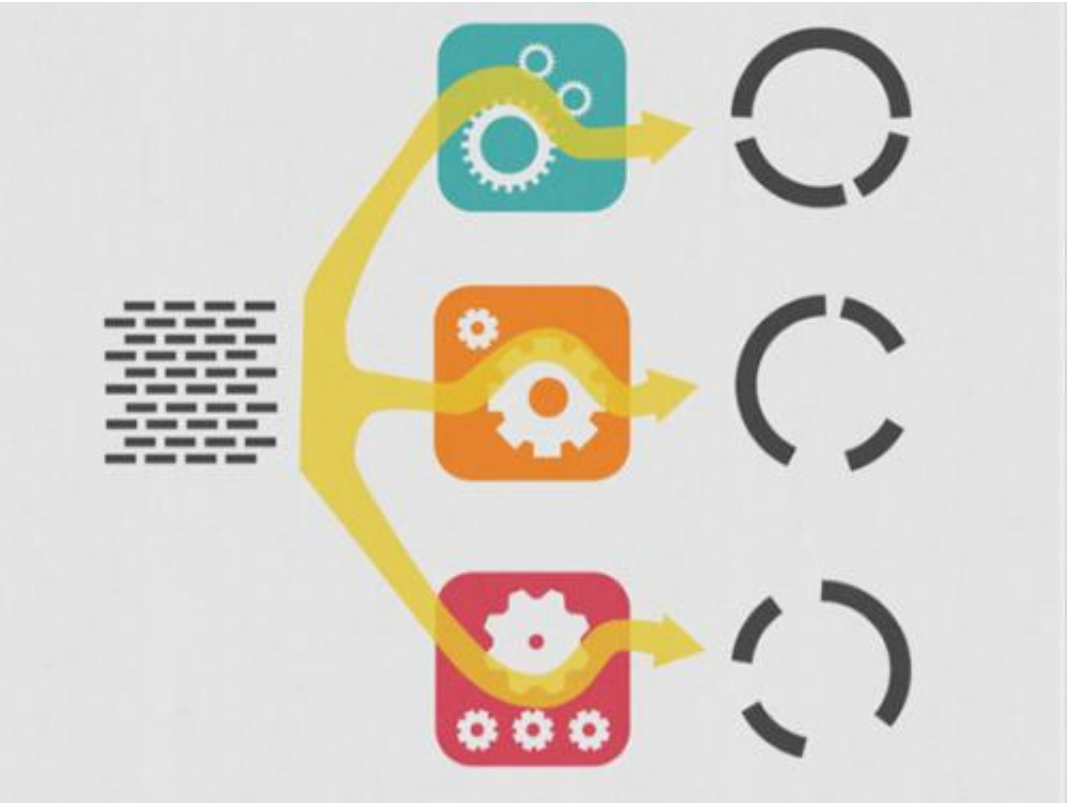
Automating the selection process for a genome assembler

October 20 2014



The process of selecting the right genome assembler for the job is being automated at the DOE JGI, and bioinformaticist Michael Barton welcomes other assembler submissions to the nucleotid.es repository. Credit: Michael Barton

A repository of genome assemblers is being developed to automate the process of selecting the best assembler for the task at hand.

There are many different genome assemblers being introduced and touted. On the nucleotid.es site (nucleotid.es/), the test results for various genome assemblers provide reproducible findings that genomics researchers can use to select the appropriate assembler for their needs.

After an organism's genetic code has been sequenced, researchers have to assemble the DNA fragments into a single sequence to be able to parse the information. However, selecting an assembler while considering factors such as the large number of short sequence reads generated, repeated sequences, and lack of a reference genome sequence against which to compare the draft assembly can be challenging.

At the U.S. Department of Energy Joint Genome Institute (DOE JGI), a DOE Office of Science user facility, bioinformatics systems analyst Michael Barton has been developing a repository of genome assemblers called nucleotid.es to help the DOE JGI team address these questions for sequencing projects in process. Right now, he said, the process of selecting a genome assembler is manual so an automated pipeline would be very helpful. The repository at http://nucleotid.es/ is publicly available so that other bioinformaticists can benefit from the findings being generated.

"A lot of assemblers are being produced in the bioinformatics community, and instead of reading subjective papers with assemblers, you can test the assemblers for yourself," Barton said, "with the added benefit of having reproducible research so that anyone can produce the results."

Barton started with genome assemblers that are being used by the DOE JGI, and he tested them against an internal dataset of several microbial genomes. The findings are categorized by benchmarks such as NG50 (a statistic which tracks the average length of a set of DNA sequences) on the website so that bioinformaticists can see how each assembler fared at

the criteria of interest to them.

Each of the assemblers on the nucleotid.es site is enclosed in virtual boxes called docker containers. The docker containers make it easy to share and use the software. If a bioinformaticist finds a particular assembler useful, they can easily download it from the nucleotid.es site. Conversely, if other bioinformaticists want to see another assembler on the site, Barton said, they can send him the docker container for posting.

So far, he said, the genome assemblers on nucleotid.es are testing microbial genomes that have come off Illumina sequencers. He plans to add assemblers such as meraculous, an assembler for plant genomes developed at the DOE JGI, and jigsaw and allpaths. Barton said eventually he also hopes to have assemblers for other types of genome projects on nucleotides.

Provided by DOE/Joint Genome Institute

Citation: Automating the selection process for a genome assembler (2014, October 20) retrieved 17 April 2024 from https://phys.org/news/2014-10-automating-genome.html