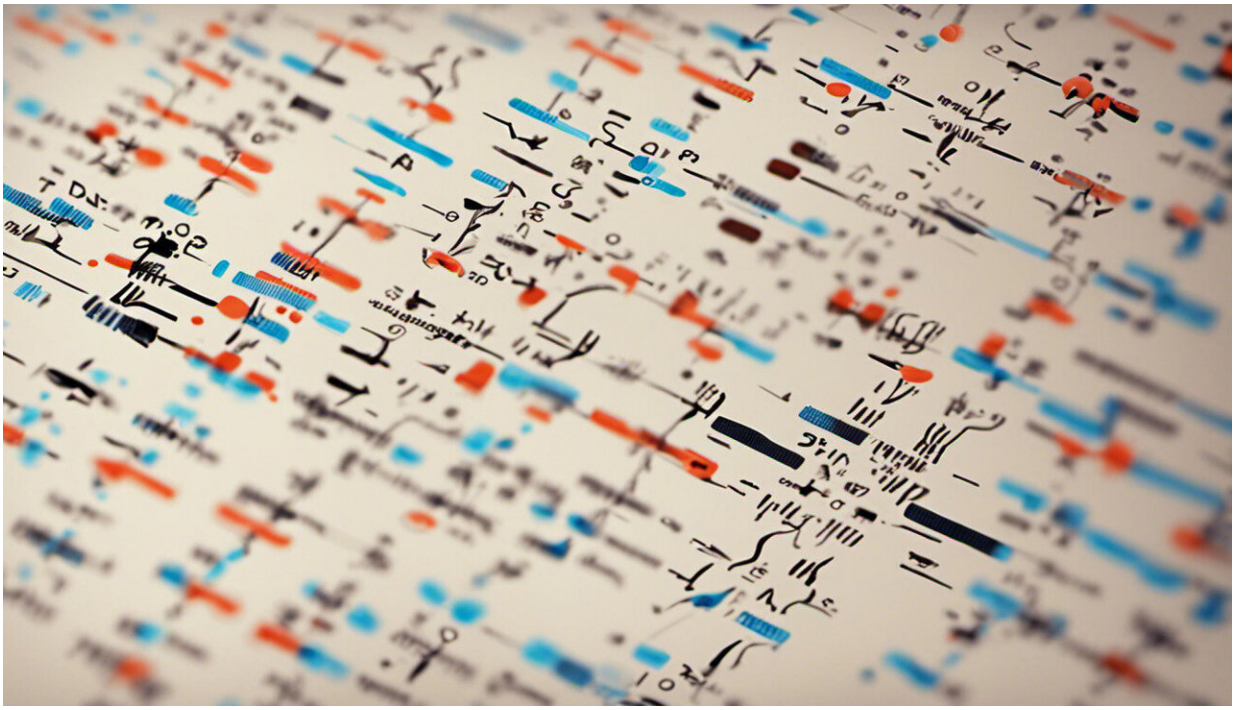
DNA sample analysis times dramatically reduced thanks to new file format

February 2 2022, by Neil Martin



Credit: AI-generated image (disclaimer)

Nanopore sequence analysis can now be done more than 30 times faster, providing quicker and better specialized treatments for patients with cancer and other diseases.

A new computer file format which helps process DNA samples 30 times

faster than existing systems has been developed by teams at UNSW and the Garvan Institute of Medical Research.

The SLOW5 format has been specifically designed to more efficiently analyze nanopore sequences, which provide a more complete view of genetic variations.

The improved efficiency not only helps medical experts analyze individual DNA samples much more quickly and provide faster and better healthcare to patients, but also allows more sampling in any given period.

The research behind the development has been published in *Nature Biotechnology*, but the software has already been made available through open source and has been downloaded more than 1,000 times in just a few weeks.

The complex nature of nanopore sequencing of DNA means huge amounts of data are created which then needs to be stored and properly analyzed.

This data has routinely been recorded in a file format called FAST5, with such complex information often producing files around 1.3 terabyte in size—roughly the equivalent of 650 hours of HD video.

Historically, it has taken around two weeks for computers to process such large FAST5 files and analyze the human genome information contained on them.

But Dr. Hasindu Gamaarachchi, previously a UNSW Ph.D. candidate working under Professor Sri Parameswaran in the School of Computer Science and Engineering, has now created a file format designed for efficient, scalable analysis of nanopore signal data.

The new SLOW5 file format not only significantly reduces the size of the files, but can also process the exact same information in around 10.5 hours—more than 30 times faster than FAST5.

Parallel computing

The key to this is that the SLOW5 format, unlike FAST5, allows for efficient parallel computing, whereby several processors can simultaneously execute multiple, smaller analyses broken down from the much bigger, more complex, complete dataset.

Credit: AI-generated image (disclaimer)

"You can think of this like trying to dig a very big hole with ten people,

but there is only one shovel they have to share round," said Dr. Gamaarachchi, lead author of the paper who is now a Genomics Computing Systems Engineer at the Garvan Institute.

"That's how it used to be with FAST5. But with SLOW5 everyone gets their own shovel and they can all dig at the same time and do the job much faster.

"The FAST5 format is slow because the data cannot be accessed in parallel. It is based around the Hierarchical Data Format which was designed in the 1990s to work on machines which at the time only had one processor, rather than the modern ones which include multiple processors.

"The Hierachical Data Format is also generic, whereas the SLOW5 is purpose-built. So in terms of the digging analogy, it's like we are also providing a shovel that is specially designed for the type of soil.

"And because the new SLOW5 can be accessed in parallel by multiple processors at the same time, the processing time has reduced by a factor of 30," he said.

"So instead of it taking around two weeks to process the data for a human genome, the time has now dropped down to less than half a day."

Nanopore sequencing itself offers a more complete view of genetic variations and the possibility to reconstruct complex genomes.

A nanopore is a nano-scale hole, over which an ionic current is passed, with alterations in current measured when biological molecules pass through. The alterations are then documented and translated to identify that molecule and base modification.

Nanopore sequencing is used to identify a range of diseases and also helps medical professionals analyze the DNA samples in greater detail to potentially offer more personalized medicine, especially in the treatment of various cancers.

The new SLOW5 file format can now help medical professionals diagnose diseases more quickly and ensure that patients are prescribed specific targeted medicine—often the most effective treatment—much faster than was previously possible.

Dr. Ira Deveson, Head of Genomic Technologies at the Garvan Institute and a co-author of the paper, said: "SLOW5 has removed one of the major bottlenecks to the use of nanopore sequencing, a new technology that has countless potential applications in clinical genetics, agriculture and other bioscience domains.

"With the development of SLOW5, our ability to process nanopore sequencing data can now keep up with our ability to generate it. This will open the door to many new applications in medical science for this exciting, emerging technology."

More information: Hasindu Gamaarachchi et al, Fast nanopore sequencing data analysis with SLOW5, *Nature Biotechnology* (2022). DOI: 10.1038/s41587-021-01147-4

Provided by University of New South Wales