MedicalXpress

Researchers build SEQSpark to analyze massive genetic data sets

June 30 2017

Uncovering rare susceptibility variants that contribute to the causes of complex diseases requires large sample sizes and massively parallel sequencing technologies. These sample sizes, often made up of exome and genome data from tens to hundreds of thousands of individuals, are often too large for current analytical tools to process. A team at Baylor College of Medicine, led by Dr. Suzanne Leal, professor of molecular and human genetics, has developed new software called SEQSpark to overcome this processing obstacle. A study on the new technology appears in The *American Journal of Human Genetics*.

"To handle these large data sets, we built the SEQSpark tool based on the commonly used Spark program, which allows SEQSpark to utilize multiple processing platforms to increase the speed and efficiency of performing data quality control, annotation and rare variant association analysis," Leal said.

To test and validate the versatility and speed of SEQSpark, Leal and her team analyzed benchmarks from the whole genome sequence data from the UK10K, testing specifically for waist-to-hip ratios.

"The analysis and related tasks took about one and a half hours to complete, in total. This includes loading the data, annotation, principal components analysis and single and rare variant aggregate association analysis for the more than 9 million variants present in this sample set," explained Di Zhang, a postdoctoral associate in the Leal lab at Baylor and first author on the paper.

To evaluate SEQSpark's performance in a larger data set, Leal and the research team generated 50,000 simulated exomes. The SEQSprak program ran the analysis for a quantitative trait using several variant aggregate association methods in an hour and forty-five minutes.

When compared to other variant association tools, SEQSpark was consistently faster, reducing computation to a hundredth of the time in some cases.

"What is unique about SEQSpark is that it is scalable, and smaller labs can run it without super specific hardware, and it can also be run in a multi-server environment to increase its speed and capacity for large genetic data sets," Zhang said. "It is ideal for large-scale genetic epidemiological studies and is highly efficient from a computational standpoint."

"We see this software as being very useful as the demand for the analysis of massively parallel sequence data grows. SEQSpark is highly versatile, and as we analyze increasingly large sets of rare variant data, it has the potential to play a key role in furthering personalized medicine," Leal said.

In the future, Leal and her team will continue to test and increase SEQSpark's capabilities and will be analyzing soon data sets that have 500,000 samples or more.

More information: Di Zhang et al. SEQSpark: A Complete Analysis Tool for Large-Scale Rare Variant Association Studies using Whole-Genome and Exome Sequence Data, *The American Journal of Human Genetics* (2017). DOI: 10.1016/j.ajhg.2017.05.017