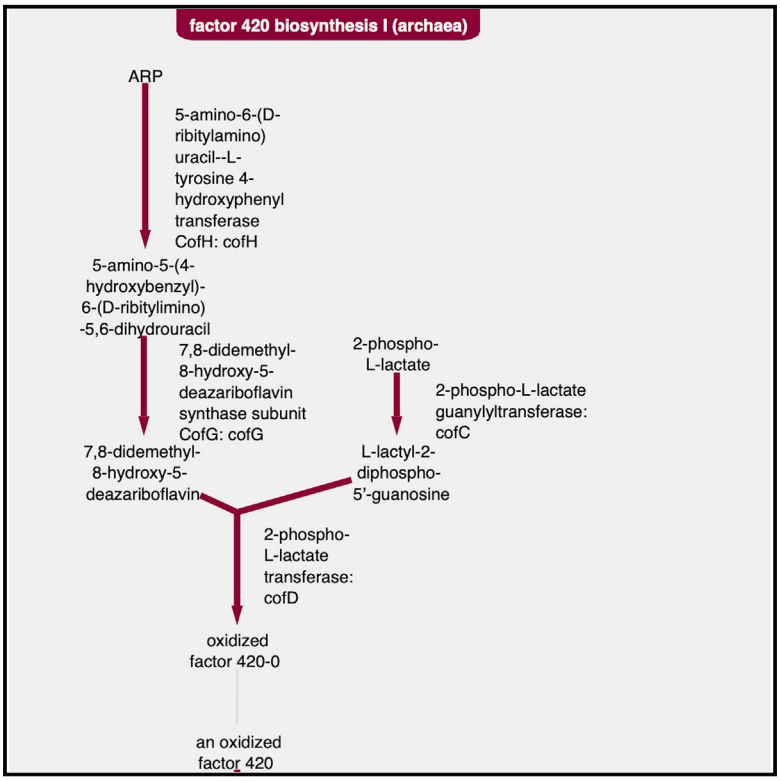
iScience

Article

MjCyc: Rediscovering the pathway-genome landscape of the first sequenced archaeon, *Methanocaldococcus (Methanococcus) jannaschii*

Graphical abstract



Highlights

- Over 600 gene products predicted with enzymatic activity in *M. jannaschii*
- About 883 enzymatic reactions inferred, involving 540 enzymes/transporters
- 104 genes had updated EC numbers with new functional descriptions

Authors

I. Baltsavia, G. Stamoulos, K. Tziavaras, ..., P.D. Karp, N.C. Kyrpides, C.A. Ouzounis

Correspondence

pkarp@ai.sri.com (P.D.K.), cao@csd.auth.gr (C.A.O.)

In brief

Gene network; Biological classification; Genomic analysis

Baltsavia et al., 2025, iScience *28*, 111546
January 17, 2025
https://doi.org/10.1016/j.isci.2024.111546

CellPress

iScience

CellPress
OPEN ACCESS

Article

MjCyc: Rediscovering the pathway-genome landscape of the first sequenced archaeon, *Methanocaldococcus (Methanococcus) jannaschii*

I. Baltsavia,[1,2] G. Stamoulos,[2,3] K. Tziavaras,[2,3] C. Bouas,[3] I. Katikaridou,[3] A. Dermaris,[3] A. Kothari,[4] I. Iliopoulos,[1] R. Caspi,[4] P.D. Karp,[4,*] N.C. Kyrpides,[5,6] and C.A. Ouzounis[2,3,5,7,*]

[1]Computational Biology Group, School of Medicine, University of Crete, Heraklion, Greece
[2]Biological Computation & Process Laboratory, Chemical Process & Energy Resources Institute, Centre for Research & Technology Hellas, Thessalonica, Greece
[3]Biological Computation & Computational Biology Group, Artificial Intelligence & Information Analysis Lab, School of Informatics, Aristotle University of Thessalonica, Thessalonica, Greece
[4]Bioinformatics Research Group, Artificial Intelligence Center, SRI International, Menlo Park, CA, USA
[5]DoE Joint Genome Institute, Berkeley, CA, USA
[6]Environmental Genomics & Systems Biology Division, Lawrence Berkeley National Lab, Berkeley, CA, USA
[7]Lead contact
*Correspondence: pkarp@ai.sri.com (P.D.K.), cao@csd.auth.gr (C.A.O.)
https://doi.org/10.1016/j.isci.2024.111546

SUMMARY

The genome of *Methanocaldococcus* (*Methanococcus*) *jannaschii* DSM 2661 was the first Archaeal genome to be sequenced in 1996. Subsequent sequence-based annotation cycles led to its first metabolic reconstruction in 2005. Leveraging new experimental results and function assignments, we have now re-annotated *M. jannaschii,* creating an updated resource with novel information and testable predictions in a pathway-genome database available at BioCyc.org. This reannotation effort has resulted in 652 function assignments with enzyme roles, accounting for a third of the total protein-coding entries for this genome. The updated resource includes 883 reactions, 540 enzymes, and 142 individual pathways. Despite notable progress in computational genomics, more than a third of the genome remains functionally uncharacterized. The publicly available MjCyc pathway-genome database holds great potential for the wider community to conduct research on the biology of methanogenic Archaea.

INTRODUCTION

Methanocaldococcus (Methanococcus) jannaschii, the first archaeon for which the whole genome was sequenced,[1] is an autotrophic hyperthermophile obligate anaerobic methanogen.[2] Physiologically, *M. jannaschii* has the ability to grow at an extreme pressure of >200 atm and a temperature of 94°C and is thus considered an extremophile.[3] Metabolic reconstruction for this organism is significant both from a scientific and a technical perspective. Scientifically, such a reconstruction will help us understand the structural and functional properties of methanogens and their relationships to other taxa within and across the Archaea. Technically, the re-annotation of a species whose genome sequence was released more than a quarter of a century ago[4] enables us to track progress in genome annotation efforts over time.[5] A metabolic reconstruction presented previously assigned enzymatic activity to 436 out of 1792 gene products.[6] While some of these annotations have been incorporated into the public databases, a reconstruction has only been available as a Tier 3 Pathway Genome Database (PGDB) since 2020.[7] Here, we present the outcome of a year-long, multi-level, labor-intensive, genome-wide annotation effort for *M. jannaschii* (NCBI accession number NC_000909.1) as an updated Tier 2, curated PGDB at BioCyc.org.

RESULTS

Sequence-level annotation: Function assignments

Of the 1847 genes in the *M. jannaschii* DSM 2661 chromosome available in this version, 769 were excluded from annotation updates as they remain hypothetical and mostly without any functional assignment, being examined only in the case of missing reactions in incomplete pathways.[8] Additionally, in this category, genes coding for tRNA and rRNA (44) and pseudogenes (46) are included, as they do not require any further annotation. The function descriptions of the remaining 1078 entries were either deemed as 'correct' (903 in total, status: annotated, stable description) or updated (175 in total, status: annotated, updated description) (Table 1). It is important to note that 28 years after the sequencing of the *M. jannaschii* genome, a little over a third of the genes remain uncharacterized and may still be part of the so-called 'microbial dark matter'.[9]

Of those 903 with a stable function description, 284 were assigned a more precise (or new) EC number. Of those 175 with

Table 1. Distribution of cases with stable or updated description and subtotals for the entire genome

status	stable description	updated description	total	annotated
unaffected by annotation	769	0	**769**	
annotated, same or no EC number	619	71	**690**	**(1078)**
annotated, updated EC number	284	104	**388**	
total	**1672**	**175**	**1847**	

an updated function description, 104 were assigned a more precise (or new) EC number (Table 1). These EC number updates for both sets are jointly available at: https://BioCyc.org/group?id=samo-63545-3898603617. These numbers reflect the status as of mid-2023; the annotation process for a PGDB is considered continuous, and numbers may change at the time of publication and subsequent updates in the near future (currently: 383 genes with 417 EC numbers).

We recorded over 600 gene products with enzymatic activity predictions (614 gene-reaction assignments with at least one EC number, amounting to 652 reactions, as some genes are multi-functional), a third of the protein-coding entries. Of those re-assessed (1078 gene products), 690 are recorded with the same (or no) EC number and 388 are updated (frequently equivalent to previous EC assignments as the EC number shifted to a new class) (Table 1).

The total number of inferred enzymatic reactions is 883, plus 25 transport reactions, corresponding to 540 enzymes/transporters (which include multimeric complexes, i.e., more than one gene per enzyme). By EC class,[10] we recorded 98 in EC 1 (oxidoreductases), 231 in EC 2 (transferases), 99 in EC 3 (hydrolases), 70 in EC 4 (lyases), 36 in EC 5 (isomerases), 73 in EC 6 (ligases), and just 7 in the new class EC 7 (translocases), 614 in total.

Finally, we have obtained additional bibliographic citations for 188 cases, supporting ~10% of the genome or 17% of those genes (188/1078) that were re-assessed. Another 460 publications are included from the automated process for relevant entries such as reactions or pathways, resulting in a total of 648 publications supporting the functional assignments and metabolic predictions in MjCyc, out of approximately 1000 publications on *M. jannaschii* available in PubMed.

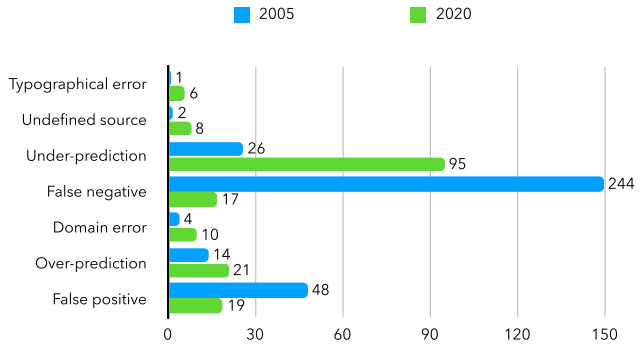


Figure 1. Comparison of annotation performance using TABS
The annotation lists from Tier-2 2005 and Tier-3 2020 are compared to Tier-2 2023. False negatives for 2005 are clipped at 150.

A wider set of database statistics are available at the following URL: https://biocyc.org/comp-genomics?tables=organism&tables=pathway&orgids=(MJ).

TABS-level scoring: Annotation quality

To assess progress in the encoded information and the potential sources of error between the (curated) Tier 2 MjCyc-2005 data[6] and the (computed) Tier 3 MjCyc-2020 data,[7] the TABS assignments[5] were calculated for both, with the final 2023 dataset used as the gold standard (see Methods). For the 2005 dataset, a total penalty score of 1499 was recorded for 339 cases (average value 4.42), compared to the 2020 dataset, where a total penalty score of 684 was recorded for 176 cases (average value 3.88), signifying genuine progress that derives from recent experimental characterization rather than mere computational annotation (Figure 1). The comparison and the TABS assignments are available as Data S1.

Most of the issues encountered in the 2005 Tier 2 dataset, for example the 244 false negative cases, derive from functional information that was not available more than twenty years ago, and do not necessarily represent technical artifacts. However, they are graded with regard to the most up-to-date dataset curated in 2023 and reported using the TABS scheme, to assess progress and accuracy among the different annotation datasets. The sharp contrast between 95 under-predictions and 17 false negatives against just 21 over-predictions and 19 false positives for the 2020 dataset (Figure 1) also shows that automated systems should tend to make conservative predictions, aiming for precision rather than coverage.

Pathway-level annotation: Archaeal biochemistry

For the 1078 genes that were assessed in detail, there were 104 cases (10%) with an updated EC number and a novel functional description, beyond the additional 284 updated EC numbers that kept the same functional description (Table 1). The total number of metabolic reactions inferred by (and curated with) Pathway Tools reached 883, corresponding to 540 enzymes and 142 individual pathways (101 complete and 41 incomplete pathways). This step links enzymatic reactions to 699 individual chemical compounds. In addition, 88 protein complexes were inferred by Pathway Tools (Figure 2).

Pathway quality control and metrics

To further assess the quality of annotation at the pathway level, pathway scores were plotted against the total number of reactions present in the corresponding pathway. Pathway score is a number between 0 and 1 inclusive indicating the likelihood (not a probability) that a pathway is present.[7] A contour plot reveals that high

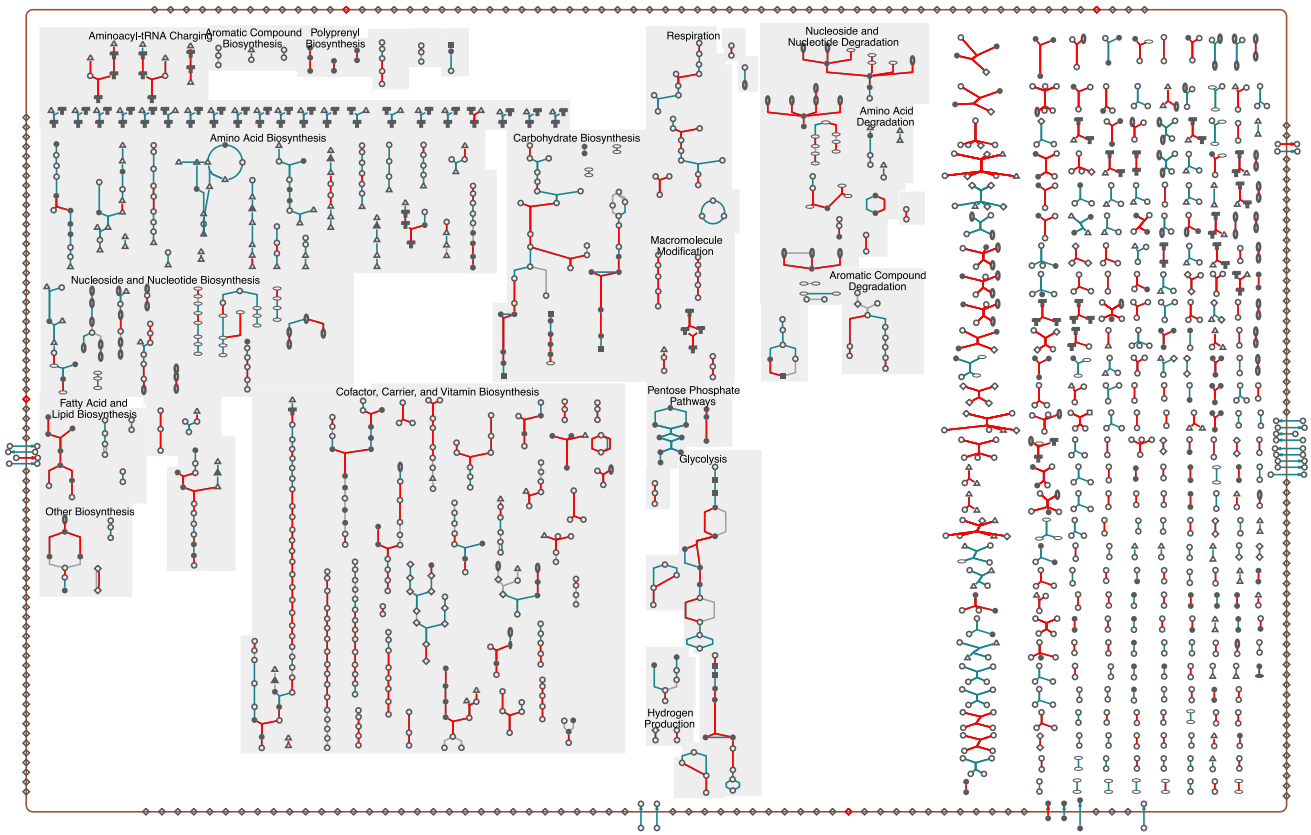


Figure 2. A snapshot of curated pathways for MjCyc

Reactions are marked according to their level of detection by automatic (blue) or manual curation (red) processes. Missing reactions are also shown (gray), defined as those reactions that have not been associated with a gene.

scores are present for even complex pathways, defined here as those with more than five reactions (Figure 3). Those include the pathways for histidine, arginine, lysine and tryptophan biosynthesis, as well as methanogenesis, factor 430 biosynthesis, tetrapyrrole biosynthesis and others, providing strong evidence for completeness and accuracy (Table 2). Additional statistics are available through Special SmartTables accessible at: https://biocyc.org/groups?tab=SPECIAL&orgid=MJ for example, all pathways are listed at: https://BioCyc.org/group?id=:ALL-PATHWAYS&orgid=MJ. Interestingly, although non-detection (and possibly absence) of function assignments for some reactions of longer pathways is expected, there is a substantial group of fully detected complete or near-complete pathways (right part, Figure 3), contrasted with a group of typically short pathways that remain incomplete, and uncharacterized. All incomplete pathways, especially those with one or two missing reactions, form an important target for future enzyme discovery efforts in methanogenic Archaea.

Detection and discovery of novel enzymes for archaea

Pathway holes are prime targets of potential false negative cases in function assignment and have been examined at the final cycle of the PGDB annotation. An exemplary case is represented by MJ0395 (NCBI protein identifier: WP_010869894.1), that was detected as the dephospho-CoA kinase catalyzing the last reaction of the archaeal coenzyme A biosynthesis III pathway (EC 2.7.1.237),[11] available at: https://biocyc.org/pathway?orgid=MJ&id=PWY-8342.

The pathway for adenosyl cobinamide-GDP salvage from cobinamide II is detected in its entirety thanks to the novel functional assignment of gene MJ1613 (UniProt: uncharacterized protein) as EC 3.5.1.90, adenosyl cobinamide amidohydrolase (CbiZ), based on its similarity to the *Methanosarcina mazei* CbiZ (25% identity, UniProt accession: Q8Q0G3). The downstream steps are catalyzed by the gene products MJ1314 (UniProt: probable cobalamin biosynthesis protein, CobD) which has been annotated as EC 6.3.1.10, adenosyl cobinamide-phosphate synthase, and MJ1117 (UniProt: adenosyl cobinamide-phosphate guanylyltransferase, CobY), which has been annotated as EC 2.7.7.62, adenosyl cobinamide-phosphate guanylyltransferase.[12]

Another finding obtained by the pathway hole filling tool is the detection of MJ1598 (UniProt: UPF0284 protein MJ1598; PDB identifier: 3L0Z) as EC 2.4.2.21, nicotinate-nucleotide-dimethylbenzimidazole phosphoribosyltransferase (CobT), an enzyme that participates in three distinct biochemical pathways involving cobamide biosynthesis.[13] Related steps are catalyzed by MJ1438 (UniProt: adenosyl cobinamide-GDP ribazoletransferase,

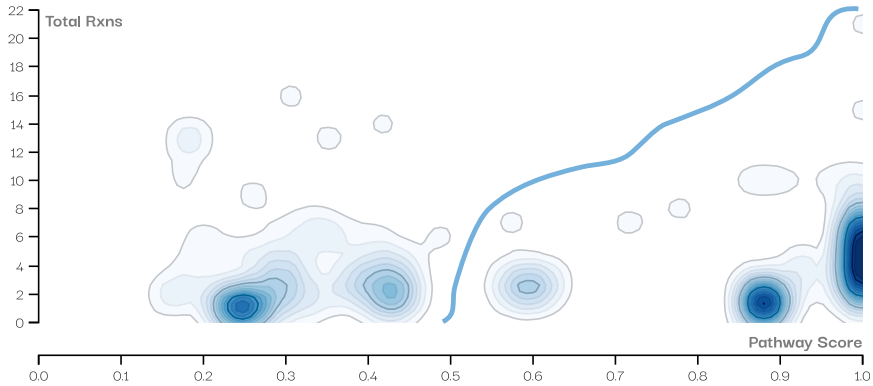


Figure 3. Contour plot for pathway score versus complexity in MjCyc

Pathway scores signify the likelihood of pathway presence (x axis), pathway complexity is measured as the number of reactions (y axis). A freehand curve represents the group of pathways with score>0.5.

CobS), which has been annotated as EC 2.7.8.26, adenosyl cobinamide-GDP ribazoletransferase, and MJ1330 (CobZ) (UniProt: uncharacterized protein MJ1330), which has been annotated as EC 3.1.3.73, adenosyl cobamide phosphatase, due to its similarity to the *Methanothrix soehngenii* (*Methanosaeta concilii*) CobZ (38% identity, UniProt accession: F4BUV0).[14,15]

Resolution of paralogous function assignments is also possible on a genome scale. An example is the pair of MJ0865 (UniProt: putative methylthiotransferase MJ0865) and MJ1487 (UniProt: uncharacterized methyltransferase), both radical-SAM proteins (32% identity) involved in 5,6-dimethylbenzimidazole biosynthesis II (anaerobic). Our analysis suggested that gene product MJ0865 is associated with the 5-methoxy-6-methylbenzimidazole synthesis step (EC 2.1.1.-) as 5-methoxybenzimidazole C-methyltransferase (24% identity to BzaD from *Eubacterium callanderi*, UniProt accession: E3GQB2), while gene product MJ1487 is associated with the downstream, final step (no EC number available) of 5,6-dimethylbenzimidazole synthesis as anaerobic 5,6-dimethylbenzimidazole synthase (29% identity to BzaE from *E. callanderi*, UniProt accession: E3GQB3).[16]

Finally, one striking example of a novel function through the combined use of sequence analysis and metabolic reconstruction is represented by MJ0570 (UniProt: uncharacterized protein MJ0570), as EC 6.3.1.14, diphthamide synthase (30% identity to diphthine-ammonia ligase from yeast, UniProt accession: Q12429).[17] This assignment completes the previously incomplete diphthamide biosynthesis I pathway for archaea.[18]

The selected examples above provide strong predictions that go beyond simple sequence similarity and take into account the context at the genomic (e.g., paralogs) and metabolic (e.g., pathway holes) levels on a genome-wide scale.[19] While these function assignments represent falsifiable predictions, the evidence from multiple sources such as sequence comparisons, literature searches, pathway scores, and other elements suggests that they form the best possible choices in the case of *M. jannaschii* and may be extended throughout the archaeal domain, once verified experimentally.

DISCUSSION

High-quality, expert curation is of major importance in metabolic reconstruction aiming at precision rather than coverage, as missing reactions can be discovered in cycles of annotation when pathway inference is in place, providing a bird's-eye view of cellular metabolism. We note that in the absence of experimental confirmation, all assignments are considered as testable predictions. In the case of methanogens, critical components participating in methanogenesis can be misidentified by a purely automated approach.

For instance, our findings suggest that the product of gene MJ0879 should be assigned the function of one of the subunits of Ni-sirohydrochlorin a,c-diamide reductive cyclase (EC 6.3.3.7), an enzyme critical to the biosynthesis of the nickel-containing tetrapyrrole cofactor factor 430, which is required by methyl coenzyme M reductase (EC 2.8.4.1).[20] The latter catalyzes the final step in methanogenesis and the production of methane (CH_4, see also: https://biocyc.org/pathway?orgid=MJ&id=METHFORM-PWY). Yet, MJ0879 was previously identified as a general-purpose nitrogenase iron protein (UniProt: Nitrogenase iron protein, EC 1.18.6.1 as 'nitrogenase'), inferred by the protein family motifs (nifH), without a specific functional assignment to class 6 as a ligase.

Regarding progress in annotation efforts, it is evident that both the accumulation of experimental data and the improvement of comparative genomics methods furnish novel information that can be captured in PGDBs for milestone species such as *M. jannaschii*, the first archaeal species whose genome was sequenced in its entirety, back in 1996 (Figure 4).

The current version of MjCyc provides the latest high-quality metabolic reconstruction for *M. jannaschii*, made available to the wider community for research and development efforts in methanogen biology.

RESOURCE AVAILABILITY

Lead contact

Further information and requests about this study should be directed to and will be fulfilled by the lead contact, Christos A. Ouzounis (cao@csd.auth.gr).

Materials availability

This study did not generate reagents.

Data and code availability

- Data: PGDB available at: https://BioCyc.org/MJ/.
- Code: Conversion to PathoLogic available at: https://github.com/GioStamoulos/PLFG-Toolkit.

Table 2. Complex pathways with score 1.0 and specific reactions detected

Pathway	Total	Present	Other
cob(II)yrinate a,c-diamide biosynthesis I (early cobalt insertion)	15	15	0
L-histidine biosynthesis	10	10	0
L-arginine biosynthesis II (acetyl cycle)	9	9	4
factor 430 biosynthesis	7	7	0
L-lysine biosynthesis VI	7	7	0
methanogenesis from H2 and CO2	6	6	0
C20,20 CDP-archaeol biosynthesis	6	6	0
reductive acetyl coenzyme A pathway II (autotrophic methanogens)	6	6	0
UMP biosynthesis III	6	6	1
L-tryptophan biosynthesis	6	6	0
tetrapyrrole biosynthesis I (from glutamate)	6	6	0
inosine-5′-phosphate biosynthesis III	6	6	0

Total: total number of reactions in pathway; Present: detected reactions in pathway; Other: reactions present in other pathways

- Other: Any additional information required to analyze the data or code reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank the members of the BCCB (CSD-AUTH) and BRG groups (AIC-SRI) and multiple collaborators in past efforts. The authors recognize the contributions of various whose work is cited within the PGDB only, as it cannot be included in the bibliography due to space limitations. This paper is dedicated to the memory of Carl R. Woese.

Funding information: CAO acknowledges support from Elixir-GR (MIS 5002780), Action "Reinforcement of the Research and Innovation Infrastructure" the Operational Program Competitiveness, Entrepreneurship, and Innovation (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). P.D.K. acknowledges support from grant R01AI160719 from the National Institute of Allergy and Infectious Diseases in the National Institutes of Health. N.C.K. was supported by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

Conceptualization, I.B., R.C., P.D.K., N.C.K., and C.A.O.; Methodology, I.B., P.D.K., and C.A.O.; Software, I.B., G.S., C.B., A.K., and C.A.O.; Validation, I.B., K.T., and C.A.O.; Investigation, I.B., I.K., and A.D.; Writing – Original Draft, I.B. and C.A.O.; Writing – Review and Editing, I.B., R.C., P.D.K., N.C.K., and C.A.O.; Funding Acquisition, I.I., P.D.K., N.C.K., and C.A.O.; Resources, R.C., P.D.K., N.C.K., and C.A.O.; Supervision, P.D.K. and C.A.O.; Project administration, C.A.O.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Re-annotation approach for updates: Overview
  - Initial genome data
  - Annotation updates
  - Enzyme annotation
  - Literature searches
  - Curation of PGDB
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - PGDB iteration
  - Annotation metrics
  - Notes

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111546.

Received: September 11, 2023
Revised: August 1, 2024
Accepted: December 3, 2024
Published: December 5, 2024

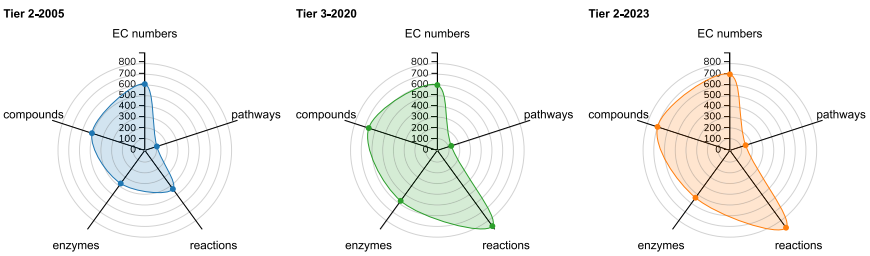


Figure 4. Annotation levels for five classes in *M. jannaschii* PGDBs

The number of pathways, reactions, enzymes, compounds and EC numbers are shown for the curated version of MjCyc (2005, blue), the automatically derived version (2020, green) and the current version (2023, orange). These radar graphs facilitate the comparison of the three annotation efforts reflected in the corresponding datasets.

REFERENCES

1. Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science *273*, 1058–1073. https://doi.org/10.1126/science.273.5278.1058.
2. Jeanthon, C., L'Haridon, S., Reysenbach, A.L., Vernet, M., Messner, P., Sleytr, U.B., and Prieur, D. (1998). *Methanococcus infernus* sp. nov., a novel hyperthermophilic lithotrophic methanogen isolated from a deep-sea hydrothermal vent. Int. J. Syst. Bacteriol. *48*, 913–919. https://doi.org/10.1099/00207713-48-3-913.
3. Jones, W.J., Leigh, J.A., Mayer, F., Woese, C.R., and Wolfe, R.S. (1983). *Methanococcus jannaschii* sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. Arch. Microbiol. *136*, 254–261. https://doi.org/10.1007/BF00425213.
4. Kyrpides, N.C., Olsen, G.J., Klenk, H.P., White, O., and Woese, C.R. (1996). *Methanococcus jannaschii* genome: revisited. Microb. Comp. Genomics. *1*, 329–338.
5. Ouzounis, C.A., and Karp, P.D. (2002). The past, present and future of genome-wide re-annotation. Genome Biol. *3*, COMMENT2001. https://doi.org/10.1186/gb-2002-3-2-comment2001.
6. Tsoka, S., Simon, D., and Ouzounis, C.A. (2004). Automated metabolic reconstruction for *Methanococcus jannaschii*. Archaea *1*, 223–229. https://doi.org/10.1155/2004/324925.
7. Karp, P.D., Midford, P.E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., Ong, W.K., Subhraveti, P., Caspi, R., Fulcher, C., et al. (2021). Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Brief. Bioinformatics. *22*, 109–126. https://doi.org/10.1093/bib/bbz104.
8. Galperin, M.Y., and Koonin, E.V. (2004). Conserved hypothetical' proteins: prioritization of targets for experimental study. Nucleic Acids Res. *32*, 5452–5463. https://doi.org/10.1093/nar/gkh885.
9. Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2019). Towards functional characterization of archaeal genomic dark matter. Biochem. Soc. Trans. *47*, 389–398. https://doi.org/10.1042/BST20180560.
10. McDonald, A.G., and Tipton, K.F. (2023). Enzyme nomenclature and classification: the state of the art. FEBS J. *290*, 2214–2231. https://doi.org/10.1111/febs.16274.
11. Shimosaka, T., Makarova, K.S., Koonin, E.V., and Atomi, H. (2019). Identification of dephospho-Coenzyme A (Dephospho-CoA) kinase in *Thermococcus kodakarensis* and elucidation of the entire CoA biosynthesis pathway in Archaea. mBio *10*, 10–128. https://doi.org/10.1128/mBio.01146-19.
12. Otte, M.M., and Escalante-Semerena, J.C. (2009). Biochemical characterization of the GTP:adenosylcobinamide-phosphate guanylyltransferase (CobY) enzyme of the hyperthermophilic archaeon *Methanocaldococcus jannaschii*. Biochemistry *48*, 5882–5889. https://doi.org/10.1021/bi8023114.
13. Jeter, V.L., Schwarzwalder, A.H., Rayment, I., and Escalante-Semerena, J.C. (2022). Structural studies of the phosphoribosyltransferase involved in cobamide biosynthesis in methanogenic archaea and cyanobacteria. Sci. Rep. *12*, 17175. https://doi.org/10.1038/s41598-022-21765-5.
14. Barber, R.D., Zhang, L., Harnack, M., Olson, M.V., Kaul, R., Ingram-Smith, C., and Smith, K.S. (2011). Complete genome sequence of *Methanosaeta concilii*, a specialist in aceticlastic methanogenesis. J. Bacteriol. *193*, 3668–3669. https://doi.org/10.1128/JB.05031-11.
15. Zayas, C.L., Woodson, J.D., and Escalante-Semerena, J.C. (2006). The cobZ gene of *Methanosarcina mazei* Go1 encodes the nonorthologous replacement of the alpha-ribazole-5'-phosphate phosphatase (CobC) enzyme of *Salmonella enterica*. J. Bacteriol. *188*, 2740–2743. https://doi.org/10.1128/JB.188.7.2740-2743.2006.
16. Hazra, A.B., Han, A.W., Mehta, A.P., Mok, K.C., Osadchiy, V., Begley, T.P., and Taga, M.E. (2015). Anaerobic biosynthesis of the lower ligand of vitamin B12. Proc. Natl. Acad. Sci. USA *112*, 10792–10797. https://doi.org/10.1073/pnas.1509132112.
17. Su, X., Lin, Z., Chen, W., Jiang, H., Zhang, S., and Lin, H. (2012). Chemogenomic approach identified yeast YLR143W as diphthamide synthetase. Proc. Natl. Acad. Sci. USA *109*, 19983–19987. https://doi.org/10.1073/pnas.1214346109.
18. Su, X., Lin, Z., and Lin, H. (2013). The biosynthesis and biological function of diphthamide. Crit. Rev. Biochem. Mol. Biol. *48*, 515–521. https://doi.org/10.3109/10409238.2013.831023.
19. Promponas, V.J., Iliopoulos, I., and Ouzounis, C.A. (2015). Annotation inconsistencies beyond sequence similarity-based function prediction - phylogeny and genome structure. Stand. Genomic Sci. *10*, 108. https://doi.org/10.1186/s40793-015-0101-2.
20. Moore, S.J., Sowa, S.T., Schuchardt, C., Deery, E., Lawrence, A.D., Ramos, J.V., Billig, S., Birkemeyer, C., Chivers, P.T., Howard, M.J., et al. (2017). Elucidation of the biosynthesis of the methane catalyst coenzyme F430. Nature *543*, 78–82. https://doi.org/10.1038/nature21427.
21. Iliopoulos, I., Tsoka, S., Andrade, M.A., Enright, A.J., Carroll, M., Poullet, P., Promponas, V., Liakopoulos, T., Palaios, G., Pasquier, C., et al. (2003). Evaluation of annotation strategies using an entire genome sequence. Bioinformatics *19*, 717–726. https://doi.org/10.1093/bioinformatics/btg077.
22. Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M., and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. *44*, 6614–6624. https://doi.org/10.1093/nar/gkw569.
23. Bairoch, A. (2000). The ENZYME database in 2000. Nucleic Acids Res. *28*, 304–305. https://doi.org/10.1093/nar/28.1.304.
24. Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P., and Karp, P.D. (2020). The MetaCyc database of metabolic pathways and enzymes–a 2019 update. Nucleic Acids Res. *48*, D445–D453. https://doi.org/10.1093/nar/gkz862.
25. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.
26. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584. https://doi.org/10.1093/nar/30.7.1575.
27. Green, M.L., and Karp, P.D. (2004). A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC. Bioinformatics. *5*, 76. https://doi.org/10.1186/1471-2105-5-76.
28. Mauri, M., Elli, T., Caviglia, G., Uboldi, G., and Azzi, M. (2017). RAWGraphs: A Visualisation Platform to Create Open Outputs. In Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter (ACM), pp. 28:1–28:5. https://doi.org/10.1145/3125571.3125585.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
pathway-tools	biocyc	NA
Other		
M. jannaschii genome sequence	NCBI	GCF_000091665.1

METHOD DETAILS

Re-annotation approach for updates: Overview

We started with a merge of the manually curated Tier 2 MjCyc-2005[6] and the automatically generated Tier 3 MjCyc-2020[7] annotations, prioritized according to functional descriptions (hypothetical proteins were seen as the last priority). Each protein-coding gene was subjected to manual inspection and comparison of available description lines; if a conflict was present, a number of steps were taken, including sequence similarity searches and assessment against the corresponding records in UniProtKB, Pfam and MetaCyc. Particular attention was paid to those protein sequences likely to be associated with an Enzyme Commission (EC) number, as this impacts their qualification as candidate enzymes. Transporter Classification (TC) was not recorded, as this system cannot be used to compare with previous efforts (prior to 2010). Some genes considered critical, e.g., those involved in methanogenesis or multiple paralogs, were examined more thoroughly. A full genome clustering was also performed to confirm annotation consistency and functional diversification among paralogous genes, available as Data S2. Finally, extensive literature searches were executed to support the annotation process and provide functional information based on reported experiments.

In a number of cases, functionally annotated proteins from closely related organisms provided clues, as *M. jannaschii* is not currently being studied through systematic omics experiments. When an enzymatic activity in other organisms was available, reverse sequence searches against the full genome were used with the corresponding enzyme sequence to detect the highest-similarity protein-coding gene, typically amongst diverse paralogs. To maintain attention to detail, a proposed 'buddy' system was adopted,[21] i.e., two persons were present when committing a correction or update at all times. In all, we aimed for precision rather than coverage to achieve completion within a reasonable time frame, with multiple, regular sessions during 2022–2023. A final cycle of annotation was carried out for the sequence-based assignments, using Pathway Tools, the software suite that powers BioCyc.org and integrates functional predictions into metabolic pathways, including support for pathway hole-filling.

Initial genome data

The genome sequence of *M. jannaschii* DSM 2661[1] was directly imported from NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000091665.1/) and annotated using Prokaryotic Genome Annotation Pipeline (PGAP) 6.1[22] for gene finding (see NCBI record for updates https://www.ncbi.nlm.nih.gov/nuccore/NC_000909, and comparison to GenBank annotation details).

Annotation updates

The Tier 3 MjCyc-2020 catalog for *M. jannaschii* DSM 2661 was exported to PathoLogic (PF) format (latest annotation date: 2022-04-11). This catalog was converted to a spreadsheet format and was compared to a list of legacy annotations from 2005.[6] All manual updates were recorded during the extended curation process and were re-converted back to PF for import into BioCyc via Pathway Tools.[7] The elements of manual curation, re-annotation and function assignments on a genome scale have been documented elsewhere.[21] The toolkit for a first-pass conversion of a custom spreadsheet to PF is available here: https://github.com/GioStamoulos/PLFG-Toolkit.

Enzyme annotation

All genes for possible enzyme assignments were compared to entries in the ENZYME database[23] on a case-by-case basis; this step issued re-assignments when EC numbers have been updated. Pathway holes (see below) were investigated using MetaCyc information.[24] Reverse BLAST searches[25] with query enzymes of interest against the *M. jannaschii* genome were also performed, in specific cases. An all-against-all comparison using BLAST and clustering with MCL was also executed,[26] to delineate genome-wide paralog groups and protein families.

Literature searches

All PubMed entries referring to *M. jannaschii* DSM 2661 (approx. 1000 publications) were investigated in parallel with the curation of individual genes; in some cases, assignments were performed from the literature to the gene. All literature records used in the process were recorded in the spreadsheet and imported into the PGDB. A total of 188 citations were captured, corresponding to 10% of the gene number. The total number of publications currently supporting the PGDB contents exceeds 700 references.

Curation of PGDB

Using the web-server edit mode of BioCyc.org that provides significant functionality for Pathway Tools, the PGDB was remotely curated by two persons at all times. Function assignment updates, EC numbers and citations to the literature were recorded for hundreds of genes (see Tables 1 and 2).

QUANTIFICATION AND STATISTICAL ANALYSIS

PGDB iteration

An important aspect of genome-wide annotation and in particular pathway inference is the step for pathway-hole fillling.[27] Pathway Tools enables this functionality for PGDB curation.[7] Starting from 110 incomplete pathways, missing enzymes were discovered (in some cases as original reports, see Results), thus reducing the number of incomplete pathways – now reported as: 64 pathway holes (13.2%) are present in the 484 total reactions of 137 pathways (https://biocyc.org/MJ/missing-rxns.html).

Annotation metrics

To assess progress of annotation with regard to transitive annotation issues, the TABS scheme was used,[5] with the curated 2023 annotation list as gold-standard. This qualitative scheme has been proposed as a way to monitor annotation 'distance' with the highest penalty recorded for a false positive case (score 7), followed by an over-prediction (score 6), a domain error (multi-domain architecture) (score 5), a false negative case (score 4), an under-prediction (score 3), cases with an undefined source (score 2), a typographical error (score 1) and finally total agreement (score 0). This scale reflects the risk of error propagation in databases and provides a simple measure of annotation consistency.

Radar graphs for the three versions of PGDBs including the latest update used in this study to depict annotation progress were generated by RawGraphs 2.0 beta.[28]

Notes

Fraction of genes devoted to metabolism 559 of 1886 (30%). Total reactions in the base pathways 513. Tier 3 status: base pathway reactions that are pathway holes 118 (23%), base pathway reactions that are not pathway holes 395 (77%). Tier 2 status: base pathway reactions that are pathway holes 68 (13%); base pathway reactions that are not pathway holes 445 (87%). Improvement: 10% over all reactions.