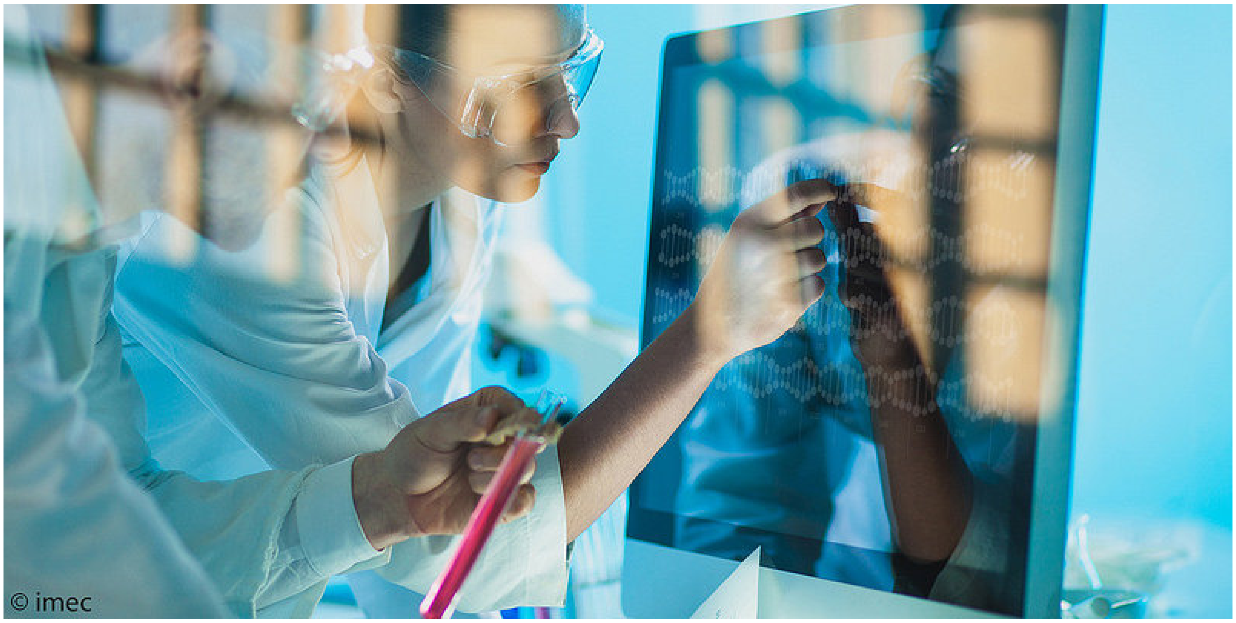
PHYS ORG

Imec's elPrep software significantly speeds up genome sequencing analysis

October 18 2018



Credit: IMEC

This week at ITF Health 2018, imec, the world-leading research and innovation hub in nanoelectronics and digital technologies, showcases elPrep 4.0, a powerful software tool to speed up human DNA sequencing analysis. elPrep accelerates whole genome and exome processing pipelines up to an order of magnitude, saving a typical lab hundreds of hours of computer processing and allowing more and faster DNA tests. elPrep 4.0 is designed as a drop-in replacement for preparation steps defined by the GATK (Genome Analysis Toolkit) Best Practices pipelines for variant calling, while delivering identical results.

DNA sequencing involves splitting a human genome into thousands of fragments, which are then fed to the sequencing machines to identify the individual bases. This results in huge data files that are processed through a pipeline of tools to reconstruct the original DNA sequence from the fragments and to flag variants that may point to e.g. genetic disorders (also known as variant calling). Data sets for human whole genome DNA are usually on the order of several hundreds of GB of uncompressed data, resulting in processing runtimes typically on the order of tens of hours per genome.

elPrep software is designed to speed up DNA sequencing analysis up to an order of magnitude. The new version 4.0 executes all preparation steps until variant calling. It replaces other DNA sequencing analysis software such as GATK4.0, Picard, and SAMtools while producing identical results. What sets elPrep apart is its architecture that allows executing pipelines by making only a single pass through the data, no matter how long the pipeline is.

elPrep is designed as a multi-threaded application that runs entirely in memory, avoids repeated file I/O, and merges the computation of data of several DNA sequencing preparation steps. As a result, in a typical run, elPrep is up to ten times faster than other software tools using the same resources. It is designed as a seamless replacement that delivers the exact same results as GATK4.0 developed by the Broad Institute. elPrep has been written in the Go programming language and is available through the open-source GNU Affero General Public License v3 (AGPL-3.0).

Imec's ExaScience Life Lab is an imec lab focused on providing software solutions for data-intensive high-performance computing problems, primarily in the life sciences domain. It solves data-intensive computational bottlenecks and by doing so helps companies develop solutions for complex problems involving multiple disciplines. Examples of successful projects include large-scale machine learning for pharmaceutical companies, DNA sequencing software for hospitals and pharmaceutical companies, assay image feature extraction, advanced biostatistics and data analytics, and even multi-physics space weather simulations. The work on elPrep 4.0 was partially funded through the imec.icon research project GAP, a research project to optimize the ICT infrastructure for whole genome sequencing in hospitals, in collaboration with Bluebee, Western Digital, Agilent, Ghent University, KU Leuven, and the academic hospital UZ Leuven.

Provided by IMEC

Citation: Imec's elPrep software significantly speeds up genome sequencing analysis (2018, October 18) retrieved 23 June 2024 from https://phys.org/news/2018-10-imec-elprep-software-significantly-genome.html