HybPiper: A bioinformatic pipeline for processing target-enrichment data

November 1 2016

With the rapid rise of next-generation sequencing technologies, disparate fields from cancer research to evolutionary biology have seen a drastic shift in the way DNA sequence data is obtained. It is now possible to sequence many genes across large numbers of species in an incredibly short period of time. And the price tag keeps getting smaller and smaller. However, the deluge of sequence data obtained using these high-throughput sequencing techniques requires a substantial amount of computational input to process—a daunting task for many biologists. A recently developed bioinformatics pipeline allows researchers with limited computational skills to quickly and efficiently extract gene regions of interest from data obtained with the increasingly popular targeted sequence capture approach.

Targeted sequence capture is a technique used to focus sequencing efforts on specific regions of the genome. By reducing the size of the target genome to only those gene regions of interest, many more samples can be sequenced concurrently. A recent study led by scientists at the Chicago Botanic Garden and available in *Applications in Plant Sciences* describes the pipeline, HybPiper, for recovering gene regions from sequence data obtained using this technique.

"We set out to design a tool to reliably extract gene sequences from high-throughput sequencing projects to build phylogenetic trees," explains Dr. Matthew Johnson, lead author of the study. "Scientists using next-generation sequencing technologies get their data delivered in a big pile of DNA fragments. HybPiper decides which fragments belong to which gene, assembles the fragments into a gene region, and returns the full gene sequence, including introns, in a format that can be used for downstream analysis."

The pipeline brings together a number of Python scripts and free-standing programs to create a simple-to-use workflow for processing large amounts of sequence data. "We used a variety of tools at each phase, and tweaked the parameter settings until we were consistently recovering the right sequence. We also tried to be sensitive to different targeted sequencing designs—for example, not everyone will be able to design probes from a closely related genome. This flexibility is reflected in a large number of customizable parameters in HybPiper to better fit each individual project," explains Johnson.

One feature that is particularly useful, especially for those researchers working with plants, is HybPiper's ability to detect duplicate genes. Because all flowering plants, for example, have at least one whole genome duplication in their shared evolutionary history, the detection of paralogous gene copies is an essential part of accurately estimating species relationships. This, however, can be an exceedingly difficult and time-consuming task. Enter HybPiper. Built into the pipeline is the ability to detect duplicate genes within a molecular dataset. Johnson explains, "Sorting DNA sequencing fragments can be tricky when what seems like one gene is really two closely related genes. HybPiper has tools that will allow users to avoid this issue and detect whether a gene has been duplicated in their study organism."

Dr. Johnson concludes, "Development of HybPiper is ongoing. We have set up a website (github.com/mossmatters/HybPiper) that helps users with installation issues and a comprehensive tutorial using an example dataset. We encourage users to provide feedback and suggest new features that will help them with their target enrichment analysis."

More information: Matthew G. Johnson et al, HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment, *Applications in Plant Sciences* (2016). DOI: 10.3732/apps.1600016

Provided by Botanical Society of America

Citation: HybPiper: A bioinformatic pipeline for processing target-enrichment data (2016, November 1) retrieved 27 April 2024 from https://phys.org/news/2016-11-hybpiper-bioinformatic-pipeline-target-enrichment.html