UK Effort to Sequence 3K Bacterial Genomes Nears Goal

Apr 23, 2018 | Ben Butkus

MADRID (GenomeWeb) – A Wellcome Trust-funded collaboration between Public Health England and the Sanger Institute is most of the way to its goal of sequencing 3,000 bacterial strains culled from the National Collection of Type Cultures, a PHE-run non-profit biorepository.

Having sequenced more than 2,700 distinct strains from NCTC, the scientists behind the project are currently busy assembling genomes and ensuring their scientific quality and accessibility by uploading them to a public database, thereby filling important gaps in terms of reference genomes for these organisms.

In a presentation at the European Conference of Clinical Microbiology and Infectious Diseases held here over the weekend, Sarah Alexander, a clinical scientist at PHE, provided an update on the project, which began in 2013 with sequencing services provided by the Sanger Institute and powered by Pacific Biosciences.

"We think this is the largest project in the world generating whole-genome sequences for prokaryotic organisms," Alexander said.

The NCTC is a bacterial strain collection founded in 1920 that has evolved into a diverse and dynamic collection of some 5,200 bacterial strains. The goal of the NCTC 3000 project, she noted, was to generate reference genomes for 3,000 of these strains and embed the data in an accessible resource to "enhance the value of the collection."

One of the key challenges in this undertaking, Alexander noted, was extracting high molecular weight DNA from NCTC strains for whole-genome sequencing, a task made easier by generating DNA quality profiles on Agilent Technologies' TapeStation platform for NGS quality control.

This non-trivial step in the process has resulted in DNA extractions for more than 3,200 strains to date, representing 852 different bacterial species and 82 families, Alexander said.

Of the more than 2,700 strains sequenced thus far, more than 56 percent were assembled with a single contig and more than 92 percent were assembled with fewer than five contigs.

One of the major scientific impacts of the project so far is the generation of sequences of 852, or about 95 percent, of the "type strains" in the NCTC. These type strains, Alexander explained, are the "reference point" to which all other strains are compared "to know whether they belong to that species." The group has paid special attention to the curation of these strains due to their scientific importance, she said.

To highlight the importance of this aspect of the project, Alexander pointed to current sequence availability in the US National Center for Biotechnology Information database, where 108, or 12 percent of those type strains, have no whole-genome sequence available for the species at all, while 298, or 30 percent of the type strains, have no WGS available

for those strains in particular. About 44 percent of the 852 types strains listed in NCBI have only draft genome sequences available.

The type strains are "an essential requirement to describe new bacterial species," Alexander said, noting that NCTC 3000 is filling important gaps for available reference genomes.

The NCTC 3000 data that has been generated has also already been applied by the scientific community and utilized in multiple external peer-reviewed studies. Further, scientists were able to identify antimicrobial-resistance genes in Enterobacteriaceae in comparison to the date when those strains were isolated, providing valuable insight into drug resistance in this organism.

Finally, the group has generated multiple genomic datasets for many species of high interest to human health — for example, they've generated more than 250 reference genomes for Escherichia coli and 167 for Staphylococcus aureus.

Having made tremendous progress in their effort over the last few years, the NCTC 3000 scientists now believe they will be able to sequence about 3,300 strains representing some 852 species and are encouraging scientists worldwide to use the data, which is hosted on the Sanger Institute website. They are also encouraging scientists to deposit clinically relevant strains into the NCTC for inclusion in the sequencing project.