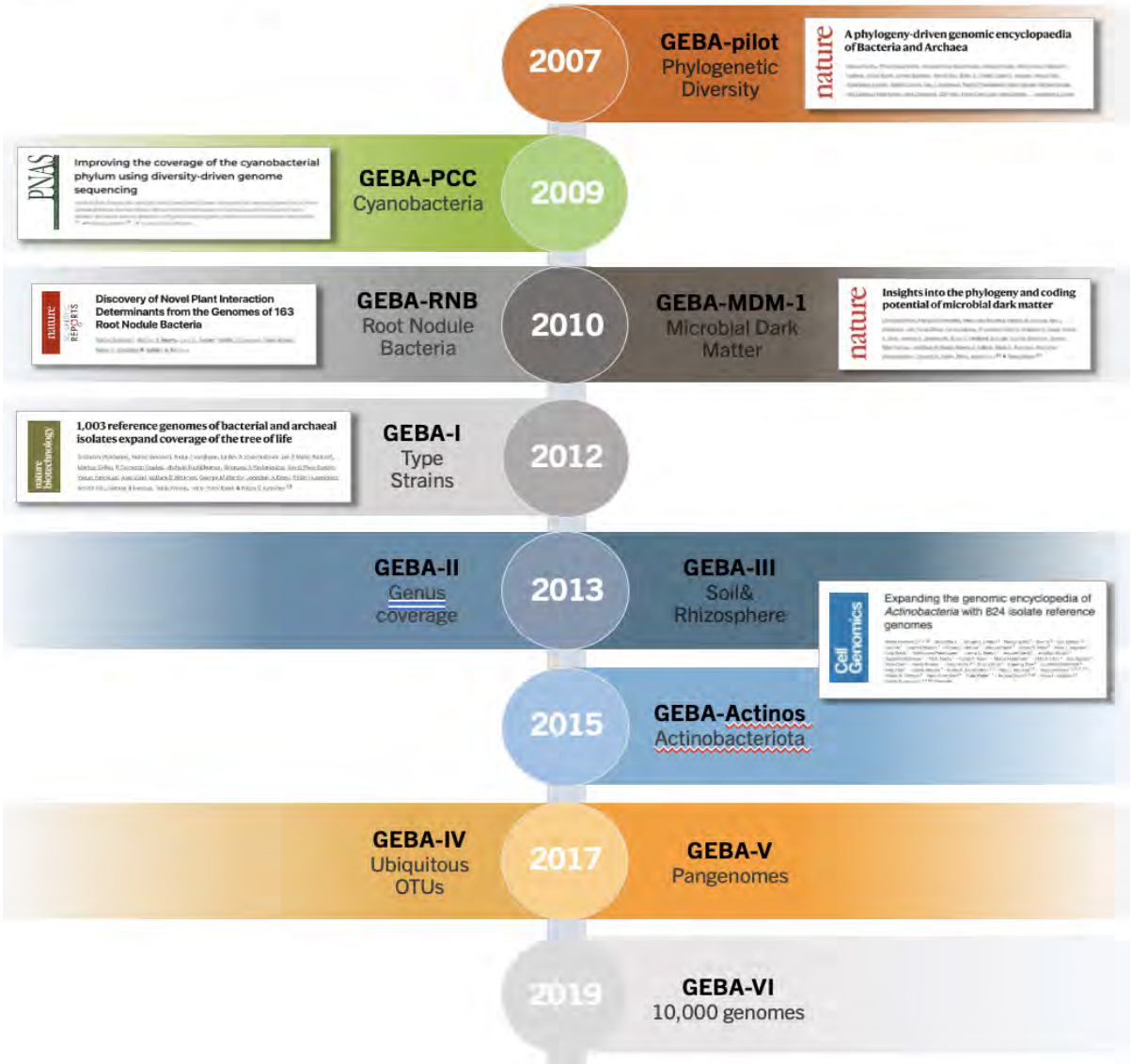
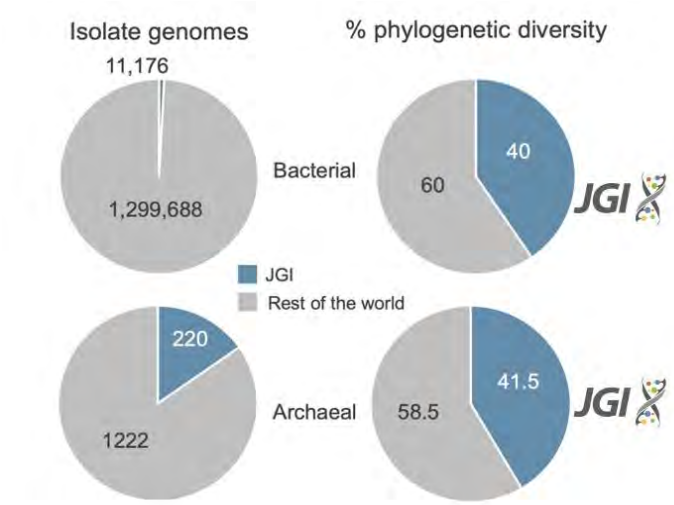
JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

A lifelong commitment to expanding genomic diversity



- ***Catalogs of Cultivated Bacteria & Archaea***
- ***Genomic Encyclopedia of Bacteria and Archaea (GEBA)***
- ***Beyond GEBA: Rumen and Rhizosphere Collections***
- ***Beyond Cultivated Taxa***

Catalogs of Cultivated *Bacteria* & *Archaea*



PD estimated by summing branch lengths of universally conserved single copy marker genes encoded by over 1.3 million isolate genomes in GenBank (Wu et al).

The biodiversity within the bacterial and archaeal domains is immense. These microbes play critical roles in sustaining the biosphere and supporting the health of our planet and ourselves. Whole genome sequences are essential tools for studying these organisms, revolutionizing our understanding of their roles in processes including energy production, bioremediation, global nutrient cycles, as well as the origins, evolution, and diversity of life. Over the past 20 years, the JGI has generated numerous catalogs of cultivated *Bacteria* and *Archaea*, with some of these efforts highlighted below.

Isolate sequencing has been democratized globally and while the JGI contributes a minority of isolate genomes to public databases, the diversity of these isolates accounts for approximately 40% of the total phylogenetic diversity (PD) represented by over 1.3 million genomes in GenBank.

Genomic Encyclopedia of *Bacteria* and *Archaea* (GEBA) - responding to the bias in phylogenetic diversity of sequenced genomes

The GEBA projects were initiated to address a recognized bias in early genome sequencing efforts, disproportionately focused on a limited range of taxonomic groups, primarily those with

clinical or biotechnological relevance. In neglecting the vast diversity of known prokaryotic organisms, this bias hindered our understanding of both microbial genome complexity, as well as of evolution, physiology, and metabolic capacity of microbes.

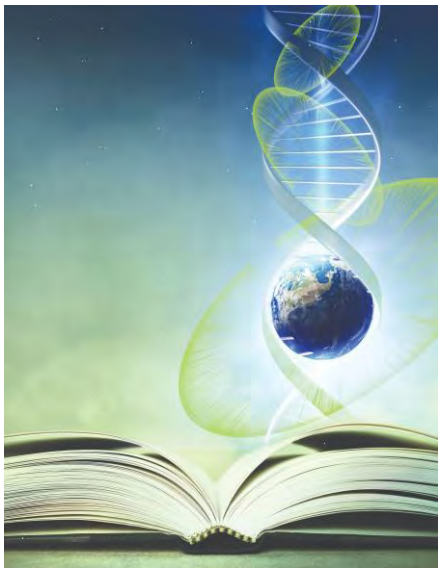
Early GEBA projects (Pilot, GEBA-I, GEBA-II) focused on sequencing thousands of type strains of high phylogenetic diversity (e.g., new phyla and genera) relative to existing microbial genome data. These efforts represented first steps toward achieving a phylogenetically balanced representation of microbial sequence space. Subsequent projects targeted isolates from environments relevant to DOE mission areas, such as soil and the rhizosphere (GEBA-III), nitrogen-fixing root nodules (GEBA-RNB), -or novel taxonomic groups with biofuel or other biotechnological potential (CyanoGEBA and GEBA-Actinobacteria). Thousands of reference genomes of isolates have since been generated and published in various compendia articles , with many more still in progress.

Type strains in ampules such as these were used to generate reference microbial genomes. (Photo Courtesy of the DSMZ. Design credit: Zosia Rostomian, Berkeley Lab Creative Services.)

The GEBA collections mainly consist of type strains obtained from the Leibniz Institute DSMZ and other international culture collections. These type strains, which serve as permanent reference points for bacterial and archaeal species classification, are meticulously maintained and well-characterized by phenotype, isolation source, and other attributes. A comprehensive repository of high-quality reference genomes from these strains, combined with their existing biochemical and genetic data, offers a robust foundation for a wide array of experiments and investigations.

Results from various GEBA projects included:

- Cataloging bacterial and archaeal diversity
- Unraveling novel functions and biosynthetic gene clusters (natural products)
- Expanding sequence diversity of protein families and ortholog groups
- Linking or correlating phenotypes to genotypes
- Improving phylogenetic anchoring of metagenomic data
- Helping identify and develop new microbial model systems
- Expanding our understanding of evolutionary history and diversification of species (and underlying mechanisms)



JGI's Genomic Encyclopedia of Bacteria and Archaea (GEBA) initiative. (Zosia Rostomian, Berkeley Lab Creative Services.)

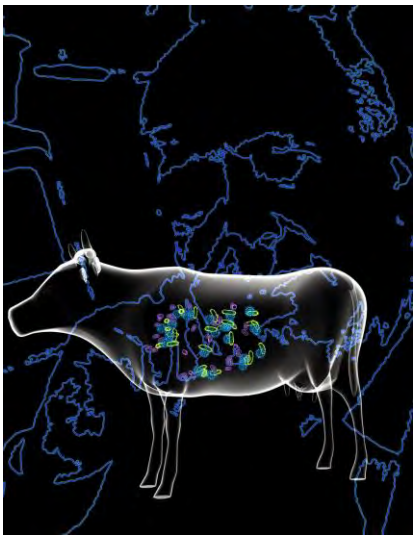
Nominate your strains for sequencing at the Joint Genome Institute

Do you have unsequenced bacterial or archaeal **isolates** and an interesting scientific question addressable via genomics? Are the questions relevant to the DOE mission? If YES, please nominate your isolates for sequencing at JGI. **Please include details** like strain count, environmental origin, taxonomic breadth, etc.

Beyond GEBA: Rumen and Rhizosphere Collections

Beyond GEBA, the JGI is committed to helping develop reference genome catalogs from select environments with DOE mission relevance such as:

In 2018, through a JGI Community Science Program, researchers put together a collection of 410 reference genomes from the rumen microbiome, a leading emitter of anthropogenic methane. To date it is the single largest effort to provide a cataloged and curated culture- and genome sequence resource of rumen microbes. The project, dubbed "Hungate1000," relied involved 54 researchers from 14 organizations across nine countries. A saturated genome collection is fundamental to developing any strategies to modify and harness the rumen microbiome toward addressing food security and climate resilience concerns. The Hungate1000 is a continuing effort in partnership with the Rumen Gateway Project. Efficient enzymes encoded by rumen microbes for degrading plant biomass is of significant additional interest for the efficient engineering of biofuels. Learn more in this Genome Insider podcast.



Depiction of rumen microbiome (Ella Maru Studio, www.scientific-illustrations.com)

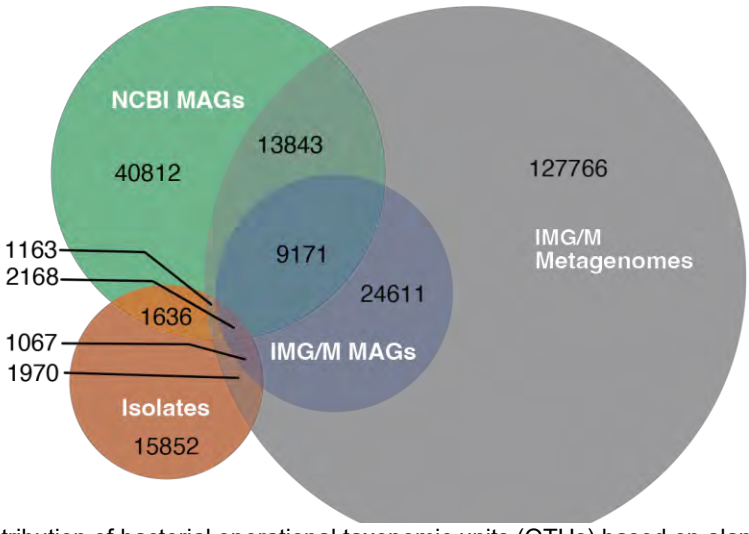
Another target environment was the plant-associated microbiome. In collaboration with researchers from the University of North Carolina at Chapel Hill, Oak Ridge National Laboratory, the University of Washington, and the Max Planck Institute, 377 novel bacteria were isolated and sequenced from the rhizosphere of *Brassicaceae*, poplar trees, and maize. These organisms, along with their enriched host-adaptive functions, provided an unprecedented resource for the rational design of "plant probiotics" to support the growth of bioenergy crops with reduced reliance on chemical inputs such as fertilizers and pesticides.

Beyond Cultivated Taxa: GEBA-MDM (Microbial Dark Matter) illuminates uncultivated candidate phyla using single-cell genomics

The "great plate count anomaly" experiment revealed that over 99% of microbial lineages are uncultivated and, consequently, unstudied. This concept is often illustrated by the metaphor of an iceberg, where the disproportionately larger mass of submerged ice represents the vast unseen microbial diversity. To explore these uncultivated taxa beyond taxonomic surveys, the GEBA-MDM project leveraged the JGI's culture-independent high-throughput single-cell genomics capabilities to recover 200 single-cell genomes (SAGs) from 29 underexplored branches of the tree of life, including candidate novel phyla representatives. These SAGs, along with subsequent culture-independent efforts, have provided unprecedented insights into previously inaccessible microbial "dark matter," greatly enriching our understanding of microbial diversity and metabolic potential while challenging long-held assumptions derived from the study of isolated or domesticated species. For example, we discovered novel amino acid usage for the opal stop codon, an archaeal-type purine synthesis pathway in *Bacteria*, and complete sigma factors in *Archaea* similar to those found in *Bacteria*.

Please see our metagenome program pages for notes on complementary efforts to recover uncultivated metagenome assembled genomes (MAGs) and more.

Charting the path forward with a comprehensive census of microbial genomes



Distribution of bacterial operational taxonomic units (OTUs) based on alanyl-tRNA synthetase (COG0013) gene sequences. The figure highlights shared and unique species-level OTUs across various genome categories, with numbers indicating the counts of spec of species-level OTUs per genome category (Wu et al.). The gray area represents the uncaptured "dark matter" of the microbial world, identified only through marker genes derived from metagenomes.

With the bounty of sequenced genomes from both cultivated and uncultivated sources, we aimed to conduct a comprehensive census of *Bacteria* and *Archaea* and assess the biodiversity represented by all available genome sequences. The study conservatively estimates that over 40% of bacterial and archaeal phylogenetic diversity (PD) is not represented by any genomic sequences. While cultivated isolates are essential for laboratory research, they account for only a small fraction of total

microbial diversity (9.73% in bacteria and 6.55% in archaea). The study highlights several environments with substantial uncaptured diversity, including freshwater, marine subsurface, sediment, and soil, and identifies prominent uncaptured taxonomic groups within these habitats.

Using the results of this analysis as a roadmap, we advocate for renewed efforts in cultivation and targeted sequencing of underexplored habitats that harbor novel taxa. Achieving the ultimate goal of a saturated reference genome database and culture collection will require coordinated collaboration, resource sharing, and expertise across research institutions to develop a comprehensive understanding of microbial life and its role in the biosphere.

Publications:

- Wu D et.al. A Metagenomic Perspective on the Microbial Prokaryotic Genome Census. Science Advances. 2025 Jan 17; 11,eadq2166.
- Seshadri R, et. al. Expanding the genomic encyclopedia of Actinobacteria with 824 isolate reference genomes. Cell Genom. 2022 Nov 11; 2(12):100213.
- Levy A, Salas Gonzalez I, et. al. Genomic features of bacterial adaptation to plants. Nat Genet. 2017 Dec 18; 50(1):138-150.
- Seshadri R, Leahy SC, et. al. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. Nat Biotechnol. 2018 Apr; 36(4):359-367.
- Mukherjee S, Seshadri R, et. al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat Biotechnol. 2017 Jul; 35(7):676-683.
- Seshadri R et. al. Discovery of Novel Plant Interaction Determinants from the Genomes of 163 Root Nodule Bacteria. Scientific Reports. 2015. 5: 16825.
- Rinke C et. al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013 Jul 25;499(7459):431-7.
- Shih et. al. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. PNAS. 2013 Jan 15;110(3):1053-8.
- Wu D et. al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature. 2009 Dec 24;462(7276):1056-60.