn.py** from a file containing a list of reaction, it returns the pathways where these reactions are involved.
- **padmet_stats.py** creates a PADMET stats file (named padmet_stats.tsv) containing the number of pathways, reactions, genes and compounds inside the one or several PADMET files.
- **pathway_production.py** compares 1-n PADMET objects to show the pathway input/output for them.
- **prot2genome.py** contains function to search a genome using protein sequences and Gene-Protein-Reaction associations. It is used in the structural search step of AuCoMe.
- **report_network.py** creates reports of a PADMET file, and it writes three TSV files (all_metabolites.tsv, all_pathways.tsv, and all_reactions.tsv).
- **visu_network.py** allows to visualize a metabolic network on a compounds perspectives.
- **visu_path.py** allows to visualize a pathway in PADMET network.

- **visu_similarity_gsmn.py** visualize similarity between metabolic networks using MDS.

* **management** this subdirectory contains 5 files:
  - **__init__.py** indicates the directory as a Python module.
  - **manual_curation.py** updates a PadmetSpec object by filling specific forms. It either creates new reaction(s) to PADMET file, or it adds/removes reaction(s) from a PadmetRef.
  - **padmet_compart.py** for a given PADMET file, it checks and updates compartment.
  - **padmet_medium.py** for a given set of compounds representing the growth medium (or seeds), it creates two reactions in order to maintain consistency of the network for flux analysis.
  - **relation_curation.py** for a given PADMET file, it adds or removes relations between nodes.

4.1.3 Content of the scripts_analyses subdirectory

The scripts repertory contains 12 files:

- **bacteria_random_degradation.py** was used to degrade the *E. coli K–12 MG1655* genome. The procedure for the genome degradation is described in the algorithm 1.
- **figure_2_algal_dataset.py** for each species of the algal dataset, and at each AuCoMe step. This script allows to generate the figure 2D.
- **figure_2_bacterial_dataset.py** for each species of the bacterial dataset, and at each AuCoMe step. This script allows to generate the figure 2B.
- **figure_2_fungal_dataset.py** for each species of the fungal dataset, and at each AuCoMe step. This script allows to generate the figure 2C.
- **figure_3_degradation.py** allows to generate the figure 3B from the figure_3_fmeasure_steps.tsv file (described above).
- **figure_5_mds.py** allows to generate the figure 5A from two reactions.tsv files of the algal dataset (annotation-based and final).

- **figure_S4_comparison_bacteria.py** computes statistics on all the 29 bacterial metabolic networks reconstructed with AuCoMe, CarveMe, gapseq and ModelSEED, it uses the mapping_modelseed_ec.tsv, soft_stat.tsv files and bacteria/networks_soft directories, then it creates the Supplemental Fig. S4 and files inside the analyses/bacteria/Figure_S4_output repertory.
- **figure_S5_reference_catalog.py** reads the ecocyc.padmet, kegg_ecs.txt files, and the jsons_bigg/, jsons_modelseed/ directories, then it creates the Figure_S5_refence_ec_catalog_K12MG1655.tsv and the Supplemental Fig. S5.
- **figure_S6.py** reads the Figure_S5_refence_ec_catalog_K12MG1655.tsv file and those inside the bacteria/networks_soft/ directories about *E. coli* K–12 MG1655 , it generates the Supplemental Fig. S6.
- **figure_S7_comparison_pathway_fungi.py** reads metacyc_23.5.padmet file and those inside the fungi/networks_soft/ directories, then it create five completion_pathway_species.svg pictures which composed the Supplemental Fig. S7.
- **figures_S8_S9.py** reads the metacyc_23.5.padmet, and metabolic networks of *S. cerevisiae S288* reconstructed with AuCoMe, gapseq, and YeastCyc. For all pathways of both AuCoMe and gapseq networks, it also computes their completion rates. Then it compares the results obtained with AuCoMe and gapseq on *S. cerevisiae S288C* to YeastCyc according to the completion rates of their pathways. It generates Supplemental Fig. S8 and S9.
- **figure_S12_supervenn.py** allows to generate the figure S12, it reads the reactions.tsv file of the algal dataset at the final AuCoMe step, and another tabular file that contains abbreviated names of species.

4.2 Content of the Belcour_Supplemental_File_S2.zip file

- **metacyc_23.5.padmet** the version 23.5 of the MetaCyc database (`https://metacyc.org/`) in the PADMET format. It was used by AuCoMe to reconstruct all the metabolic networks. Hence metacyc_23.5.padmet is required to reproduce the article results.
- **bacterial** this directory gathers the bacterial dataset on which AuCoMe was run.

4.2.1 Content of the bacterial directory

It contains a file and 8 subdirectories.

- **Table_S1_description_bacterial_dataset.xlsx** is the Supplemental Table S1.
- **FASTA** contains the proteome of each species as a FASTA file.
- **cleaned_GBKs** for each species, it contains the annotated genome, with the protein sequences in a GenBank format file.
- **dictionaries** for some species, genes needed to be renamed for compatibility reasons. This folder contains CSV files with the mapping between the old names of genes and the new ones.
- **annotated_DATs** contains a subdirectory per species with all the output files from Pathway Tools v23.5, without any post-treatment, in the DAT format.
- **annotated_PADMETs** for each species, it contains a metabolic network of the draft reconstruction step of AuCoMe, in the PADMET format.
- **final_PADMETs** for each species, it contains a metabolic network generated by the AuCoMe workflow, at the PADMET format.
- **final_SBMLs** for each species, it contains a metabolic network generated by the AuCoMe workflow, in the SBML format.
- **panmetabolism** is composed of 7 files describing the final metabolic networks:
  - **genes.tsv** contains, for each organism, the list of genes and the associated reactions.
  - **metabolites.tsv** contains the list of metabolites present in the panmetabolism. Then, for each metabolite and for each organism, it lists the reactions that produced this compound and the reactions that consumed it.
  - **pathways.tsv** contains the list of pathways present in the panmetabolism. For each pathway and for each organism, it indicates the number of reactions present in this pathway, and the names of these reactions.
  - **reactions.tsv** contains the list of reactions present in the panmetabolism. Then for each reaction, it indicates whether or not it belongs to an organism. If a reaction is found in a species, the genes associated with the reaction are also listed.

– **pvclust_reaction_dendrogram.png** based on the presence/absence matrix of reactions in different species of the dataset, it computes the Jaccard distances between these species, and it applies a hierarchical clustering on these data with a complete linkage to create a dendrogram. The R package pvclust is used to create the dendrogram, with bootstrap resampling. For each node, a p-value indicates how strong the cluster is supported by data. This dendrogram is provided as a PNG picture.

4.3 Content of the Belcour_Supplemental_File_S3.zip file

This repertory includes the file **SupFile_S4_description_synthetic_bacterial_dataset.xlsx** that describes the 32 in silico bacterial datasets (Supplemental Table S4). It also contains 32 subdirectories named Run_00, Run_01, ..., etc, Run_31. Each subdirectory is composed of 9 files:

- **K_12_MG1655.gbk** the annotated genome of *E. coli* K–12 MG1655 to which degradation of the functional and/or structural annotations was applied.
- **annotated_K_12_MG1655.sbml** the metabolic network of *E. coli* K–12 MG1655 output of the draft reconstruction step of AuCoMe in the SBML format.
- **annotated_K_12_MG1655.padmet** the metabolic network of *E. coli* K–12 MG1655 output of the draft reconstruction step of AuCoMe in the PADMET format.
- **orthology_K_12_MG1655.sbml** the metabolic network of *E. coli* K–12 MG1655 output of the orthology propagation step of AuCoMe in the SBML format.
- **orthology_K_12_MG1655.padmet** the metabolic network of *E. coli* K–12 MG1655 output of the orthology propagation step of AuCoMe in the PADMET format.
- **structural_K_12_MG1655.sbml** the metabolic network of *E. coli* K–12 MG1655 output of the structural verification step of AuCoMe in the SBML format.
- **structural_K_12_MG1655.padmet** the metabolic network of *E. coli* K–12 MG1655 output of the structural verification step of AuCoMe in the PADMET format.
- **final_K_12_MG1655.sbml** the metabolic network of *E. coli* K–12 MG1655 output of the AuCoMe workflow in the SBML format.

- **final_K_12_MG1655.padmet** the metabolic network of *E. coli* K–12 MG1655 output of the AuCoMe worflow in the PADMET format.

4.4 Content of the Belcour_Supplemental_File_S4.zip file

It contains a part of the fungal dataset and AuCoMe result files from *A. clavatus* to *C. tropicalis* ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details). It also encompasses Table_S2_description_fungal_dataset.xlsx which is the Supplemental Table S2.

4.5 Content of the Belcour_Supplemental_File_S5.zip file

It gathers a part of the fungal dataset and AuCoMe result files from *C. globosum* to *E. gossypii* ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details).

4.6 Content of the Belcour_Supplemental_File_S6.zip file

It includes a part of the fungal dataset and AuCoMe result files from *F. graminearum* to *N. crassa* ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details).

4.7 Content of the Belcour_Supplemental_File_S7.zip file

It gathers a part of the fungal dataset and AuCoMe result files from *P. brasiliensis* Pb03 to *S. sclerotiorum* ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details).

4.8 Content of the Belcour_Supplemental_File_S8.zip file

It includes a part of the fungal dataset and AuCoMe result files from *S. nodorum* to *Y. lipolytica* ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details). It also encompasses the panmetabolism of all the fungal species (see section 4.2.1).

4.9 Content of the Belcour_Supplemental_File_S9.zip file

It contains a part of the algal dataset and AuCoMe result files from *A. protothecoides* to *C. merolae*, except the GBK file *C. braunii* which is too bigger, ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details). It also encompasses Table_S3_description_algal_dataset.xlsx which is the Supplemental Table S3.

4.10 Content of the Belcour_Supplemental_File_S10.zip file

It gathers a part of the algal dataset and AuCoMe result files from *C. paradoxa* to *F. cylindrus* ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details).

4.11 Content of the Belcour_Supplemental_File_S11.zip file

It includes a part of the algal dataset and AuCoMe result files from *G. sulphuraria* to *N. gaditana* ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details).

4.12 Content of the Belcour_Supplemental_File_S12.zip file

It encompasses a part of the algal dataset and AuCoMe result files from *N. decipiens* to *S. japonica* ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details).

4.13 Content of the Belcour_Supplemental_File_S13.zip file

It contains a part of the algal dataset and AuCoMe result files from *S. cerevisiae* to *V. carteri* ordered alphabetically in 7 directories: FASTA, cleaned_GBKs, dictionaries, annotated_DATs, annotated_PADMETs, final_PADMETs, and final_SBMLs (see section 4.2.1 for more details). It also encompasses the panmetabolism of all the algal species (see section 4.2.1).

4.14 Content of the Belcour_Supplemental_File_S14.zip file

It is composed of three subdirectories: algae, bacteria, and fungi.

4.14.1 Content of the algae repertory

It encompasses 9 files.

- **Figure_2_algal_nb_reactions.tsv** for each species of the algal dataset, this file gives the number of reactions at each AuCoMe step. It was used to create figure 2D.
- **Figure_S10_Deepec_algal.tsv** for each species of the algal dataset, at each AuCoMe step (robust orthology, non-robust orthology, and annotation or orthology), several measures were computed, i.e.: the number of reactions, the number of ECs, the number of ECs validated by DeepEC, and the ratio number of ECs valided by DeepEC / number of ECs. It was used to design figure S10(b).
- **Table_S6_50_random_reactions_found.xlsx** contains manual validation of 50 randomly chosen reactions found in any of the species (is the Supplemental Table S6).
- **Table_S7_50_random_reactions_absent.xlsx** includes manual validation of 50 reactions absent from a species and randomly chosen (is the Supplemental Table S7).
- **Table_S8_reactions_common_only_Cokamuranus_Sjaponica.xlsx** encompasses reactions common to *Saccharina japonica* and *Cladosiphon okamuranus* but not found in other brown algae (is the Supplemental Table S8).
- **Table_S9_homologues_Esiliculosus_Sjaponica.xlsx** contains additional homologs in *E. siliculosus* found by BLASTP searches for sequences inferred to be present only in *C. okamuranus* and *S. japonica* (is the Supplemental Table S9).
- **Table_S10_o-aminophenol_Esiliculosus_holomogues.xlsx** includes additional o-aminophenol oxidases from *E. siliculosus* and their homologs in other stramenopiles. It is the Supplemental Table S10 with more detail (like the amino acid sequences).
- **Table_S11_reactions_cryptophytes_haptophytes_stramenopiles_archeplastida.xlsx** encompasses reactions distinguishing the cryptophyte, haptophyte, stramenopile, and archeplastida groups (is the Supplemental Table S11).

- **Table_S12_pathways_cryptophytes_haptophytes_stramenopiles_archeplastida.xlsx** contains shared metabolic pathways as well as the absence of pathways between chryptophytes, haptophytes, stramenopiles, and archaeplastida (is the Supplemental Table S12).

4.14.2 Content of the bacteria directory

It encompasses 12 files and 9 repertories.

- **aucome_final.tsv** output file of the figure_S4_comparison_bacteria.py script, for each of the 29 bacterial metabolic networks produced with AuCoMe, this table contains the number of ECs, the number of unique ECs, the number of total reactions, the number of enzymatic reactions with genes, the number of enzymatic reactions without genes, and the number of spontaneous reactions.

- **carveme_stat.tsv** output file of the figure_S4_comparison_bacteria.py script, for each of the 29 bacterial metabolic networks produced with CarveMe, this table contains the number of ECs, the number of unique ECs, the number of total reactions, the number of enzymatic reactions with genes, the number of enzymatic reactions without genes, and the number of spontaneous reactions.

- **ecocyc.padmet** contains the EcoCyc database version 23.5 at the PADMet, is used to generate the Supplemental Fig. S5.

- **Figure_2_bacterial_nb_reactions.tsv** for each species of the bacterial dataset, this file gives the number of reactions at each AuCoMe step. It was used to create figure 2B.

- **Figure_3_nb_reactions_step.tsv** for each dataset of the 32 synthetic bacterial datasets, this file enumerates the number of reactions at each AuCoMe step. It was used to create figure 3A.

- **Figure_3_fmeasure_steps.tsv** for each dataset of the 32 synthetic bacterial datasets, this file indicates the values of the F-measures resulting of the comparison of the GSMNs recovered for each *E. coli* K–12 MG1655 genome replicate with the gold-standard network EcoCyc. It was used to create figure 3B.

- **Figure_S4_output** contains 3 output files of the figure_S4_comparison_bacteria.py script:

  - **Figure_S4_boxplot_networks.svg** is the Supplemental Figure S4 in high resolution.

  - **Figure_S4_boxplot_networks.tsv** contains the number of reactions, the type of reactions (All, Reactions with genes, ...), and the used software. Thes data were produced and used in the figure_S4_comparison_bacteria.py script.
  - **Figure_S4_barplot_time_networks.svg** for each software, shows the required time in seconds used to reconstruct these bacterial metabolic networks.

- **Figure_S5_output** encompasses 3 output files of figure_S5_reference_catalog.py script:
  - **Figure_S5_ec_union.svg** is the Supplemental Figure S5 in high resolution.
  - **Figure_S5_ec_union_venn.svg** another visualisation of presenting the results of the Supplemental Fig. S5.
  - **Figure_S5_refence_ec_catalog_K12MG1655.tsv** contains an EC catalog to *E. coli* K-12 MG1655 from the BIGG, EcoCyc, KEGG, and ModelSEED databases. This file is used to produce the Supplemental Figure S5.

- **Figure_S6_output** includes 2 output files of the figure_S6.py script:
  - **Figure_S6_comparison_all.svg** is the Supplemental Figure S6 in high resolution.
  - **Figure_S6_comparison_all.tsv** contains data used to produce the Supplemental Figure S6.

- **gapseq_stat.tsv** output file of the figure_S4_comparison_bacteria.py script, for each of the 29 bacterial metabolic networks produced with gapseq, this table contains the number of ECs, the number of unique ECs, the number of total reactions, the number of enzymatic reactions with genes, the number of enzymatic reactions without genes, and the number of spontaneous reactions.

- **jsons_bigg** todate contains the five metabolic networks of *E. coli* K–12 MG1655 that can find in BIGG at JSON format. These files correspond to the BIGG reference metabolic network on the Supplemental Figure S5.

- **jsons_modelseed** todate includes the metabolic network of *E. coli* K–12 MG1655 that can find in ModelSEED at JSON format. It is the ModelSEED reference metabolic network on the Supplemental Figure S5.

- **kegg_ecs.txt** input file of the figure_S5_reference_catalog.py script, it contains matches between EC numbers and all the entries of *E. coli* K–12 MG1655 in the KEGG database.

- **mapping_modelseed_ec.tsv** input file of the figure_S4_comparison_bacteria.py script, it encompasses matches between ModelSEED reactions and EC numbers.
- **modelseed_stat.tsv** output file of the figure_S4_comparison_bacteria.py script, for each of the 29 bacterial metabolic networks produced with ModelSEED, this table contains the number of ECs, the number of unique ECs, the number of total reactions, the number of enzymatic reactions with genes, the number of enzymatic reactions without genes, and the number of spontaneous reactions.
- **networks_aucome** for each of the 29 bacteria, contains a metabolic networks at the PADMet format obtained with AuCoMe.
- **networks_carveme** for each of the 29 bacteria, contains a metabolic networks at the SBML format got to CarveMe.
- **networks_gapseq** composes of 29 subdirectories (one for each bacterium). All these subdirectories contain 10 files about the metabolic networks a obtained with gapseq:
  - **species-all-Pathways.tbl** encompasses data on pathways at TBL format.
  - **species-all-Reactions.tbl** includes data on reactions at TBL format.
  - **species-draft.RDS** is a draft metabolic network at RDS (R Data Format).
  - **species-draft.xml** is a draft metabolic network at SBML format.
  - **species-medium.csv** encompasses all the metabolites allow the default medium.
  - **species.RDS** is the final metabolic network at RDS (R Data Format).
  - **species-rxnWeights.RDS** is a temporary file nedeed to gapseq fill at RDS (R Data Format).
  - **species-rxnXgenes.RDS** is a temporary file nedeed to gapseq fill at RDS (R Data Format).
  - **species-Transporter.tbl** includes data on transporters at TBL format.
  - **species.xml** is the final metabolic network at SBML format.
- **networks_modelseed** includes two subdirectories:
  - **sbml** for each of the 29 bacteria, encompasses a metabolic networks at the SBML format got to ModelSEED.
  - **tsv** for each of the 29 bacteria, contains two TSV files:

    * **genomeset__species.gbk_genome.fbamodel-compounds.tsv** includes data on compounds at TSV format.
    * **genomeset__species.gbk_genome.fbamodel-reactions.tsv** encompasses data on reactions at TSV format.

- **time_carveme.txt** input file of the figure_S4_comparison_bacteria.py script, for each of the 29 bacteria it stores the running time of CarveMe (in seconds) to reconstruct a metabolic network.
- **time_gapseq.txt** input file of the figure_S4_comparison_bacteria.py script, for each of the 29 bacteria it stores the running time of gapseq (in seconds) to reconstruct a metabolic network.

4.14.3 Content of the fungi repertory

It contains three files and five directories.

- **All-pathways-of-S.-cerevisiae-S288c.txt** encompasses all the YeastCyc pathways.
- **Figure_2_fungal_nb_reactions.tsv** for each species of the fungal dataset, this file gives the number of reactions at each AuCoMe step. It was used to create figure 2C.
- **Figure_S7_output** contains 11 output files of the figure_S7_comparison_pathway_fungi.py script:
  - **completion_pathway_species.svg** for each of the 5 fungi (*L. bicolor*, *N. crassa*, *R. oryzae*, *S. cerevisiae S288C*, and *S. pombe*), contains a subfigure of the Supplemental Fig. S7.
  - **fungi_stats.tsv** is the Supplemental Table S5.
  - **pathway_venn_species.png** for each of the 5 fungi (*L. bicolor*, *N. crassa*, *R. oryzae*, *S. cerevisiae S288C*, and *S. pombe*), includes a Venn diagram about all the pathways found with the 3 software (AuCoMe, gapseq, and ModelSEED).
- **Figures_S8_S9_output** contains 11 files, in all these files, a comparison of all pathways of metabolic networks of *S. cerevisiae S288C* obtained with AuCoMe and gapseq to those of YeastCyc was released.
  - **comparison_yeastcyc.png** is a picture about number of pathways true positive, false positive, and false negative are found, according the used method (AuCoMe and gapseq).
  - **completion_pathway_gapseq.svg** includes the number of pathways common or specific to YeastCyc and gapseq with their completeness ratio predicted by gapseq.

- **Figure_S8_completion_pathway_aucome.svg** contanis the number of pathways common or specific to YeastCyc and AuCoMe with their completeness ratio predicted by AuCoMe, is the Supplemental Figure S8.
- **Figure_S9_venn_diagram_70_100.svg** is the Supplemental Figure S9. All pathways of AuCoMe, gapseq and YeastCyc with a completion rate between 50% and 70% are compared.
- **venn_diagram.svg** in this picture, all pathways are compared.
- **venn_diagram_50.svg** all pathways of AuCoMe, gapseq and YeastCyc with a completion rate less than 50% are compared.
- **venn_diagram_50_gapseq.svg** all pathways of gapseq whatever their completion rate are compared to the AuCoMe and YeastCyc pathways with a completion rate less than 50%.
- **venn_diagram_50_70.svg** all pathways of AuCoMe, gapseq, and YeastCyc with a completion rate between 50% and 70% are compared.
- **venn_diagram_50_70_gapseq.svg** all pathways of gapseq whatever their completion rate are compared to the AuCoMe and YeastCyc pathways with a completion rate between 50% and 70%.
- **venn_diagram_70_100_gapseq.svg** all pathways of gapseq whatever their completion rate are compared to the AuCoMe and YeastCyc pathways with a completion rate between 70% and 100%.
- **yeast_cyc_comparison.tsv** contains the number of pathways true positive, false positive, and false negative are found, according the used method (AuCoMe and gapseq).

- **Figure_S10_Deepec_fungal.tsv** for each species of the fungal dataset, at each AuCoMe step (robust orthology, non-robust orthology, and annotation or orthology), several measures were computed, i.e.: the number of reactions, the number of ECs, the number of ECs valided by DeepEC, and ratio number of ECs validated by DeepEC / number of ECs. It was used to design figure S10(a).

- **networks_aucome** for each of the 5 fungi (*L. bicolor*, *N. crassa*, *R. oryzae*, *S. cerevisiae S288C*, and *S. pombe*), contains a metabolic networks at the PADMet format obtained with AuCoMe.

- **networks_gapseq** is composed of 5 subdirectories (one for each fungus). All these subdirectories contain two files about the metabolic networks a obtained with gapseq:
  - **species-all-Pathways.tbl** encompasses data on pathways at TBL format.

  - **species-all-Reactions.tbl** includes data on reactions at TBL format.

- **networks_modelseed** for each of the 5 fungi (*L. bicolor*, *N. crassa*, *R. oryzae*, *S. cerevisiae S288C*, and *S. pombe*), contains two TSV files:

  - **species.gbk_genome.draftModel-compounds.tsv** includes data on compounds at TSV format.

  - **species.gbk_genome.draftModel-reactions.tsv** encompasses data on reactions at TSV format.

References

[1] A. Bateman, M. J. Martin, S. Orchard, M. Magrane, R. Agivetova, S. Ahmad, E. Alpi, E. H. Bowler-Barnett, R. Britto, B. Bursteinas, H. Bye-A-Jee, R. Coetzee, A. Cukura, A. Da Silva, P. Denny, T. Dogan, T. Ebenezer, J. Fan, L. G. Castro, P. Garmiri, G. Georghiou, L. Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, P. Jokinen, V. Joshi, D. Jyothi, A. Lock, R. Lopez, A. Luciani, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, M. Menchi, A. Mishra, K. Moulang, A. Nightingale, C. S. Oliveira, S. Pundir, G. Qi, S. Raj, D. Rice, M. R. Lopez, R. Saidi, J. Sampson, T. Sawford, E. Speretta, E. Turner, N. Tyagi, P. Vasudev, V. Volynkin, K. Warner, X. Watkins, R. Zaru, H. Zellner, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, J. Bolleman, E. Boutet, L. Breuza, C. Casals-Casas, E. de Castro, K. C. Echioukh, E. Coudert, B. Cuche, M. Doche, D. Dornevil, A. Estreicher, M. L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, G. Keller, A. Kerhornou, V. Lara, P. Le Mercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. B. Neto, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, M. Pozzato, M. Pruess, C. Rivoire, C. Sigrist, K. Sonesson, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, J. Zhang, P. Ruch, and D. Teodoro. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, 49(D1):D480–D489, 01 2021. [PubMed Central:PMC7778908] [DOI:10.1093/nar/gkaa1100] [PubMed:29959318].

[2] D. Bhattacharya, H. S. Yoon, and J. D. Hackett. Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *Bioessays*, 26(1):50–60, Jan 2004. [DOI:10.1002/bies.10376] [PubMed:14696040].

[3] Fedor Indukaev. Supervenn python package. *Zenodo*, 2021.

[4] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36:D480–484, 2008.

[5] Peter D Karp, Peter E Midford, Richard Billington, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Wai Kit Ong, Pallavi Subhraveti, Ron Caspi, Carol Fulcher, Ingrid M Keseler, and

Suzanne M Paley. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, 12 2019. bbz104.

[6] Marie-Françoise Liaud, Ulrike Brandt, Margitta Scherzinger, and Rüdiger Cerff. Evolutionary origin of cryptomonad microalgae: Two novel chloroplast/cytosol-specific GAPDH genes as potential markers of ancestral endosymbiont and host cell components. *Journal of Molecular Evolution*, 44(1):S28–S37, 1997.

[7] N. C. Rockwell and J. C. Lagarias. Ferredoxin-dependent bilin reductases in eukaryotic algae: Ubiquity and diversity. *J Plant Physiol*, 217:57–67, Oct 2017. [PubMed Central:PMC5603387] [DOI:10.1016/j.jplph.2017.05.022] [PubMed:16380422].

[8] N. C. Rockwell and J. C. Lagarias. Phytochrome evolution in 3D: deletion, duplication, and diversification. *New Phytol*, 225(6):2283–2300, 03 2020. [PubMed Central:PMC7028483] [DOI:10.1111/nph.16240] [PubMed:26635768].

[9] N. C. Rockwell, J. C. Lagarias, and D. Bhattacharya. Primary endosymbiosis and the evolution of light and oxygen sensing in photosynthetic eukaryotes. *Front Ecol Evol*, 2(66), 2014. [PubMed Central:PMC4343542] [DOI:10.3389/fevo.2014.00066] [PubMed:12218172].

[10] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proceedings of the National Academy of Sciences*, 116(28):13996–14001, 2019.

[11] I. Schomburg, L. Jeske, M. Ulbrich, S. Placzek, A. Chang, and D. Schomburg. The BRENDA enzyme information system-From a database to an expert system. *J Biotechnol*, 261:194–206, Nov 2017. [DOI:10.1016/j.jbiotec.2017.04.020] [PubMed:28438579].