



Whole-genome sequencing of diverse 351 cultured prokaryotes including as-yet-unsequenced type strains

Shingo Kato, Sachiko Masuda, Arisa Shibata, et al.

Genome Res. 2026 36: 875-884

Access the most recent version at doi:[10.1101/gr.281026.125](https://doi.org/10.1101/gr.281026.125)

References

This article cites 69 articles, 8 of which can be accessed free at:
<http://genome.cshlp.org/content/36/4/875.full.html#ref-list-1>

Creative Commons License

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

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



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

Resource

Whole-genome sequencing of diverse 351 cultured prokaryotes including as-yet-unsequenced type strains

Shingo Kato,¹ Sachiko Masuda,² Arisa Shibata,² Takashi Itoh,¹ Mitsuo Sakamoto,¹ Ken Shirasu,² and Moriya Ohkuma¹

¹Japan Collection of Microorganisms, RIKEN BioResource Research Center, Tsukuba, Ibaraki 305-0074, Japan; ²Plant Immunity Research Group, RIKEN Center for Sustainable Resource Science, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Genome sequences provide fundamental information for both basic and applied life sciences. Whole-genome sequencing is now requested for describing novel prokaryotic species and designating their type strains, which serve as representative and well-characterized strains of the species. Indeed, the number of sequenced prokaryotic genomes has been rapidly increasing. However, a considerable number of isolated strains, particularly technically demanding strains such as strict anaerobes and slow growers, remain without genome sequence information. Here we report the whole-genome sequencing of 290 bacterial and 61 archaeal strains, including technically demanding type strains, obtained from the Japan Collection of Microorganisms (JCM) using a combination of short- and long-read sequencing technologies. The data set includes 284 type-strain genomes and 235 complete genomes. Notably, in the data set, genomes of more than 200 strains, including more than 150 type strains, have not been made publicly available. Comparative genomic analysis suggests that some strains need to be assigned to novel taxa or reclassified. Functional gene survey indicates that some strains possess previously unrecognized potential for carbon fixation or bioactive secondary metabolite production. Our data set will contribute to more accurate taxonomic classification, fill gaps in the phylogeny of prokaryotes, and provide insights into their physiology and ecology.

[Supplemental material is available for this article.]

Whole-genome sequences serve as a fundamental blueprint for life, providing the foundation for a broad range of life science research. The genomes of prokaryotes, that is, archaea and bacteria, are typically smaller (<10 Mbp) and simpler in structure, consisting of a single circular chromosome in most cases, than those of eukaryotes. The rapid development of DNA sequencing technologies and the increase in computational power have driven a significant expansion in the number of both sequenced prokaryotic genomes from cultivated strains and from culture-independent microbiomes. In particular, type strains, which are representatives of species with well-characterized phenotypic information, have been a primary focus of large-scale whole-genome sequencing projects (Mukherjee et al. 2017; Wu et al. 2018). In addition, since 2018, the determination of the genome sequence of the type strain and comparison with those of related species have been required when describing novel prokaryotic species (Chun et al. 2018; Riesco and Trujillo 2024), resulting in a substantial accumulation of type-strain genomes in public databases. Furthermore, recent advances in long-read sequencing technologies have enabled the resequencing and completion of type-strain genomes (Dicks et al. 2023; Nguyen et al. 2024). However, a considerable number of genomes from technically demanding strains, such as slow-growing strict anaerobes with low maximum cell density, remain unsequenced owing to the challenges associated with their cultivation and DNA extraction. Therefore, gaps in the prokaryotic

phylogeny remain to be filled, and the full extent of their metabolic potentials remains unclear.

Here, we report the whole-genome sequencing of 351 publicly available prokaryotic strains, including technically demanding and/or previously unsequenced type strains, obtained from the Japan Collection of Microorganisms (JCM), one of the international culture collections. Despite the rapid accumulation of metagenome-assembled genomes (MAGs), genomes obtained from isolated strains remain invaluable, as they are guaranteed to be derived from a single organism and can be directly linked to experimentally validated phenotypes. The genome analysis of the isolated strains will help to uncover biological characteristics that are difficult to resolve from MAGs alone owing to their incompleteness, assembly artifacts, and ambiguous phenotypic predictions.

Results

We determined the whole-genome sequences of 351 microbial strains, including 284 type strains (Fig. 1A; Supplemental Tables S1, S2). Of the 351 genomes, 235 genomes (including 189 type-strain genomes) were completed. The size of the determined genomes ranged from 1,265,941 bp for *Methanothermobacter sociabilis* JCM 10723^T to 10,631,713 bp for *Streptomyces phaeofaciens* JCM 4814^T (Supplemental Fig. S1). The genome size of *M. sociabilis*

Corresponding authors: skato@riken.jp, mohkuma@riken.jp

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.281026.125>.

© 2026 Kato et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <https://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

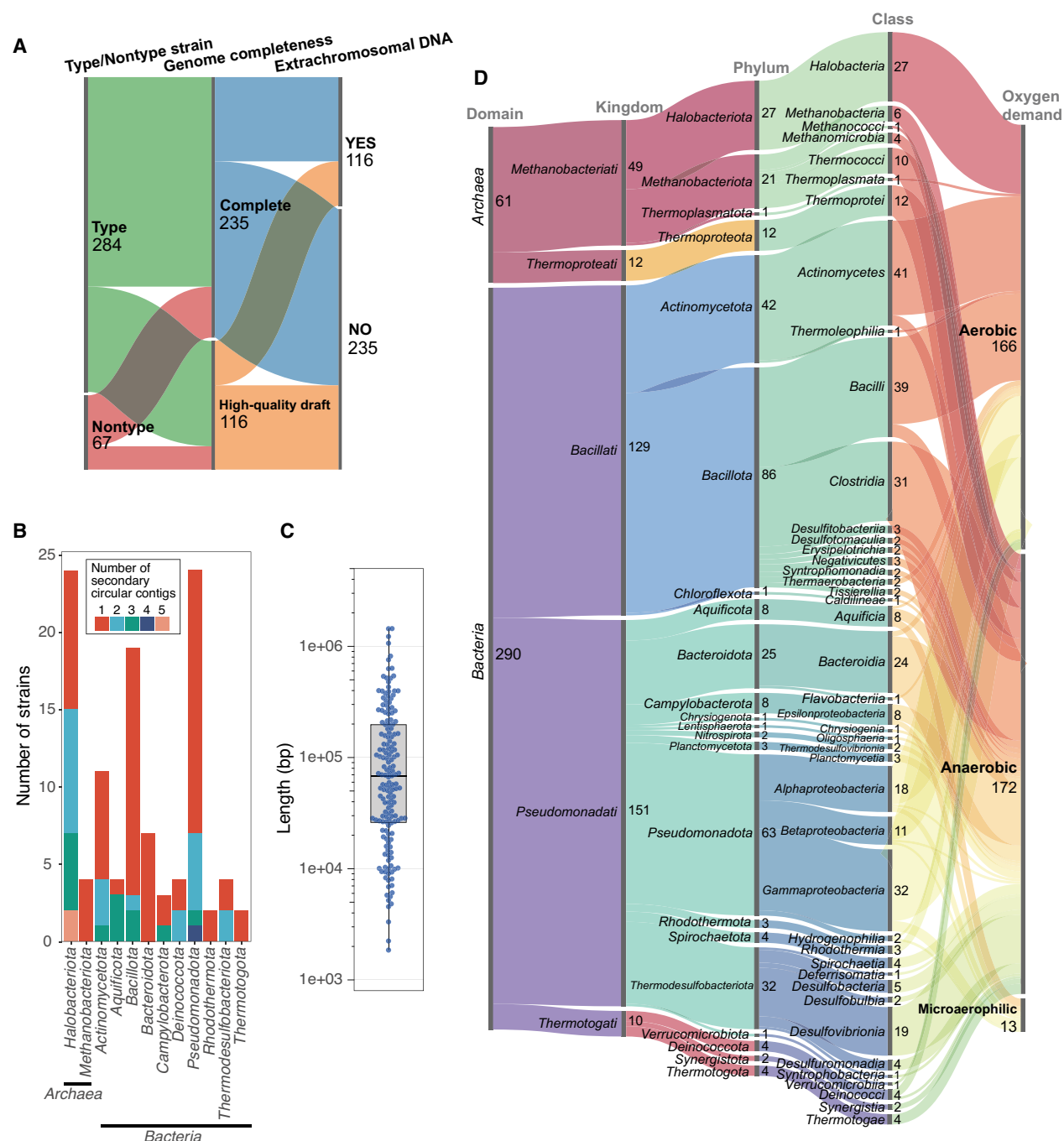


Figure 1. Overview of the determined genomes of the 351 strains. (A) Alluvial diagram showing the distribution of type and nontype strains, genome completeness, and the presence of extrachromosomal DNA. (B) Bar chart showing the number of strains harboring one or more secondary circular contigs. The number of secondary circular contigs is indicated by different colors as shown in the inset box. (C) Box plot showing the size distribution of the secondary circular contigs. (D) Alluvial diagram showing the taxonomic classification and oxygen requirement of the 351 strains.

JCM 10723^T is comparable to that of other small genomes of free-living organisms, such as “*Nitrosopelagicus brevis*” CN25 (1.232 Mbp) and *Methanothermobacter ferredoxin* V24S^T (1.243 Mbp). The G+C content of the genomes varied widely, ranging from 24% for *Exilispira thermophila* JCM14728^T to 76% for *Cellulomonas pakistensis* JCM18755^T. The values of genome completeness and contamination levels ranged from 84.3% to 100% and from zero to 13.7%, respectively (Supplemental Table S2). A total of 209 (38 ar-

chaeal and 171 bacterial) genomes, including 166 type-strain genomes, were being released from public databases for the first time (Fig. 2). One or more (up to 16) copies of 16S rRNA genes were found in 331 genomes.

A total of 76 complete genomes contained two or more contigs (Fig. 1B; Supplemental Table S2). In particular, 68 of these genomes included secondary circular contigs. Furthermore, 40 draft genomes contained smaller circular contigs than each of the

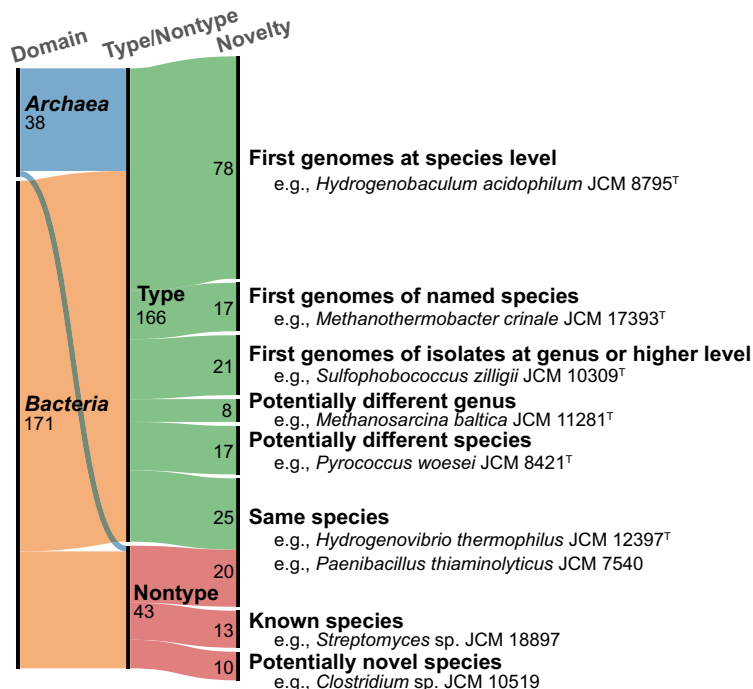


Figure 2. Genomic and taxonomic novelty. Alluvial diagram showing the genomic and taxonomic novelty of the 209 strains with newly released genomes in this study. Details are provided in the main text and in Supplemental Table S3.

longest contig. Up to five secondary circular contigs were detected per genome (Fig. 1B). The size of these circular contigs ranged from 1847 to 1,436,886 bp (Fig. 1C).

The data set encompassed phylogenetically diverse prokaryotes, comprising 61 archaeal and 290 bacterial strains (Fig. 1D; Supplemental Table S3). The taxonomic classification, based on the International Code of Nomenclature of Prokaryotes (ICNP), showed that our data set spanned four archaeal phyla with seven classes and 18 bacterial phyla with 36 classes. The scientific names for most species were validly published; however, the proposed names for 19 species were not yet validated. The results of the Genome Taxonomy Database (GTDB)-based taxonomic classification (Parks et al. 2022) and from the Type Strain Genome Server (Meier-Kolthoff et al. 2022) were also summarized in Supplemental Table S3. Moreover, the determined genomes represented physiologically diverse prokaryotes, including anaerobes and microaerophiles (Fig. 1D; Supplemental Table S2), and mesophiles to hyperthermophiles (Supplemental Table S1). The predicted optimal growth temperature (OGT) of each strain ranged

from 21.8°C to 95.8°C (Supplemental Table S4).

In the determined genomes, we identified genes associated with a variety of known metabolic functions (Supplemental Table S5). For example, key genes for methanogenesis (*mcr* for methyl-coenzyme M reductase) and methane oxidation (*pmo* for particulate methane monooxygenase) were detected in the genomes of methanogenic archaea belonging to *Methanobacteriota*, such as *Methanobacterium movens* JCM 15415^T, and methane-oxidizing bacteria belonging to *Pseudomonadota*, such as *Methylosoma difficile* JCM 14076^T, respectively. Regarding the potential for carbon fixation, we showed that 67 genomes possessed the complete or nearly complete gene set with the key genes for the Calvin-Benson-Bassham (CBB) cycle, reverse tricarboxylic acid (rTCA) cycle, Wood-Ljungdahl (WL) pathway, and reversed oxidative TCA (roTCA) cycle (Fig. 3A; Supplemental Table S6). We also showed that key genes for carbon fixation pathways were missing in the genomes of four strains previously reported as autotrophs (Fig. 3B). Regarding

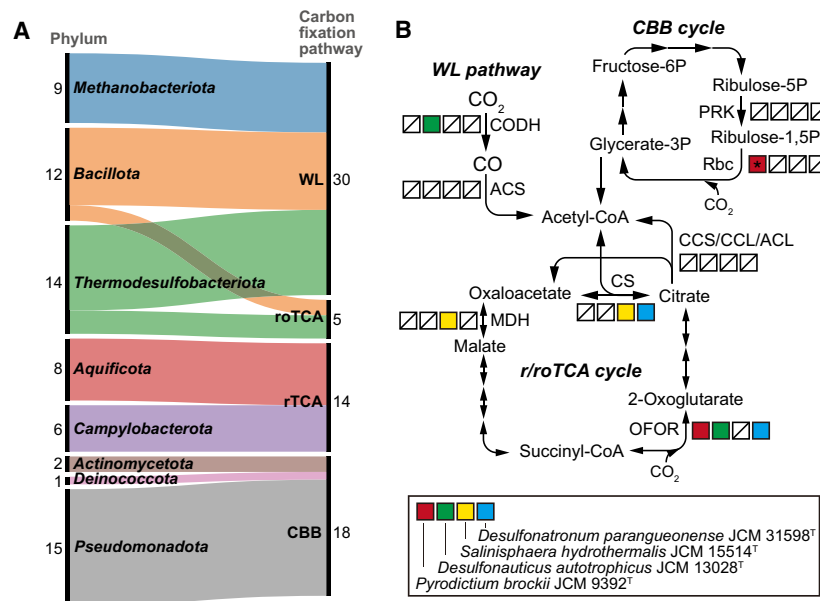


Figure 3. Carbon fixation pathways encoded in the genomes. (A) Alluvial diagram showing the phylum-level classification of the 67 strains possessing key genes for each carbon fixation pathway. (WL) Wood-Ljungdahl pathway; (roTCA) reversed oxidative tricarboxylic acid cycle; (rTCA) reverse tricarboxylic acid cycle; (CBB) Calvin-Benson-Bassham cycle. (B) Presence of key genes for each carbon fixation pathway in the genomes of the four autotrophic strains highlighted in the box. (*) A gene for RuBisCO form III, but not forms I or II, was detected in the genome of *Pyrodicticum brockii* JCM 9392^T (for details, see the main text). (CCS) Citryl-CoA synthetase; (CCL) citryl-CoA lyase; (ACL) ATP-citrate lyase; (CODH) carbon monoxide dehydrogenase; (ACS) acetyl-CoA synthase; (PRK) phosphoribulokinase; (Rbc) ribulose-bisphosphate carboxylase; (CS) citrate synthase; (OFOR) 2-oxoacid:ferredoxin oxidoreductases; (MDH) malate dehydrogenase.

the potential for secondary metabolite production, we identified 1696 biosynthetic gene clusters (BGCs) in 281 genomes (Fig. 4; Supplemental Table S7). Of the 235 complete genomes, 222 (94.9%) genomes possessed one or more BGCs. The BGC counts were roughly correlated with the genome sizes in our data set ($r^2 = 0.524$) (Fig. 4A), a trend that has been previously reported for *Actinomyces* genomes (Seshadri et al. 2022). The types of BGCs, such as nonribosomal peptide synthetases (NRPSs), polyketide synthases (PKSs), and ribosomally synthesized and post-translationally modified peptides (RiPPs), varied among taxa (Fig. 4B). For instance, PKS-categorized BGCs were only found in the bacterial genomes. Terpene-categorized BGCs were rare in the genomes of *Methanobacteriota* and *Thermodesulfobacteriota*, both of which include obligate anaerobes. In contrast, RiPP-categorized BGCs were widely detected among the bacterial and archaeal phyla.

Discussion

In this study, we determined the whole-genome sequences of 351 prokaryotic strains, spanning a broad range of archaeal and bacterial taxa, including previously unsequenced type strains. The data set of complete or near-complete genomes provides novel insights into taxonomy and biology of prokaryotes, which are difficult to obtain from MAG-based analyses. Because the genome data set was constructed from isolated strains publicly available from culture collections, genome-driven hypotheses from our analyses can be verified by cultivation experiments with the isolates.

Contamination-level assessment

Although some genomes showed high values of contamination level, many of these genomes were completed. For instance, the determined genome of “*Desulfosporomusa polytropia*” JCM 32836^T was fully completed, comprising two circular contigs of 5,636,488 bp and 261,751 bp, but showed contamination values of up to 12.1%. Notably, the strain heterogeneity value was zero, indicating no actual contamination. Therefore, we considered that the high values of contamination level observed in this strain and others were not caused by actual contamination but rather by a limitation of the checking ability in the software used in this study. Indeed, no 16S rRNA gene sequences from potential con-

taminants were detected in any of the assemblies. Besides, some complete genomes showed low values of completeness level, for example, <95% for *Oligosphaera ethanolica* JCM 17152^T, which can also be attributed to the inadequacy of the software. Thus, our results highlight the limitations of the widely used software for assessment of genome completeness/contamination and also emphasize the necessity of more complete genome data set for more accurate estimation.

Detection of extrachromosomal DNA fragments

The genome sequencing of the isolated strains indicated the presence of extrachromosomal DNA fragments, such as secondary replicons (diCenzo and Finan 2017) including plasmids and chromids, or phages/viruses, in addition to primary chromosomes. Metagenomics of complex microbial communities in natural environments could detect such extrachromosomal DNA fragments but often fail to identify their host organisms. In this study, we detected two or more contigs in the 76 complete genomes, suggesting the presence of extrachromosomal DNA fragments. Furthermore, the smaller circular contigs than each of the longest contigs found in 40 draft genomes could also represent extrachromosomal DNA fragments. Of the putative circular extrachromosomal DNAs, 35 were not identified as plasmids or phages/viruses using existing tools, suggesting that these were novel types of secondary replicons or phages/viruses. The results will contribute to a greater understanding of interactions between extrachromosomal DNAs and their prokaryotic hosts.

To illustrate the importance of detection of extrachromosomal DNA fragments in identified host organisms, we focus on a gene (locus tag, JCM30955_02990) encoding copper-containing nitrite reductase (NirK), which catalyzes the reduction of nitrite (NO_2^-) to nitric oxide (NO) during denitrification, found in the plasmid of *Halorubrum pallidum* JCM 30955^T. In the GTDB, 146 genomes including six MAGs belonging to the genus *Halorubrum* (“g_Halorubrum” in the GTDB taxonomy) were available. Consistent with the patchy distribution of *nirK* within this genus (Miralles-Robledillo et al. 2021), the *nirK* homologs were found in 25 draft genomes of isolates but were not found in any of the six MAGs. Notably, in the recently reported complete genome (obtained from the NCBI Genome [https://www.ncbi.nlm.nih.gov/

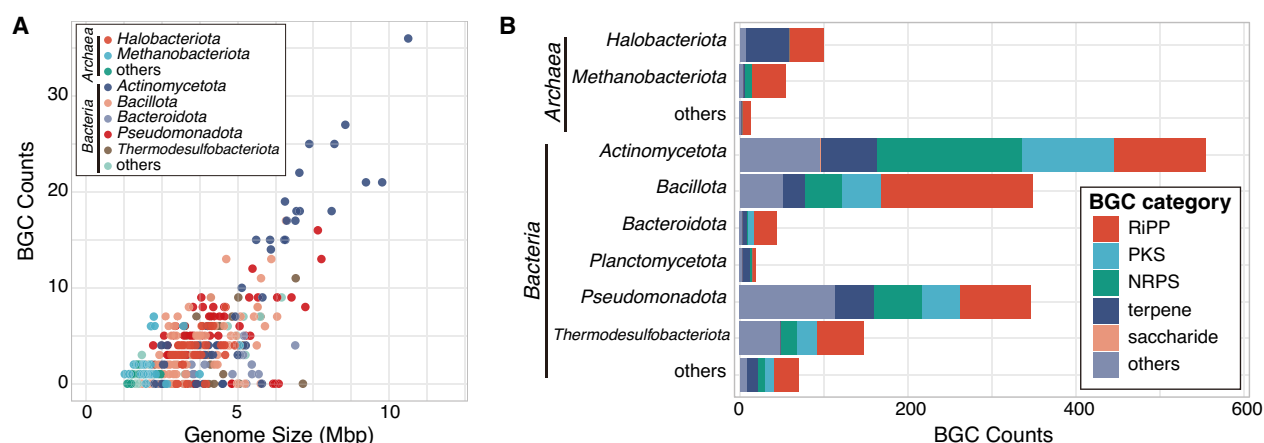


Figure 4. Biosynthetic gene clusters (BGCs) encoded in the genomes. (A) Plot showing the relationship between genome size and the number of BGCs. (B) Bar chart showing the number of BGCs for each category. (RiPP) Ribosomally synthesized and post-translationally modified peptide; (PKS) polyketide synthase; (NRPS) nonribosomal peptide synthetase.

home/genomes/] database under accession number GCF_051122765.1) of *Halorubrum* sp. AS12 (Runzheimer et al. 2025), a *nirK* homolog was found only in its plasmid, like the case of *H. pallidum* JCM 30955. These results suggest that plasmids play a role in the distribution of *nirK* among *Halorubrum* spp. and contribute to nitrogen cycling.

Recovery of 16S rRNA genes

Another limitation during MAG construction by short-read metagenomics is the difficulty in recovering multiple, nearly identical gene copies, such as 16S rRNA genes. For example, the determined complete genome of *Lactivibrio alcoholicus* JCM 17151^T contains three copies of the 16S rRNA gene. Based on the GTDB-based taxonomic classification, the genus *Lactivibrio* is corresponding to “g_CALF_XN01,” which includes three MAGs (obtained from NCBI Genome database under accession numbers GCA_937957815.1, GCA_937870605.1, and GCA_036788195.1), all with genome completeness values exceeding 90%. Notably, none of these MAGs contain 16S rRNA genes, likely owing to the difficulty of recovering these genes during MAG construction. Because 16S rRNA genes serve as essential taxonomic markers for PCR amplicon-based analyses in microbial ecology, the recovery of 16S rRNA genes from the clades previously lacking them highlights the significance of our data set.

Potential proposal of novel taxa and reclassification

Despite being much smaller number than the total of publicly available genomes, our data set covers a broad range of prokaryotic taxa and contributes to accurate taxonomic classification. Of the 166 type-strain genomes, the genomes of 78 strains, for example, *Hydrogenobaculum acidophilum* JCM 8795^T, showed low similarity (<95% average nucleotide identity [ANI] or <70% digital DNA–DNA hybridization [dDDH]), commonly used thresholds for species-level definition) (Konstantinidis and Tiedje 2005; Goris et al. 2007) to any of publicly available genomes derived from isolates or MAGs. This indicates that these genomes are the first to be publicly released at the species level. Other 17 genomes showed high similarity (>95% ANI) to previously reported MAGs or genomes of as-yet-unnamed isolates. For instance, the genome of *Methanothermobacter crinale* JCM 17393^T showed high similarity (99.6% ANI) to the MAG 41_258 (obtained from NCBI Genome database under accession number GCA_001507955.1) recovered from an oil reservoir (Hu et al. 2016). It should be noted that some of the counts of genomes reported in the present study may be subject to change owing to ongoing updates of public databases.

Of the 166 newly released type-strain genomes, other 21 genomes were the first type-strain genomes to be publicly released at the genus level, based on the GTDB-based taxonomic classification. For instance, the genome of *Sulfophobococcus zilligii* JCM 10309^T did not match any genomes of isolates but showed high similarity (98.7% ANI) to a MAG (UBA285, obtained from NCBI Genome database under accession number GCA_002495025.1), which was assigned to the genus-level clade “g_UBA285” within the family-level clade “f_Desulfurococcaceae.” This result indicates that the genus-level clade “g_UBA285” corresponds to the genus *Sulfophobococcus*. Similarly, we released the first type-strain genomes for other 20 validly named genera, including *Stetteria* and *Thermotomaculum* in the domain *Archaea*, as well as *Vulcanithermus* and *Thioreductor* in the domain *Bacteria*. Moreover, some genomes of the 21 type strains had a potential for the assignment to novel taxa at

the family or higher level. According to the GTDB-based classification, our genome data set potentially encompassed one novel class, three novel orders, and four novel families. For instance, *E. thermophila* JCM 14728^T, a member of the phylum *Spirochaetota*, could represent a novel family, order, or even class. Similarly, *Endothiovibrio diazotrophicus* JCM 17961^T belonging to the class *Gammaproteobacteria* could represent a novel family and order. *Thiopropfundum lithotrophicum* JCM 14586^T, belonging to the family *Thiopropfundaceae*, could represent a novel order. Both *Brassicibacter thermophilus* JCM 30480^T and *Thiofractor thiocaminus* JCM 15747^T could represent a novel family. Further phylogenetic analysis is needed to support the assignment to these novel taxa.

The GTDB-based classification suggested that other eight type-strain genomes, currently assigned to already known genera, could represent novel genera. For example, *Methanosarcina baltica* JCM 11281^T was classified as “g_JAQBVP01,” which is distinct from “g_Methanosarcina” and the corresponding genus *Methanosarcina*. This result is consistent with a previous study reporting that *M. baltica* can be physiologically and phylogenetically distinguished from other species in *Methanosarcina* (von Klein et al. 2002). Among species in *Methanosarcina*, the most similar type strain to *M. baltica* JCM 11281^T was *Methanosarcina subterranea* JCM 15540^T, whose genome was determined in this study and classified as “g_Methanosarcina.” The comparisons between these two strains showed the values of 94.9% (16S rRNA gene similarity), 72.9% (amino acid identity [AAI]), and 61.9% (percentage of conserved proteins [POCP]). Given the reported values for genus boundary (90%–99% for 16S rRNA gene similarity, 65%–72% for AAI, and 50%–60% for POCP) (Konstantinidis and Tiedje 2007; Qin et al. 2014; Yarza et al. 2014; Hackmann 2025), *M. baltica* JCM 11281^T is standing on the boundary of the genus threshold. These results imply that *M. baltica* JCM 11281^T belongs to a distinct genus from *Methanosarcina*, although further careful analysis is needed to conclude this notion.

Other 17 type-strain genomes showed high similarity (≥95% ANI or ≥70% dDDH) to those of different species, suggesting that taxonomic reclassification should be considered. For example, the determined genome of *Pyrococcus woesei* JCM 8421^T showed high similarity to that of *Pyrococcus furiosus* DSM 3638^T (obtained from NCBI Genome database under accession number GCA_000007305.1), with ANI and dDDH values of 99.6% and 95.2%, respectively. This result supports the notion that these two strains belong to the same species, namely, *P. furiosus*, which has been previously noted owing to the high similarities on their gene sequences and physiological characteristics (Kanoksilapatham et al. 2004). Similarly, the genome of “*Thermococcus marinus*” JCM 11825^T, a proposed novel species with an as-yet unvalidated name (Jolivet et al. 2004), showed high similarity to that of *Thermococcus eurythermalis* A501^T (obtained from NCBI Genome database under accession number GCA_000769655.1), with ANI and dDDH values of 99.2% and 90.5%, respectively. Our data set includes 14 additional strains that have been proposed to represent novel species but whose scientific names have not yet been validated. Given the strong recommendation for whole-genome sequences in the field of prokaryotic taxonomy since 2018, our data set is expected to facilitate taxonomic reclassification.

Among the newly released 43 genomes of nontype strains, 10 genomes showed low similarity to any publicly available or determined genomes of type strains, indicating that they represented probable novel species. For instance, the genome of *Clostridium* sp. JCM 10519 (=NkU-1) showed a low similarity to its closest relative, *Lacrimispora saccharolytica* (formerly *Clostridium*

saccharolytica) strain WM1^T (obtained from NCBI Genome database under accession number GCA_000144625.1), with ANI and dDDH values of 85.0% and 29.8%, respectively. Indeed, despite a high 16S rRNA gene similarity of 98.9% between the two strains, their physiological characteristics are distinct, as reported previously (Thong-On et al. 2012). Further analyses of *Clostridium* sp. JCM 10519 and the other nine strains are expected to provide additional supporting evidence for describing novel species. Other 13 genomes of unidentified species showed high similarities ($\geq 95\%$ ANI or $\geq 70\%$ dDDH) to those of known species. For instance, the genome of *Streptomyces* sp. JCM 18897 was highly similar to that of *Streptomyces albidoflavus* NRRL B-1271^T with 99.0% ANI and 95.9% dDDH, and thus, *Streptomyces* sp. JCM 18897 could be assigned to the species *S. albidoflavus*.

The remaining 25 genomes of type strains showed high similarities to the genomes of the same species of nontype strains. For instance, the determined genome of *Hydrogenovibrio thermophilus* JCM 12397^T showed a high similarity (97.1% ANI) to the genome of *H. thermophilus* JR-2 (a nontype strain; obtained from NCBI Genome database under accession number GCF_004028275.1) recovered from a deep-sea hydrothermal vent field (Scott et al. 2018). The remaining 20 genomes of nontype strains, for example, *Paenibacillus thiaminolyticus* JCM 7540, showed high similarities to the previously reported genomes of the same species of type or nontype strains.

Actual and predicted OGT

The strains used in this study represent a wide range of OGT, from mesophiles to hyperthermophiles. Based on the determined genome sequences, we assessed the accuracy of bioinformatic tools for OGT prediction. Although generally consistent, the predicted OGTs for several strains deviated from their actual OGT by $>10^{\circ}\text{C}$. For instance, *Desulfuromusa ferrireducens* JCM 12926^T and *Desulfuromonas svalbardensis* JCM 12927^T with the actual OGTs of 14°C – 17°C (Vandieken et al. 2006) were predicted to have OGTs of 25°C – 30°C . Similarly, *Piezobacter thermophilus* JCM 14636^T with the actual OGT of 50°C – 55°C (Takai et al. 2009) was predicted to have OGTs of 35°C – 37°C . These results highlight the limitations of current OGT prediction tools, and our data set provides a valuable resource for the development of more accurate predictive models. Such accurate optimum growth temperature predictions will help to obtain novel isolates from natural environments.

Potential for previously unrecognized metabolism

Genome analysis of isolates not only enables the prediction of previously unrecognized metabolic processes by identifying known genes but also facilitates the discovery of novel pathways underlying experimentally validated metabolisms. Overall, the gene context of the determined genomes is largely consistent with reported metabolic activities characterized through cultivation. However, several predicted metabolic capabilities remain unverified. In this study, we further focused on the potential for carbon fixation and secondary metabolite production, both of which are critical for applications in carbon neutrality and medical and agricultural sciences, directly relevant to human activities. The genome analysis will facilitate experimental validation of the potential for carbon fixation and secondary metabolite production using the corresponding strains.

Potential for carbon fixation

To date, seven pathways are known to be involved in autotrophic carbon fixation (Hügler and Sievert 2011; Sánchez-Andrea et al. 2020), including the CBB cycle, the rTCA cycle, and the WL pathway. Additionally, some autotrophs can fix CO_2 via the roTCA cycle, which does not require key enzymes such as citryl-CoA synthetase (CCS) or ATP-citric lyase (ACL) in the rTCA cycle but is driven only by the enzymes used in the “normal” oxidative TCA cycle (Mall et al. 2018; Nunoura et al. 2018) with a key enzyme of ferredoxin-dependent 2-oxoglutarate synthase (Steffens et al. 2021). As expected, most strains possessing the complete or nearly complete gene set with the key genes for carbon fixation have been reported to be autotrophs. However, we also found the key genes in the genomes of some strains that have not been reported as autotrophs. For example, *Mycolicibacterium crocinum* JCM 16369^T and *Mycobacterium pallens* JCM 16370^T have not been tested for the capability of carbon fixation (Hennessee et al. 2009), whereas their complete genomes possessed the complete gene set for the CBB cycle. Indeed, some autotrophic species in the genus *Mycolicibacterium* have been reported (Gomila et al. 2008), suggesting that the two strains may also be capable of autotrophic growth.

Notably, the key genes for the seven carbon fixation pathways, and even roTCA cycle, were incompletely found in the genomes of the four strains, namely, *Pyrodictium brockii* JCM 9392^T (Stetter et al. 1983), *Desulfonauticus autotrophicus* JCM 13028^T (Mayilraj et al. 2009), *Desulfonatronum parangueonense* JCM 31598^T (Pérez Bernal et al. 2017), and *Salinisphaera hydrothermalis* JCM 15514^T (Crespo-Medina et al. 2009), all of which have been reported as autotrophs. Except for JCM 31598^T, complete genomes of the three strains were determined in this study. Therefore, they potentially fix CO_2 via unknown carbon fixation pathways. In the case of *S. hydrothermalis* JCM 15514^T, the gene for ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), a key enzyme of the CBB cycle, has been detected by PCR cloning-sequencing (Crespo-Medina et al. 2009). However, neither the RuBisCO gene nor other key genes for carbon fixation were found in the determined complete genome sequence or in the previously reported draft genome sequence (obtained from the NCBI Nucleotide database [https://www.ncbi.nlm.nih.gov/nucleotide] under accession number APNE000000000.1). In the complete genome of *P. brockii* JCM 9392^T, we found a gene for RuBisCO form III, which may not be involved in carbon fixation (Sato et al. 2007), and no gene for phosphoribulokinase (PRK), another key enzyme of the CBB cycle. For the above four strains, further analyses including re-evaluation of their autotrophy will be needed to reveal the presence of novel carbon fixation pathways.

Potential for secondary metabolite production

Bioactive secondary metabolites, including antibiotics, are crucial targets for applications in medicine and agriculture, as well as for understanding interactions among organisms in ecology. The genes responsible for producing secondary metabolites are often encoded in BGCs, which have been found in the genomes of diverse bacteria and archaea (Dinglasan et al. 2025; Udway et al. 2025; Zdouc et al. 2025). We identified nearly 1700 BGCs in the 281 genomes, not only from aerobic bacteria including well-studied *Streptomyces* species of the phylum *Actinomycetota* but also from archaea and anaerobic bacteria, which have been less studied (Scherlach and Hertweck 2021).

More than 20 BGCs per genome were exclusively detected from the phylum *Actinomycetota*, with the highest number (36 BGCs) found in the genome of *S. phaeofaciens* JCM 4814^T, although some *Streptomyces* species have been reported to harbor more than 70 BGCs per genome (Belknap et al. 2020; Seshadri et al. 2022). Notably, 10 or more BGCs per genome were detected from the phyla *Bacillota*, *Pseudomonadota*, and *Thermodesulfobacteriota*, in addition to *Actinomycetota*. In obligate anaerobic bacteria, up to 11 BGCs were found in the genomes of *Desulfoconvexum algidum* JCM 16085^T (*Thermodesulfobacteriota*) and *Clostridium nitrophenolicum* JCM 14030^T (*Bacillota*), whose genomes were first determined and completed in this study. In contrast, archaeal genomes had fewer BGCs, which correlated with their smaller genome sizes. The highest number (eight BGCs) was found in the aerobic halophile *Halorubellus litoreus* JCM 17117^T, followed by seven BGCs in the obligate anaerobic methanogen *M. movens* JCM 15415^T. Although the functions of most secondary metabolites produced by BGCs remain unknown, the determined genomes of the isolated strains, especially those of not-well-studied archaea and anaerobic bacteria, will serve as a valuable foundation for genome-based mining of novel bioactive compounds.

Methods

Microbial strains and DNA extraction

We obtained 203 microbial strains from the JCM at the RIKEN BioResource Research Center (BRC) in Tsukuba, Japan. These strains were cultivated under the growth conditions summarized in Supplemental Table S1. Detailed information on the media used to culture each strain is available on the JCM website (<https://jcm.brc.riken.jp/en/>) and also deposited on the figshare website (<https://doi.org/10.6084/m9.figshare.28756649.v1>). Cells were harvested by centrifugation at 10,000g and stored at –80°C until DNA extraction. Genomic DNAs were extracted from the harvested cells using a DNeasy PowerLyzer microbial kit (Qiagen). Additionally, we obtained pre-extracted genomic DNAs for other 148 microbial strains from the DNA bank of RIKEN BRC (Supplemental Table S1). Thus, in total, we sequenced the genomes of 351 strains. These strains were selected based on the absence of previously published genome sequences or the availability of only draft genome sequence data. The selected strains primarily comprised type strains but also included nontype strains potentially representing novel species or higher taxonomic ranks. The purity and quantity of the DNA samples were assessed using a Qubit 3.0 fluorometer (Thermo Fisher Scientific) with dsDNA quantification assay kits, as well as a Multiskan SkyHigh microplate spectrophotometer (Thermo Fisher Scientific) with a µDrop plate (Thermo Fisher Scientific).

Genome sequencing

We performed short-read sequencing on MiSeq or NextSeq 1000 instruments (Illumina). DNA libraries were constructed from aliquots of the DNA samples using the QIAseq FX DNA library kit (Qiagen) for sequencing on MiSeq and NextSeq instruments and were sequenced with the MiSeq reagent kit version 3 (600 cycles, 300 bp paired-end reads; Illumina) and the NextSeq 1000/2000 P1 XLEAP-SBS reagent kit (600 cycles, 300 bp paired-end reads; Illumina).

For long-read sequencing, we used a MinION device (Oxford Nanopore Technologies [ONT]) and a Revio system (Pacific Biosciences [PacBio]). For ONT sequencing, DNA libraries were constructed using the rapid barcoding kit 24 V14 (SQK-RBK114.24)

and sequenced with R10.4.1 flow cells (FLO-MIN114). The raw data were basecalled using Dorado version 0.7.1 (<https://github.com/nanoporetech/dorado>) with the “sup” model. For PacBio sequencing, we first assessed the quality of the DNA samples using a Femto Pulse system (Agilent Technologies). Subsequently, SMRTbell libraries were prepared using the SMRTbell express template prep kit v2.0. On several libraries when necessary, size selection was performed on the BluePippin system using a 0.75% agarose cassette (Sage Science) with a 5–50 kb high-pass cutoff. The SMRTbell libraries were then bound to the sequencing polymerase enzyme using a Revio polymerase kit. Shotgun genomic DNA sequence data were collected on SMRT cells using HiFi sequencing protocols and a Revio sequencing plate (PacBio). HiFi reads were generated using SMRTLink v13.1 with default parameters and extracted with a quality score >Q30.

Genome assembly

The short reads were filtered using fastp version 0.23.4 (Chen 2023) with the option “-l 200.” The long reads were filtered using Chopper version 0.7.0, implemented in NanoPack2 (De Coster and Rademakers 2023), with the options “-q 8 -l 1000.” In some cases, long-read data exceeding the desired size were downsized to approximately 100× coverage using Filtlong (<https://github.com/rrwick/Filtlong>). The filtered long reads were assembled using Flye version 2.9.3 (Kolmogorov et al. 2019) or the Improved Phased Assembler version 1.8.0 (<https://github.com/PacificBiosciences/pbipa>) equipped with the PacBio Revio system. If closed contigs for putative primary chromosomes were not obtained, hybrid assembly was performed using Unicycler version 0.5.0 (short-read-first hybrid assembling) (Wick et al. 2017) or Hybracter version 0.7.3 (long-read-first hybrid assembling) (Bouras et al. 2024) with the short reads. For cases in which only short reads were available, Unicycler was used for short-read-only assembly. The assembly software used for each strain is listed in Supplemental Table S2. In this study, when the longest contig for the primary chromosome in the assembly was circularized, the genome was treated as a “complete” genome, with the exception of *Streptomyces* species, which may have linear chromosomes. For *Streptomyces* assemblies, the presence of terminal inverted repeats detected by BLASTN version 2.14.0 (Altschul et al. 1997) was used to assess whether the linear genomes were complete.

Genome analysis

The genome assemblies were then annotated using DFAST version 1.3.2 (Tanizawa et al. 2018) with the option “--minimum_length 500 --use_prodigal --use_trnascan.” For additional annotation, we used METABOLIC version 4.0 (Zhou et al. 2022) equipped with KEGG (Ogata et al. 1999), HydDB (Søndergaard et al. 2016), dbCAN2 (Zhang et al. 2018) and MEROPS (Rawlings et al. 2018), eggNOG-mapper version 2.1.12 (Cantalapiedra et al. 2021), and gapseq version 1.3.1 (Zimmermann et al. 2021). BGCs were detected using antiSMASH version 7.1.0 (Blin et al. 2023). Plasmids and viruses/phages were identified using geNomad version 1.8.1 (Camargo et al. 2024), mob-suite version 3.1.9 (Robertson and Nash 2018), and VirSorter version 2.2.4 (Guo et al. 2021).

The quality of the determined genomes was evaluated using CheckM version 1.2.3 (Parks et al. 2015) and CheckM2 version 1.0.2 (Chklovski et al. 2023). Taxonomic classification of strains based on the determined genomes was performed using GTDB-tk version 2.4.0 (Chaumeil et al. 2022) with the reference database R226 (Parks et al. 2022). ANI values among the determined genomes were calculated using FastANI version 1.34 (Jain et al. 2018). Average AAI values among the determined genomes were

calculated using EzAAI version 1.2.3 (Kim et al. 2021). Values of the POCF were calculated using POCF-nf version 2.3.4 (Hölzer 2024). dDDH values were determined using the Genome-to-Genome Distance Calculator (GGDC) version 3.0 (Meier-Kolthoff et al. 2022). We used the Type Strain Genome Server (Meier-Kolthoff et al. 2022) and GTDB-tk to detect the closest publicly available genomes to determined ones, based on dDDH and ANI values, respectively. The OGT of each strain was predicted using Tome version 1.0.0 (Li et al. 2019) and OGTFinder version 0.0.2 (Colette et al. 2025). Alluvial diagrams were generated using RAWGraphs (<https://www.rawgraphs.io>). Plots and bar charts were generated using ggplot2 (<https://ggplot2.tidyverse.org>).

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (<https://www.ncbi.nlm.nih.gov/bioproject/>) under accession numbers PRJDB20344, PRJDB20346, PRJDB20693, and PRJDB20694. The Supplemental Materials are also available at Figshare (<https://doi.org/10.6084/m9.figshare.28756649>).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Hiromi Omokawa, Nahomi Noda, Kai Zhang, Naomi Sakurai, Nagisa Sato, Michiru Shimizu, and Koji Suzu for their technical assistance. We also thank the anonymous reviewers for their helpful and constructive comments and suggestions. This work was supported by the value addition subprogram of National BioResource Project (NBRP) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, and partially by Japan Science and Technology Agency (JST), Green Technologies of Excellence (GteX) Program, Biomufacturing Area (JPMJGX23B0, JPMJGX23B2), RIKEN TRIP initiative fieldomics, and Japan Society for the Promotion of Science KAKENHI grant number JP24K00747. The microbial strains and genomic DNAs used in this study were provided by the RIKEN BioResource Research Center through NBRP.

Author contributions: S.K. was responsible for conceptualization, formal analysis, methodology, validation, writing the original draft, and reviewing and editing. S.M., A.S., T.I., and M.S. were responsible for formal analysis, methodology, and reviewing and editing. K.S. and M.O. were responsible for conceptualization, validation, and reviewing and editing.

References

Altschul SE, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402. doi:10.1093/nar/25.17.3389

Belknap KC, Park CJ, Barth BM, Andam CP. 2020. Genome mining of bio-synthetic and chemotherapeutic gene clusters in *Streptomyces* bacteria. *Sci Rep* **10**: 2003. doi:10.1038/s41598-020-58904-9

Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, Fetter A, Terlouw BR, Metcalf WW, Helfrich EJJ, et al. 2023. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res* **51**: W46–W50. doi:10.1093/nar/gkad344

Bouras G, Houtak G, Wick RR, Mallawaarachchi V, Roach MJ, Papudeshi B, Judd LM, Sheppard AE, Edwards RA, Vreugde S. 2024. Hybracter: enabling scalable, automated, complete and accurate bacterial genome assemblies. *Microb Genom* **10**: 001244. doi:10.1099/mgen.0.001244

Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PSG, Nayfach S, Kyrpides NC. 2024. Identification of mobile genetic elements with geNomad. *Nat Biotechnol* **42**: 1303–1312. doi:10.1038/s41587-023-01953-y

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* **38**: 5825–5829. doi:10.1093/molbev/msab293

Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2022. GTDB-Tk v2: memory friendly classification with the Genome Taxonomy Database. *Bioinformatics* **38**: 5315–5316. doi:10.1093/bioinformatics/btac672

Chen S. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**: e107. doi:10.1002/imt2.107

Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. 2023. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* **20**: 1203–1212. doi:10.1038/s41592-023-01940-w

Chun J, Oren A, Ventosa A, Christensen H, Arahall DR, da Costa MS, Rooney AP, Yi H, Xu XW, De Meyer S, et al. 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* **68**: 461–466. doi:10.1099/ijsem.0.002516

Colette S, François J, De Moor B, van Noort V. 2025. Machine learning for optimal growth temperature prediction of prokaryotes using amino acid descriptors. *bioRxiv* doi:10.1101/2025.03.03.640802

Crespo-Medina M, Chatziefthimiou A, Cruz-Matos R, Perez-Rodriguez I, Barkay T, Lutz RA, Starovoytov V, Vetriani C. 2009. *Salinisphaera hydrothermalis* sp. nov., a mesophilic, halotolerant, facultatively autotrophic, thiosulfate-oxidizing gammaproteobacterium from deep-sea hydrothermal vents, and emended description of the genus *Salinisphaera*. *Int J Syst Evol Microbiol* **59**: 1497–1503. doi:10.1099/ijms.0.005058-0

De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**: btad311. doi:10.1093/bioinformatics/btad311

diCenzo GC, Finan TM. 2017. The divided bacterial genome: structure, function, and evolution. *Microbiol Mol Biol Rev* **81**: e00019-00017. doi:10.1128/MMBR.00019-17

Dicks J, Fazal MA, Oliver K, Grayson NE, Turnbull JD, Bane E, Burnett E, Deheer-Graham A, Holroyd N, Kaushal D, et al. 2023. NCTC3000: a century of bacterial strain collecting leads to a rich genomic data resource. *Microb Genom* **9**: mgen000976. doi:10.1099/mgen.0.000976

Dinglasan JLN, Otani H, Doering DT, Udwardy D, Mouncey NJ. 2025. Microbial secondary metabolites: advancements to accelerate discovery towards application. *Nat Rev Microbiol* **23**: 338–354. doi:10.1038/s41579-024-01141-y

Gomila M, Ramirez A, Gasco J, Lalucat J. 2008. *Mycobacterium llatzerense* sp. nov., a facultatively autotrophic, hydrogen-oxidizing bacterium isolated from haemodialysis water. *Int J Syst Evol Microbiol* **58**: 2769–2773. doi:10.1099/ijms.0.65857-0

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**: 81–91. doi:10.1099/ijms.0.64483-0

Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, Pratama AA, Gazitua MC, Vik D, Sullivan MB, et al. 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**: 37. doi:10.1186/s40168-020-00990-y

Hackmann TJ. 2025. Setting new boundaries of 16S rRNA gene identity for prokaryotic taxonomy. *Int J Syst Evol Microbiol* **75**: 006747. doi:10.1099/ijsem.0.006747

Hennessee CT, Seo JS, Alvarez AM, Li QX. 2009. Polycyclic aromatic hydrocarbon-degrading species isolated from Hawaiian soils: *Mycobacterium crocinum* sp. nov., *Mycobacterium pallens* sp. nov., *Mycobacterium rutilum* sp. nov., *Mycobacterium rufum* sp. nov. and *Mycobacterium aromaticivorans* sp. nov. *Int J Syst Evol Microbiol* **59**: 378–387. doi:10.1099/ijms.0.65827-0

Hölzer M. 2024. POCF-nf: an automatic Nextflow pipeline for calculating the percentage of conserved proteins in bacterial taxonomy. *Bioinformatics* **40**: btae175. doi:10.1093/bioinformatics/btae175

Hu P, Tom L, Singh A, Thomas BC, Baker BJ, Piceno YM, Andersen GL, Banfield JF. 2016. Genome-resolved metagenomic analysis reveals roles for candidate phyla and other microbial community members in biogeochemical transformations in oil reservoirs. *mBio* **7**: e01669-15. doi:10.1128/mBio.01669-15

Hügler M, Sievert SM. 2011. Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. *Ann Rev Mar Sci* **3**: 261–289. doi:10.1146/annurev-marine-120709-142712

Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**: 5114. doi:10.1038/s41467-018-07641-9

- Jolivet E, Corre E, L'Haridon S, Forterre P, Prieur D. 2004. *Thermococcus marinus* sp. nov. and *Thermococcus radiotolerans* sp. nov., two hyperthermophilic archaea from deep-sea hydrothermal vents that resist ionizing radiation. *Extremophiles* **8**: 219–227. doi:10.1007/s00792-004-0380-9
- Kanoksilapatham W, González JM, Maeder DL, DiRuggiero J, Robb FT. 2004. A proposal to rename the hyperthermophile *Pyrococcus woesei* as *Pyrococcus furiosus* subsp. *woesei*. *Archaea* **1**: 277–283. doi:10.1155/2004/513563
- Kim D, Park S, Chun J. 2021. Introducing EzAAI: a pipeline for high throughput calculations of prokaryotic average amino acid identity. *J Microbiol* **59**: 476–480. doi:10.1007/s12275-021-1154-0
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**: 540–546. doi:10.1038/s41587-019-0072-8
- Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci* **102**: 2567–2572. doi:10.1073/pnas.0409727102
- Konstantinidis KT, Tiedje JM. 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* **10**: 504–509. doi:10.1016/j.mib.2007.08.006
- Li G, Rabe KS, Nielsen J, Engqvist MKM. 2019. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synth Biol* **8**: 1411–1420. doi:10.1021/acssynbio.9b00099
- Mall A, Sobotta J, Huber C, Tschirner C, Kowarschik S, Bačnik K, Mergelsberg M, Boll M, Hügler M, Eisenreich W, et al. 2018. Reversibility of citrate synthase allows autotrophic growth of a thermophilic bacterium. *Science* **359**: 563–567. doi:10.1126/science.aao2410
- Mayilraj S, Kaksonen AH, Cord-Ruwisch R, Schumann P, Spröer C, Tindall BJ, Spring S. 2009. *Desulfonauticus autotrophicus* sp. nov., a novel thermophilic sulfate-reducing bacterium isolated from oil-production water and emended description of the genus *Desulfonauticus*. *Extremophiles* **13**: 247–255. doi:10.1007/s00792-008-0212-4
- Meier-Kolthoff JP, Carbasse JS, Peinado-Olarte RL, Göker M. 2022. *Nucleic Acids Res* **50**: D801–D807. doi:10.1093/nar/gkab902
- Miralles-Robledillo JM, Bernabeu E, Giani M, Martínez-Serna E, Martínez-Espinosa RM, Pire C. 2021. Distribution of denitrification among *Haloarchaea*: a comprehensive study. *Microorganisms* **9**: 1669. doi:10.3390/microorganisms9081669
- Mukherjee S, Seshadri R, Varghese NJ, Eloë-Fadrosch EA, Meier-Kolthoff JP, Göker M, Coates RC, Hadjithomas M, Pavlopoulos GA, Paez-Espino D, et al. 2017. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat Biotechnol* **35**: 676–683. doi:10.1038/nbt.3886
- Nguyen SV, Puthuvelil NP, Petrone JR, Kirkland JL, Gaffney K, Tabron CL, Wax N, Duncan J, King S, Marlow R, et al. 2024. The ATCC genome portal: 3,938 authenticated microbial reference genomes. *Microbiol Resour Annu* **13**: e0104523. doi:10.1128/mra.01045-23
- Nunoura T, Chikaraishi Y, Izaki R, Suwa T, Sato T, Harada T, Mori K, Kato Y, Miyazaki M, Shimamura S, et al. 2018. A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile. *Science* **359**: 559–563. doi:10.1126/science.aao3407
- Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**: 29–34. doi:10.1093/nar/27.1.29
- Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043–1055. doi:10.1101/gr.186072.114
- Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* **50**: D785–D794. doi:10.1093/nar/gkab776
- Pérez Bernal MF, Souza Brito EM, Bartoli M, Aubé J, Fardeau ML, Cuevas Rodriguez G, Ollivier B, Guyonaud R, Hirschler-Réa A. 2017. *Desulfonatotrom paranguonense* sp. nov., a sulfate-reducing bacterium isolated from sediment of an alkaline crater lake. *Int J Syst Evol Microbiol* **67**: 4999–5005. doi:10.1099/ijsem.0.002399
- Qin QL, Xie BB, Zhang XY, Chen XL, Zhou BC, Zhou J, Oren A, Zhang YZ. 2014. A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol* **196**: 2210–2215. doi:10.1128/JB.01688-14
- Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. 2018. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res* **46**: D624–D632. doi:10.1093/nar/gkx1134
- Riesco R, Trujillo ME. 2024. Update on the proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* **74**: 006300. doi:10.1099/ijsem.0.006300
- Robertson J, Nash JHE. 2018. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* **4**: e000206. doi:10.1099/mgen.0.000206
- Runzheimer K, Schwab L, Engel D, Schaudinn C, Laue M, Rebrosóvá K, Beblo-Vranesevic K, Azizah M, Leuko S. 2025. Life in the brine of Lunenburg, Germany: unveiling microorganisms associated with Zechstein salt deposits. *Front Microbiol* **16**: 1625916. doi:10.3389/fmicb.2025.1625916
- Sánchez-Andrea I, Guedes IA, Hornung B, Boeren S, Lawson CE, Sousa DZ, Bar-Even A, Claessens NJ, Stams AJM. 2020. The reductive glycine pathway allows autotrophic growth of *Desulfovibrio desulfuricans*. *Nat Commun* **11**: 5090. doi:10.1038/s41467-020-18906-7
- Sato T, Atomi H, Imanaka T. 2007. Archaeal type III RuBisCOs function in a pathway for AMP metabolism. *Science* **315**: 1003–1006. doi:10.1126/science.1135999
- Scherlach K, Hertweck C. 2021. Mining and unearthing hidden biosynthetic potential. *Nat Commun* **12**: 3864. doi:10.1038/s41467-021-24133-5
- Scott KM, Williams J, Porter CMB, Russel S, Harmer TL, Paul JH, Antonen KM, Bridges MK, Camper GJ, Campa CK, et al. 2018. Genomes of ubiquitous marine and hypersaline *Hydrogenovibrio*, *Thiomicrothrix* and *Thiomicrospira* spp. encode a diversity of mechanisms to sustain chemolithoautotrophy in heterogeneous environments. *Environ Microbiol* **20**: 2686–2708. doi:10.1111/1462-2920.14090
- Seshadri R, Roux S, Huber KJ, Wu D, Yu S, Udwardy D, Call L, Nayfach S, Hahnke RL, Pukall R, et al. 2022. Expanding the genomic encyclopedia of *Actinobacteria* with 824 isolate reference genomes. *Cell Genom* **2**: 100213. doi:10.1016/j.xgen.2022.100213
- Søndergaard D, Pedersen CN, Greening C. 2016. HydDB: a web tool for hydrogenase classification and analysis. *Sci Rep* **6**: 34212. doi:10.1038/srep34212
- Steffens L, Pettinato E, Steiner TM, Mall A, König S, Eisenreich W, Berg IA. 2021. High CO₂ levels drive the TCA cycle backwards towards autotrophy. *Nature* **592**: 784–788. doi:10.1038/s41586-021-03456-9
- Stetter KO, König H, Stackebrandt E. 1983. *Pyrodicticum* gen. nov., a new genus of submarine disc-shaped sulphur reducing archaeobacteria growing optimally at 105°C. *Syst Appl Microbiol* **4**: 535–551. doi:10.1016/S0723-2020(83)80011-3
- Takai K, Miyazaki M, Hirayama H, Nakagawa S, Querellou J, Godfroy A. 2009. Isolation and physiological characterization of two novel, piezophilic, thermophilic chemolithoautotrophs from a deep-sea hydrothermal vent chimney. *Environ Microbiol* **11**: 1983–1997. doi:10.1111/j.1462-2920.2009.01921.x
- Tanizawa Y, Fujisawa T, Nakamura Y. 2018. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics* **34**: 1037–1039. doi:10.1093/bioinformatics/btx713
- Thong-On A, Suzuki K, Noda S, Inoue J, Kajiwara S, Ohkuma M. 2012. Isolation and characterization of anaerobic bacteria for symbiotic recycling of uric acid nitrogen in the gut of various termites. *Microbes Environ* **27**: 186–192. doi:10.1264/jsmet.2.ME11325
- Udwardy DW, Doering DT, Foster B, Smirnova T, Kautsar SA, Mouncey NJ. 2025. The secondary metabolism collaborator: a database and web discussion portal for secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* **53**: D717–D723. doi:10.1093/nar/gkai1060
- Vandieken V, Mußmann M, Niemann H, Jørgensen BB. 2006. *Desulfuromonas svalbardensis* sp. nov. and *Desulfuromonas ferrireducens* sp. nov., psychrophilic, Fe(III)-reducing bacteria isolated from Arctic sediments, Svalbard. *Int J Syst Evol Microbiol* **56**: 1133–1139. doi:10.1099/ijfs.0.63639-0
- von Klein D, Arab H, Völker H, Thomm M. 2002. *Methanosarcina baltica*, sp. nov., a novel methanogen isolated from the Gotland Deep of the Baltic Sea. *Extremophiles* **6**: 103–110. doi:10.1007/s007920100234
- Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**: e1005595. doi:10.1371/journal.pcbi.1005595
- Wu L, McCluskey K, Desmeth P, Liu S, Hideaki S, Yin Y, Moriyo O, Itoh T, Kim CY, Lee JS, et al. 2018. The global catalogue of microorganisms 10K type strain sequencing project: closing the genomic gaps for the widely published prokaryotic and fungi species. *GigaScience* **7**: giy026. doi:10.1093/gigascience/giy026
- Yarza P, Yilmaz P, Priesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12**: 635–645. doi:10.1038/nrmicro3330
- Zdouc MM, Blin K, Louwen NLL, Navarro J, Loureiro C, Bader CD, Bailey CB, Barra L, Booth TJ, Bozhuyuk KAJ, et al. 2025. MiBig 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res* **53**: D678–D690. doi:10.1093/nar/gkai1115
- Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: a meta server for automated carbohydrate-active

- enzyme annotation. *Nucleic Acids Res* **46**: W95–W101. doi:10.1093/nar/gky418
- Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, Karaoz U, Anantharaman K. 2022. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* **10**: 33. doi:10.1186/s40168-021-01213-8
- Zimmermann J, Kaleta C, Waschina S. 2021. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol* **22**: 81. doi:10.1186/s13059-021-02295-1

Received June 10, 2025; accepted in revised form January 27, 2026.