

Scientists produce new QC tool for microbial genomes (w/ Video)

May 25 2010

More than a thousand microbial genomes have been sequenced at various sequencing centers in the past 15 years to better understand their roles in tasks ranging from bioenergy to health to environmental cleanup. Conservative estimates suggest roughly 10,000 microbial genomes will be publicly available within the next two years, but genomic standards have not caught up with the technological advances that have made the sequencing process faster and cheaper. As a result, the torrent of DNA sequences being released has varying levels of quality, which impacts researchers' ability to use this information.

To assist in checking the quality of the microbial genomic DNA sequences generated before they are submitted to the federally funded public archive GenBank, the U.S. Department of Energy (DOE) Joint Genome Institute (JGI) has introduced a quality control tool known as the Gene Prediction Improvement Pipeline or GenePRIMP.

GenePRIMP is described in a paper published online May 2 in [Nature Methods](#) and has the potential to become a standard in prokaryotic gene calling, a technique by which the start and end of potential [gene coding](#) sequences are identified.

First author Amrita Pati, a software developer in the DOE JGI's [Genome Biology](#) Program noted that GenePRIMP double-checks the gene boundaries, gene annotations and unannotated intergenic regions in genome sequences after the finishing process. She credited colleague Natalia Ivanova with establishing the biological basis of the software tool and helping to refine GenePRIMP. The program, said Pati, identifies

gene-calling errors such as potentially incorrect gene start and end positions, large overlaps between genes, fragmented genes and missed genes.

Gene-calling errors, noted Pati, can range from two percent to as much as 30 percent of the original genes identified in the genome and are dependent on many factors, such as [horizontal gene transfer](#) between species. For example, genes acquired by horizontal gene transfer could become pseudogenes (similar to genes but not coding for a gene product) and are more error-prone. Pati said GenePRIMP significantly reduces the amount of time scientists spend checking the whole genome by specifically highlighting errors that need to be manually corrected.

"Without a GenePRIMP report to work with," she said, singling out the microbe *Starkeya novella*—a soil bacterium that plays an important role in regulating the cycling of carbon and sulfur—as an example, "a DOE JGI scientist would have to examine 4,480 gene models. With GenePRIMP they have to examine less than a tenth of that number of genes." She said that the GenePRIMP report offers additional value because it already includes any anomalies found through BLAST (Basic Local Alignment Search Tool), which finds regions of local similarity between sequences, so scientists don't have to run their own BLAST search.

"With GenePRIMP we have achieved a major breakthrough in the improvement of the quality of structural annotations such as gene predictions," said Genome Biology Program head and study senior author Nikos Kyrpides. He pointed out that using GenePRIMP offers researchers three major advantages: high quality results with reduced errors; an approach that can be used regardless of the automated software originally used to check gene annotations; and finally a method to standardize gene calling.

"There are a lot of different tools used for predicting genes in prokaryotes," Kyrpides said. "The major problem we have is that they all produce very variable results. This impedes our ability to compare genomes sequenced and annotated from various sources, as they use different tools for gene prediction. GenePRIMP is not substituting any of the available methods; a user can employ any available automatic gene prediction method, and then use GenePRIMP to correct the initial output. It will generate a much more standardized output, thus not only significantly improving quality, but also significantly facilitating comparative analysis."

Dawn Field, President of the Genomics Standards Consortium, an international initiative of genomics researchers interested in establishing standards for collecting and capturing genomic data to the general community, called GenePRIMP "a great solution to a long-standing problem in computational bioinformatics - how to clean up gene calls based on comparative genomic data. Ideally," she added, "the underlying principles will pave the way for new standards in gene calling."

GenePRIMP is available for use by researchers at <http://geneprimp.jgi-psf.org/>. Genomics researchers supported by a range of federal agencies and other funding sources are expected to take advantage of this new quality control tool. The current version of the software finds and reports gene model anomalies to the scientists. Pati said that a future version of GenePRIMP will automatically find and correct said anomalies as well as report frameshifts (genetic mutation caused by insertion or deletion of nucleotides) and pseudogenes.

"Consistent high-quality annotation on microbial genomes is key to their utility," said Owen White, director of bioinformatics at the University of Maryland School of Medicine and head of the Human Microbiome Project Data Analysis and Coordination Center that tracks, stores, analyzes and distributes the data. "Software such as GenePRIMP is an

important component in our quality control toolbox."

Pati said the automated software tool is the crystallization of manual operating procedures used for more than 3 years for correcting gene models at the DOE JGI. "As such, it is also following the principles of standardization of the Genomics Standards Consortium and further development will factor in the Consortium's recommendations," she and her colleagues wrote, while enabling faster, better and cheaper analyses.

Provided by DOE/Joint Genome Institute

Citation: Scientists produce new QC tool for microbial genomes (w/ Video) (2010, May 25)
retrieved 28 April 2024 from

<https://phys.org/news/2010-05-scientists-qc-tool-microbial-genomes.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.