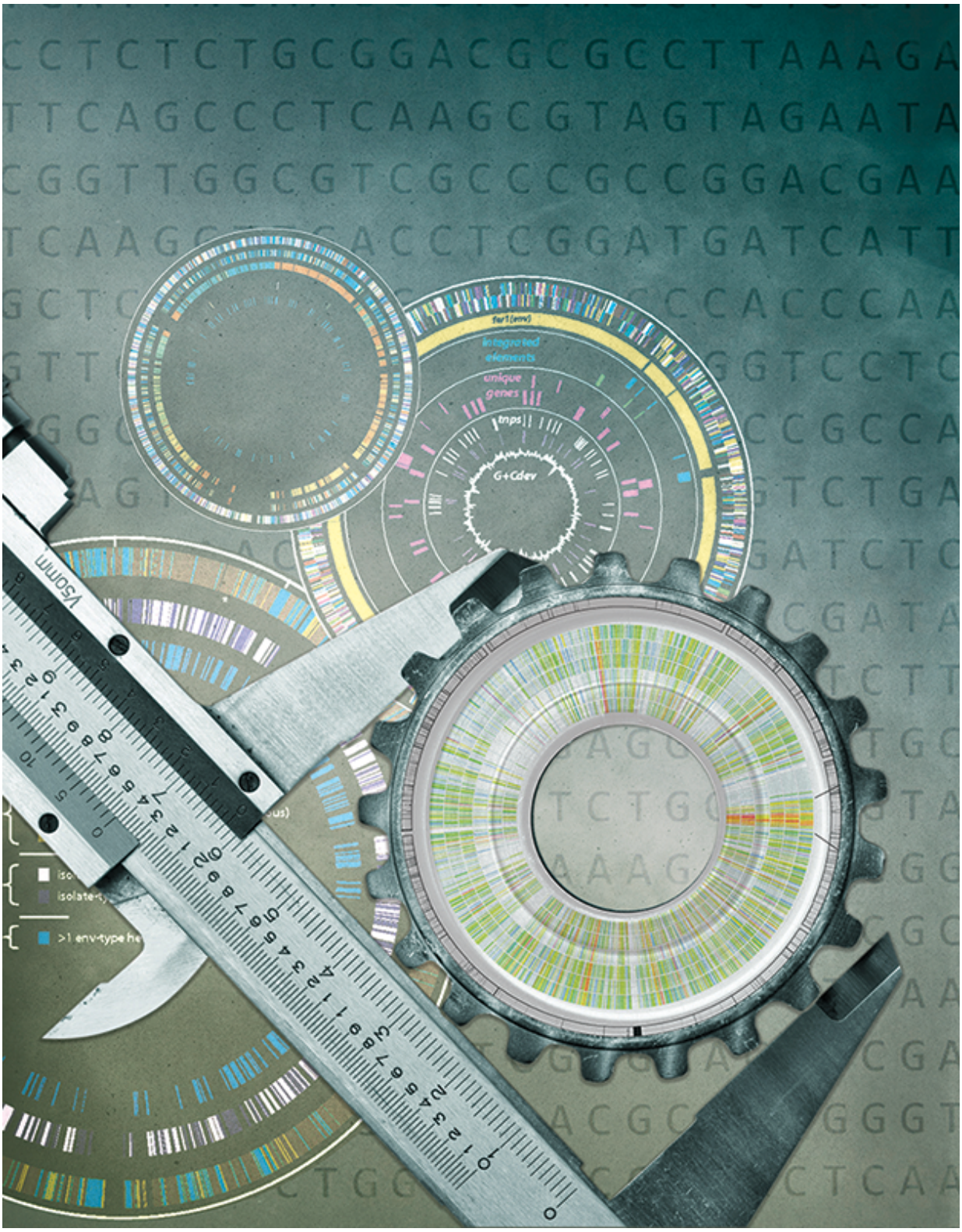
Defining standards for genomes from uncultivated microorganisms

August 9 2017



The importance of standards is dramatically illustrated when they don't exist or

are not commonly accepted. an international team led by DOE JGI researchers has developed standards for the minimum metadata to be supplied with single amplified genomes (SAGs) and metagenome-assembled genomes (MAGs) submitted to public databases. Credit: Zosia Rostomian, Berkeley Lab Creative Services

During the Industrial Revolution, factories began relying on machines rather than people for mass production. Amidst the societal changes, standardization crept in, from ensuring nuts and bolts were made identically to maintain production quality, to a standard railroad gauge used on both sides of the Atlantic. The importance of standards is dramatically illustrated when they don't exist or are not commonly accepted, e.g., Macs, vs. PCs, or even pounds vs. kilograms.

More than a century after the Industrial Revolution, advances in DNA sequencing technologies have caused similarly dramatic shifts in scientific research, and one aspect is studying the planet's biodiversity. Microbes play crucial roles in regulating global cycles involving carbon, nitrogen, and phosphorus among others, but many of them remain uncultured and unknown. Learning more about this so-called "microbial dark matter" involves extracting microbial genomes from the amplified DNA of single cells and from metagenomes. As genomic data production has ramped up over the past two decades and is being generated on various platforms around the world, scientists have worked together to establish definitions for terms such as "draft assembly" and data collection standards that apply across the board. One critical term that needs standardization is "metadata," defined simply as "data about other data." In the case of sequence data, metadata can encompass what organism or cell was sequenced, where it came from, what it was doing, quality metrics, and a spectrum of other characteristics that add value to the sequence data by providing context for it and enabling greater

biological understanding of the significance of the sequence.

Published August 8, 2017 in *Nature Biotechnology*, an international team led by researchers at the U.S. Department of Energy Joint Genome Institute (DOE JGI), a DOE Office of Science User Facility, has developed standards for the minimum metadata to be supplied with single amplified genomes (SAGs) and metagenome-assembled genomes (MAGs) submitted to public databases. "Over the last several years, single-cell genomics has become a popular tool to complement metagenomics," said study senior author Tanja Woyke, head of the DOE JGI Microbial Genomics Program. "Starting 2007, the first single-cell genomes from environmental cells appeared in public databases and they are draft assemblies with fluctuations in the data quality. Metagenome-assembled genomes have similar quality challenges. For researchers who want to conduct comparative analyses, it's really important to know what goes into the analysis. Robust comparative genomics relies on extensive and correct metadata."

Categories of Genome Quality

In their paper, Woyke and her colleagues proposed four categories of genome quality. Low-Quality Drafts would be less than 50 percent complete, with minimal review of the assembled fragments and less than 10 percent contaminated with non-target sequence. Medium-Quality Drafts would be at least 50 percent complete, with minimal review of the assembled fragments and less than 10 percent contamination. High-Quality Drafts would be more than 90 percent complete with the presence of the 23S, 16S and 5S rRNA genes, as well as at least 18 tRNAs, and with less than 5 percent contamination. The Finished Quality category is reserved for single contiguous sequences without gaps and less than 1 error per 100,000 base pairs.

The DOE JGI has generated approximately 80 percent of the over 2,800

SAGs and more than 4,500 MAGs currently accessible on the DOE JGI's Genomes OnLine Database (GOLD). DOE JGI scientist and study first author Bob Bowers said many of the SAGs already in GOLD would be considered Low-Quality or Medium-Quality Drafts. These are highly valuable datasets, though for some purposes, researchers might prefer to use High-Quality or Finished datasets. "Single cell and metagenomic datasets vary greatly in their overall quality. However, in cases where a low quality, fragmented genome is the only representative of a new branch on the tree of life, some data is better than no data," he added. "Bringing up the proposed categories will force scientists to carefully consider genome quality before submission to the public databases."

From Proposal to Community Implementation

Moving from a proposal in print to implementation requires community buy-in. Woyke and Bowers conceived of the minimum metadata requirements for SAGs and MAGs as extensions to existing metadata standards for sequence data, referred to as "MIxS," developed and implemented by the Genomic Standards Consortium (GSC) in 2011. The GSC is an open-membership working body that ensures the research community is engaged in the standards development process and includes representatives from the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI). This is important since these are the main data repositories where the minimum metadata requirements are implemented. By working directly with the data providers, the GSC can assist both large-scale data submitters and databases to align with the MIxS standard and submit compliant data.

"Other key public microbiome data management systems such as MG-RAST, IMG and GOLD have also adapted the MIxS standards," said Nikos Kyrpides, head of the DOE JGI Prokaryote Super Program and GSC Board member. He notes that as part of the DOE JGI's core

mission, the Institute has been involved in organizing the community to develop genomic standards. "The GSC has been instrumental in bringing the community together to develop and implement a growing body of relevant standards. In fact, the need to expand MIxS to uncultivated organisms was identified in one of the recent GSC meetings at the DOE JGI."

"These extensions complement the MIxS suite of metadata standards by defining the key data elements pertinent for describing the sampling and sequencing of single-cell genomes and genomes from metagenomes," said GSC President and study co-author Lynn Schriml of the Institute of Genome Sciences at University of Maryland School of Medicine. "These standards open up a whole new area of metadata data exploration as the vast majority of microbes, referred to as microbial dark matter, are currently not described within the MIxS standard."

She described the group and their mission as community-driven. "I think it helps that the people developing standards are the people conducting the studies," she said. "We have a vested interest in the data. Research is growing and expanding and it is critical that we capture this data in a rigorous way. Developing these novel metadata standards enables researchers to consistently report the most critical metadata for analysis. Capturing data using controlled vocabularies facilitates data consistency, thus making the databases richer and reusable." And in the end, it is to be hoped, sequence data accompanied by agreed-on standards for metadata will mean the same thing to everyone who wants to use it.

More information: Robert M Bowers et al, Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea, *Nature Biotechnology* (2017). DOI: 10.1038/nbt.3893

Provided by DOE/Joint Genome Institute