New software analyzes human genomes faster than other available technologies

January 30 2015

Investigators at Nationwide Children's Hospital have developed an analysis "pipeline" that slashes the time it takes to search a person's genome for disease-causing variations from weeks to hours. An article describing the ultra-fast, highly scalable software was published in the latest issue of *Genome Biology*.

"It took around 13 years and $3 billion to sequence the first human genome," says Peter White, PhD, principal investigator and director of the Biomedical Genomics Core at Nationwide Children's and the study's senior author. "Now, even the smallest research groups can complete genomic sequencing in a matter of days. However, once you've generated all that data, that's the point where many groups hit a wall. After a genome is sequenced, scientists are left with billions of data points to analyze before any truly useful information can be gleaned for use in research and clinical settings."

To overcome the challenges of analyzing that large amount of data, Dr. White and his team developed a computational pipeline called "Churchill." By using novel computational techniques, Churchill allows efficient analysis of a whole genome sample in as little as 90 minutes.

"Churchill fully automates the analytical process required to take raw sequence data through a series of complex and computationally intensive processes, ultimately producing a list of genetic variants ready for clinical interpretation and tertiary analysis," Dr. White explains. "Each step in the process was optimized to significantly reduce analysis time,

without sacrificing data integrity, resulting in an analysis method that is 100 percent reproducible."

The output of Churchill was validated using National Institute of Standards and Technology (NIST) benchmarks. In comparison with other computational pipelines, Churchill was shown to have the highest sensitivity at 99.7 percent; highest accuracy at 99.99 percent and the highest overall diagnostic effectiveness at 99.66 percent.

"At Nationwide Children's we have a strategic goal to introduce genomic medicine into multiple domains of pediatric research and healthcare. Rapid diagnosis of monogenic disease can be critical in newborns, so our initial focus was to create an analysis pipeline that was extremely fast, but didn't sacrifice clinical diagnostic standards of reproducibility and accuracy" says Dr. White. "Having achieved that, we discovered that a secondary benefit of Churchill was that it could be adapted for population scale genomic analysis."

By examining the computational resource use during the data analysis process, Dr. White's team was able to demonstrate that Churchill was both highly efficient (>90 percent resource utilization) and scaled very effectively across many servers. Alternative approaches limit analysis to a single server and have resource utilization as low as 30 percent. This efficiency and capability to scale enables population-scale genomic analysis to be performed.

To demonstrate Churchill's capability to perform population scale analysis, Dr. White and his team received an award from Amazon Web Services (AWS) in Education Research Grants program that enabled them to successfully analyze phase 1 of the raw data generated by the 1000 Genomes Project - an international collaboration to produce an extensive public catalog of human genetic variation, representing multiple populations from around the globe. Using cloud-computing

resources from AWS, Churchill was able to complete analysis of 1,088 whole genome samples in seven days and identified millions of new genetics variants.

"Given that several population-scale genomic studies are underway, we believe that Churchill may be an optimal approach to tackle the data analysis challenges these studies are presenting," says Dr. White.

The Churchill algorithm was licensed to Columbus-based GenomeNext LLC, which has built upon the Churchill technology to develop a secure and automated software-as-a-service platform that enables users to simply upload raw whole-genome, exome or targeted panel sequence data to the GenomeNext system and run an analysis that not only identifies genetic variants but also generates fully annotated datasets enabling filtering and identification of pathogenic variants. The company provides genomic data analysis solutions that simplify the process of data management and automate analysis of large scale genomic studies. The system was also developed with the research and clinical market in mind, offering a standardized pipeline that is well suited to settings where customers have to meet regulatory requirements.

  More information: Kelly BJ, Fitch JR, Hu Y, Corsmeier DJ, Zhong H, Wetzel AN, Nordquist RD, Newsom DL, White P. Churchill: an ultra-fast, deterministic, highly scalable and highly balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. *Genome Biology* 2015, 16:6. DOI: 10.1186/s13059-014-0577-x , genomebiology.com/2015/16/1/6/abstract

Provided by Nationwide Children's Hospital

January 30) retrieved 20 March 2024 from https://medicalxpress.com/news/2015-01-software-human-genomes-faster-technologies.html