METAGENOMICS

# Finding life's missing pieces

The Uncultivated Bacteria and Archaea dataset is a foundational collection of 7,903 genomes from uncultivated microorganisms. It highlights how microbial diversity is readily recovered using current tools and existing metagenomic datasets to help piece together the tree of life.

Lindsey M. Solden and Kelly C. Wrighton

Microorganisms play critical roles directing Earth's biogeochemical cycles, producing energy to sustain the planet, and maintaining human health[1]. Until recently, insights into the reactions catalysed by these microbial engines were impeded by the fact that genomic content of the microbial world was largely inaccessible. Genome database initiatives, such as the Genomic Encyclopedia of Bacteria and Archaea and the Human Microbiome Project, have helped overcome this barrier by targeting the genome sequencing of underrepresented microorganisms[2]. However, these genome inventories were only conducted on organisms that could be grown in the laboratory, leaving large knowledge gaps that would not be filled until the advent of metagenomics. Using metagenomics, the acquisition of DNA no longer required cultivated organisms, but instead could be obtained directly from the environment. Sequencing of environmental DNA results in short nucleotide reads, which can be assembled into larger fragments and ultimately pieced into genomes[3]. In this issue of *Nature Microbiology*, Parks et al. demonstrate the use of established metagenomic approaches to generate a library of 7,903 metagenome-assembled genomes (MAGs)[4], where 5,726 of these genomes were unique. Consequently, this study has unearthed microbial missing pieces from life's jigsaw puzzle.

Metagenomic studies have generally focused on the recovery of genomes from a single ecosystem, with one publication yielding 2,540 MAGs[5]. Rather than focusing on a single habitat, Parks et al. conducted the first MAG database initiative. This study mined publically available sequencing reads deposited by the scientific community into the National Center for Biotechnology Information (NCBI) archive. This pioneering approach uncovered genomes from thousands of environments, with samples spanning the globe from the deepest underwater hydrothermal vents to handrails on the New York subway. Parks et al. named this collection of genomes the Uncultivated Bacteria and Archaea (UBA) dataset. This study resulted in a 10% increase in the number of genomes currently found in repositories, vastly accelerating representation of the uncultivated microbial world (Fig. 1).

The UBA genome dataset enables analyses that alter current perceptions about the microbial tree of life. These genomes represent new pieces of the puzzle, constituting the first representatives from 20 phylum-level lineages composed exclusively of UBA genomes. Based on the absence of an accepted nomenclature for genomes from uncultivated microorganisms[6], the authors named these new uncultured bacterial or archaeal phyla with alphanumeric identifiers starting with UBP or UAP, respectively. Prior to this study, these phyla were not known. To put the significance of these findings in perspective, this is analogous to discovering the animal phylum that contains mammals, fish, birds and amphibians.

Beyond new branches, the UBA genomes also add new leaves to established branches on the microbial tree of life. For instance, over 75% of the UBA genomes belong to four known phyla (Bacteroidetes, Firmicutes, Proteobacteria and Actinobacteria). Despite the fact that these phyla already constitute the

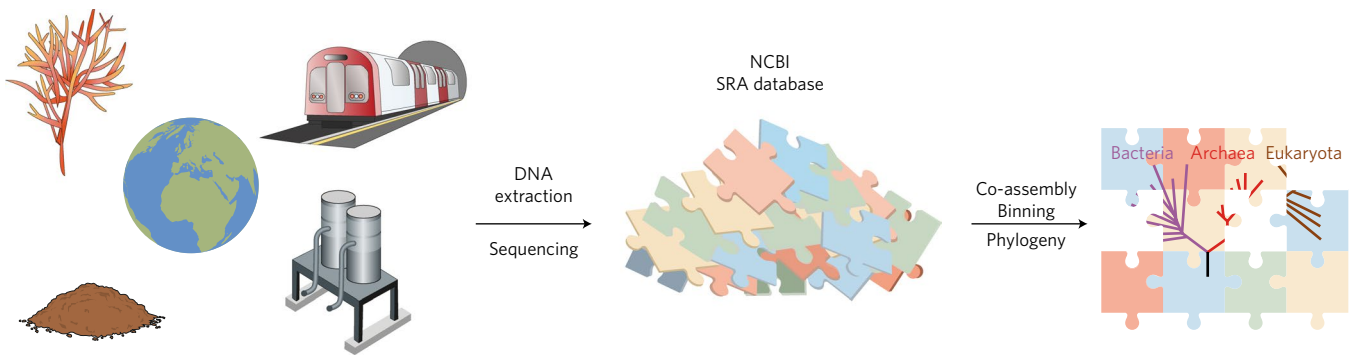


**Fig. 1 | Reconstruction of nearly 8,000 microbial genomes from a range of environments provides previously missing pieces of the tree of life.** To cast the widest global sampling, the authors accessed publically available sequencing reads deposited by the scientific community into the NCBI Sequence Read Archive (SRA). A breadth of non-human ecosystems were sampled using this approach, ranging from coral reefs to wastewater plants, ruminants and soils. Parks et al. used current assembly and binning algorithms to piece these genome fragments or puzzle pieces into thousands of near-complete genomes, providing new insights into the uncultivated microbial world.

majority of existing microbial genomic information[7], the UBA genomes expand the phylogenetic diversity of each phylum by an average of 30%. Together, the UBA genomes demonstrate that current metagenomic methodologies can readily untangle a large fraction of microbial diversity that is only accessible via cultivation-independent approaches.

This study also provides a new perspective on the geometry of the microbial tree of life. Historically, cultivated microorganisms were assigned to specific branches based on their 16S rRNA gene. As it is often difficult to obtain this gene from MAGs[2], 16 or 23 ribosomal proteins are commonly used for phylogenetic placement of genomes from uncultivated microorganisms[7–9]. Parks et al. did not limit their analyses to only ribosomal proteins, but instead used a set of 120 functionally diverse proteins. Their analyses using this protein set decreased the relative diversity of certain phylum-level lineages on the bacterial tree, challenging the current view on the microbial tree of life[9]. Further discussions on the number and types of proteins, as well as the phylogenetic best practice, are necessary to resolve these discrepant tree topologies.

As the field of metagenomics has matured, quality standards for MAGs have become a moving target. Using commonly employed MAG quality assessments[10], Parks et al. determined that 44% of the nearly 8,000 UBA genomes were estimated to be 'near complete' (greater than 90% complete and with less than 5% contamination). This statistic clearly highlights not only the scale of the UBA dataset, but also the value of the genomic content. It is important to note, however, that in the field of metagenomics, different definitions are commonly used for the same quality category (for example, medium quality in Parks et al.)[2], and therefore caution needs to be employed when comparing quality descriptors across publications. Around the time that Parks et al. went to press, the Genomic Standards Consortium (GSC) published the first quality recommendations for MAGs. These recommendations expand upon the completion and contamination metrics used by Parks et al. and prescribe additional quality estimates that incorporate genome assembly measures. Figure 1 of the Article incorporates some of these assembly metrics to provide a more comprehensive analysis of genome quality than is typically performed in most metagenome studies today. Given the increased deposition of MAGs into public databases, there is a need to standardize quality metrics and to establish consistent nomenclature for categorizing genome quality. Fortunately, this publication by Parks et al. and the recently published GSC benchmarks[2] provide a clear roadmap for the path ahead.

The UBA genomes begin to piece together the big picture of the microbial tree of life, while also serving as a valued resource for the scientific community. The authors combined short nucleotide reads into genomic prose, creating a catalogue of uncultivated microorganisms that were concealed within the NCBI Sequence Read Archive. The UBA dataset represents an important reference that provides taxonomic context for recovered genomes from uncultivated lineages. Researchers can also work backwards from the UBA dataset, using the genomes to index relevant sequencing reads that can assist in the curation of genome assemblies. This dataset dramatically increases the sampling of functional genes contained within the uncultivated world of microorganisms, expediting bioprospecting of enzymes from a range of ecosystems. Collectively, our knowledge of life's evolutionary history and the critical processes catalysed by microorganisms will be markedly improved through access to the UBA genomes. ❐

Lindsey M. Solden and Kelly C. Wrighton*
*Department of Microbiology, The Ohio State University, Columbus, OH 43210, USA.*
**e-mail: wrighton.1@osu.edu*

Published online: 25 October 2017
DOI: 10.1038/s41564-017-0048-8

## References

1. Falkowski, P. G., Fenchel, T. & Delong, E. F. *Science* **320**, 1034–1039 (2008).
2. Bowers, R. M. et al. *Nat. Biotech.* **35**, 725–731 (2017).
3. Solden, L. M., Lloyd, K. & Wrighton, K. C. *Curr. Opin. Microbiol.* **31**, 217–226 (2016).
4. Parks, D. H. et al. *Nat. Microbiol.* https://doi.org/10.1038/s41564-017-0012-7 (2017).
5. Antharaman, K. et al. *Nat. Commun.* **7**, 13219 (2016).
6. Konstantinidis, K. T., Rossello-Mora, R. & Amann, R. *ISME J.* https://doi.org/gbprgw (2017).
7. Rinke, C. et al. *Nature* **499**, 431–437 (2013).
8. Brown, C. T. et al. *Nature* **523**, 208–211 (2015).
9. Hug, L. A. et al. *Nat. Microbiol.* **1**, 16048 (2016).
10. Vanwonterghem, I. et al. *Nat. Microbiol.* **1**, 16170 (2016).

## Competing interests

The authors declare no competing financial interests.