GENOMICS

Contamination in sequence databases

A fast algorithm detects unexpected contamination events in public databases.

Biological sequences in public databases are indispensable resources for life science research. Despite our everyday reliance on these databases, there are gaps, errors and contamination in the data. "One of our research efforts for the past several years has been the detection of pathogens in humans by using metagenomic shotgun sequencing," says Martin Steinegger, who was a member in Steven L. Salzberg's lab at Johns Hopkins University and is now at the Seoul National University. "Unfortunately, in many cases we have found that contamination within the genome sequences produces false positives."

This motivated Steinegger and Salzberg to start a project to assess contamination in the GenBank, RefSeq and NR databases. Using recent fast algorithms, they developed a tool called Conterminator that enables searching for contamination across kingdoms and scales linearly. "The version of GenBank we evaluated had a size of 3.3 terabytes and contained 400 million sequences. Aligning them all-against-all would require hundreds of years using classic methods," says Steinegger. "Our algorithm only required 12 days to process all of GenBank on a single 32-core server."

They expected to see a few thousand contaminated sequences but ended up with millions. 2,161,746, 114,035 and 14,148 contaminated sequences were detected in GenBank, RefSeq and NR, respectively. "This single most surprising finding was the presence of a piece of a bacterium, *Acidithiobacillus thiooxidans*, in an alternative scaffold of the current version of the human reference genome (GRCh38)," says Steinegger. "The human genome has been around for such a long time, and so many researchers use it on a daily basis, that we didn't expect to see any contaminants there."

Steinegger hopes Conterminator can help researchers who sequence genomes and database managers to detect contamination. As a word of caution to users of genome sequences, "Many of the genomes contain contamination. One particular problem that arises, again and again, is that contamination leads to incorrect claims about horizontal gene transfer," says Steinegger.

Lin Tang

Published online: 2 July 2020
https://doi.org/10.1038/s41592-020-0895-8

Research paper
Steinegger, M. & Salzberg, S. L. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* **21**, 115 (2020).