PHYS.ORG

How to quickly and efficiently identify huge gene data sets to help coronavirus research

March 31 2020

Credit: Max Rahubovskiy from Pexels

Thanks to the advancement of sequencing technology, it's possible to produce massive amounts of genome sequence data on various species. It's crucial to examine pan-genomic data—the entire set of genes possessed by all members of a particular species—particularly in areas like bacteria and virus research, investigation of drug resistance mechanisms and vaccine development. For example, why is the coronavirus resistant to common drugs? Can big data help to rapidly identify the characteristics of such novel virus strains? A group of researchers supported by the EU-funded PANGAIA project is now tackling this challenge by developing methods for comparing gigantic gene data sets.

As explained in a news release by PANGAIA project partner Bielefeld University, scientists often use a reference genome to see whether the genetic material of an organism shows particular variations. "They combine several genomes in such a way that they exhibit the typical characteristics of an entire species. This enables researchers to compare a new influenza virus with a reference genome that summarizes the typical features of the virus strains from which it originates."

Quoted in the same news release, Prof. Dr. Jens Stoye from Bielefeld University says: "In these cases, we compare only two genomes with each other—differences and similarities are relatively easy to identify on the computer." He adds: "With the new approach, we can compare one genome to thousands of other genomes in a single step." This process involves pan-genomics. "The new technology enables a simultaneous, integrated analysis of many strains of the same organism. These can be viruses, bacteria, and sometimes even higher organisms." Prof. Stoye continues: "This makes it possible to highlight the similarities and differences between the individual members. In the case of pathogens, it is often even possible to understand and predict the processes that led to the development of particularly infectious strains."

Genetic abnormalities

The new method can also be used to detect hereditary diseases in humans or to determine genetic abnormalities, according to the same news release. "Over the next few years, we want to develop new algorithms and data structures with our project partners that will make computer-assisted pangenomics faster and more user-friendly," says Prof. Dr. Alexander Schönhuth, also from Bielefeld University. The project team hopes to develop algorithms where computers search for similarities and differences between the comparative genomes and present the results by using variation graphs. These representations enable researchers "to identify completely novel mutations, such as those that have presumably occurred in the variant of the coronavirus" that broke out in China and which "led to resistance to the usual medications," as noted by Prof. Schönhuth.

The ongoing PANGAIA (Pan-genome Graph Algorithms and Data Integration) project will run until end-December 2023. It will focus on graph-based representations of large genome data sets and demonstrate their advantages over traditional sequence-based presentations of pan-genomic data. "In this project, we will put this shift of paradigms—from sequence to graph based representations of genomes—into full effect," as stated on CORDIS. "As a result, we can expect a wealth of practically relevant advantages, among which arrangement, analysis, compression, integration and exploitation of genome data are the most fundamental points."

More information: PANGAIA project: cordis.europa.eu/project/id/872539

Provided by CORDIS

Citation: How to quickly and efficiently identify huge gene data sets to help coronavirus research (2020, March 31) retrieved 20 June 2024 from https://phys.org/news/2020-03-quickly-efficiently-huge-gene-coronavirus.html