Medical Xpress

Trillion data points to identify disease-causing genes

September 2 2020, by Ofa Fitzgibbons

Credit: CC0 Public Domain

Researchers from CSIRO, Australia's national science agency, have accomplished a world-first by processing one trillion points of genomic data through VariantSpark, an artificial intelligence-based platform, which can help pinpoint the location of specific disease-causing genes in the human genome.

The human genome is a person's complete set of DNA, which contains more than three billion DNA base pairs.

CSIRO Bioinformatics Group leader Dr. Denis Bauer said artificial intelligence (AI) could give a deeper understanding of complex diseases, in a fraction of the time compared to traditional approaches, by analyzing immense genomic datasets.

"Our VariantSpark platform can analyze traits, such as diseases or susceptibilities, and uncover which genes may jointly cause them," Dr. Bauer said.

"This can provide valuable information about how the disease works on a molecular level, which can ultimately lead to better treatments. VariantSpark is already being used to help determine what genes might be linked to cardiovascular disease, motor neurone disease, dementia, and Alzheimer's disease."

In a new study published in the technical journal *Giga Science*, the researchers analyzed a synthetic dataset of 100,000 individuals, enabled by Amazon Web Services (AWS).

Dr. Bauer said no other technology platform had been able to process one trillion data points of genomic data, over ten million variants and 100 thousand samples at once.

"Our research shows VariantSpark is the only method able to scale to ultra-high dimensional genomic data in a manageable time," Dr. Bauer said.

"It was able to process this information in 15 hours while it would take the fastest competitors likely more than 100,000 years to process such a volume of data. This is a significant milestone, as it means VariantSpark can be scaled up to analyze population-level datasets and drive better healthcare outcomes."

CSIRO's Australian e-Health Research Center CEO Dr. David Hansen said AI technologies were crucial to the future of healthcare in Australia.

"Artificial intelligence is a critical component of understanding genomic information, which is increasingly being used to guide healthcare delivery in Australia and around the world," Dr. Hansen said.

"Despite recent technology breakthroughs with whole genome sequencing studies, the molecular and genetic origins of complex diseases are still poorly understood which makes prediction, application of appropriate preventive measures and personalized treatment difficult."

VariantSpark was developed by CSIRO's digital health research team at the Australian e-Health Research Center with support from CSIRO's digital specialist arm, Data61.

It is also one of the first machine-learning based health products which allows researchers around the world to access important data critical for developing treatments and accelerating research capabilities, that is available on the AWS Marketplace.

More information: Arash Bayat et al. VariantSpark: Cloud-based machine learning for association study of complex phenotype and large-scale genomic data, *GigaScience* (2020). DOI: 10.1093/gigascience/giaa077

VariantSpark is available for download on AWS Marketplace.

Provided by CSIRO

Citation: Trillion data points to identify disease-causing genes (2020, September 2) retrieved 24 May 2024 from https://medicalxpress.com/news/2020-09-trillion-disease-causing-genes.html