Case Study

High Performance Computing
2nd Generation Intel® Xeon® Scalable Processors
Intel Server System R1208WFTYSR
Intel SSD S4610 Drives

intel®

TRON CoVigator Reveals Evolution of Spike-glycoprotein Mutations in Covid-19 Virus

High Performance Computing (HPC) cluster built on 2nd Gen Intel® Xeon® Scalable processors analyzes millions of genomic samples of SARS-CoV-2 virus to identify variants of key disease transmitter

Solution Summary:

- 10 Intel® Server System R1208WFTYSR computational nodes
- 2nd Gen Intel Xeon Gold® 6240R processors (480 cores and 960 threads)
- 2nd Gen Intel Xeon Silver 4208 processors for log-in node
- Intel SSD S4610 480 GB drives

Executive Summary

Researchers at TRON gGmbH have been studying the immuno-biology of cancers for more than ten years. As the Covid-19 pandemic raged in 2020, TRON scientists applied their expertise and knowledge in genome analytics to the SARS-CoV-2 virus. They began studying how it evolves, and specifically the variants of the virus' spike-glycoprotein that attacks host cells to infect the patient. To carry out the gene sequence analysis tasks, TRON needed to extend its computational capacity. They acquired Intel® Server System nodes built with 2nd Gen Intel Xeon® Scalable processors to run their CoVigator genome analysis pipelines. The new cluster allowed them to analyze nearly 2 million virus genomes and over 30,000 virus genome sequencing datasets, discovering many variants of the spike protein.
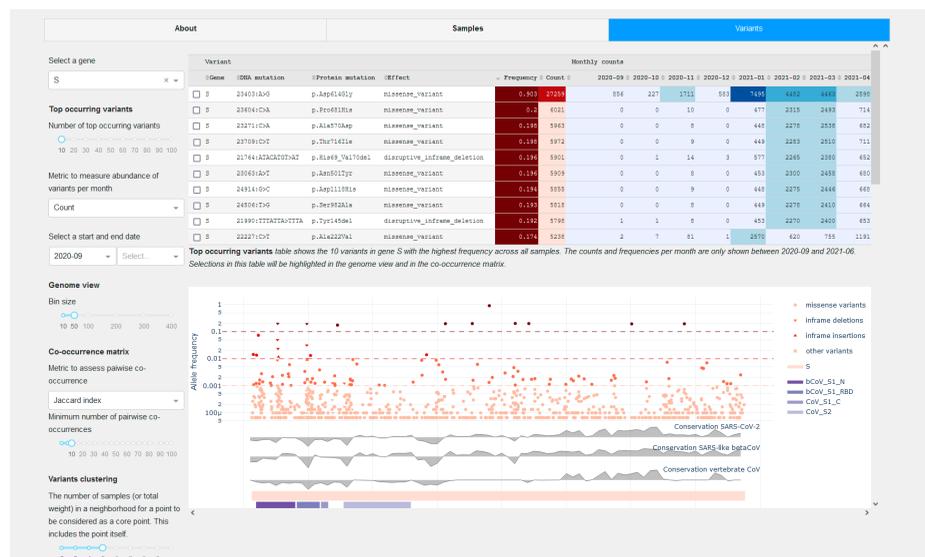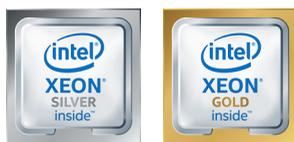
The CoVigator dashboard provides a snapshot view of variants and other key data related to the COVID virus. (Image courtesy of Thomas Bukur, TRON)

Challenge

TRON is a nonprofit research organization established as an independent spin-off of the University Medical Center of the Johannes Gutenberg University Mainz

(Germany). TRON (for Translational Oncology) bridges the gap between basic research done by academic scientists and applied therapeutic solutions created by pharmaceutical companies. Their research has been specific mainly to cancers.

"For the last ten years, our focus has been on understanding immuno-biology of cancers," explained Martin Löwer, Deputy Director of the Biomarker Development Center (BDC) at TRON. "The human immune system can recognize tumors, but tumors can evolve to evade the immune system. We wanted to study this and understand how to modify and stimulate the immune system to recognize and neutralize tumors, essentially helping to find ways to vaccinate against cancers."

Analyzing genomes is central to TRON's Biomarker Development Center, where they study tumor genomes and their evolution to look for specific features in the DNA, called biomarkers. A genetic biomarker can indicate early presence of a tumor before it shows clinical symptoms. Molecular biomarkers can also suggest a particular therapy for an individual patient. Biomarkers are important to understanding the evolution of a disease and how individualized medicine can affect the disease in different patients.

> *With a new cluster of servers built on 2nd Gen Intel Xeon Scalable processors, TRON continues to analyze existing and new samples for viral mutations. Identifying and revealing the many variants gives researchers data on how the virus changes and the impact these changes might have on vaccines.*

As COVID-19 became a worldwide pandemic, millions of people around the globe became hosts where the SARS-CoV-2 virus evolved through mutation. Simultaneously, countries began sequencing the SARS-CoV-2 genome from thousands of patients and building large genome dataset repositories for study.

Mutational variants of a disease are of considerable interest to health professionals and scientists, especially as vaccines are developed to fight against the spread of the disease. Like with cancers, TRON scientists were interested in how the evolution of the SARS-CoV-2 virus might impact the efficacy of the vaccines as they began to appear. So, they applied their knowledge and expertise to studying the SARS-CoV-2 genome and specifically the spike-glycoprotein (spike protein) that the virus uses to invade human host cells.

Looking for variants in a genome is a complex process that involves a lot of computation. The SARS-CoV-2 virus comprises 30,000 base pairs (compared to the human genome's 3.2 billion base pairs) in a sequence. The sequencing data provided by Next-Generation Sequencing (NGS) instruments is a compilation of short segments of the whole genome. Genome data repositories provide both whole genome assemblies and genomic datasets that must first be aligned, much like putting together pieces from a 30,000-piece puzzle. Once aligned, the analysis looks at differences between a reference genome and the sample being studied to find and mark the variants. The variants are further filtered to select the ones of interest and remove what is considered 'noise'.

Today, assembly, conversion, and analytics is a highly parallelized process that can be done quickly by scaling out to many High Performance Computing (HPC) servers. But, TRON's computational facilities are constantly utilized by scientists studying cancers. To complete a study of SARS-CoV-2 spike protein variants and provide a research tool for scientists worldwide would require more computing resources than they had available.

Solution

Working with primeLine Solutions and Intel's Pandemic Response Technology Initiative, TRON was able to acquire ten new Intel Server Systems nodes. These nodes are built on 2nd Gen Intel Xeon Gold 6240R processors with Intel SSD S4610 drives. A log-in server uses Intel Xeon Silver 4208 processors.

The new system gave TRON 960 dedicated threads to run the many tasks for their Corona Virus Navigator NGS (CoVigator NGS) genome alignment and analytical pipelines. TRON is able to analyze and process more than 20,000 sequencing datasets in less than three hours, providing near-real-time-analysis of the constantly growing publicly available data sets.

The CoVigator-NGS-pipeline includes trimming, alignment, variant calling, and other tasks. TRON's pipelines comprise open source tools from many genomic software repositories, including The Broad Genome Analysis Toolkit (GATK), which has been optimized for Intel architecture, BCFtools, LoFreq, and iVar. The pipelines are available to other scientists through GitHub at https://github.com/TRON-Bioinformatics.

With the new computing capabilities, TRON began its initial study of 146,917 genome assemblies and 2,393 NGS datasets that required alignment to detect non-synonymous spike protein mutations. Their research identified different types of variants, including recurrent versus individual, clonal versus sub-clonal, and those hitting T-cell or antibody target sites versus those not hitting these targets. These are all variants useful to researchers and therapy developers.

The results of the study reveal that a percentage of spike protein variants has appeared across the sample population (including many mutations occurring simultaneously within individual patients). Additionally, and importantly, the study reveals that the mutations increased over time. Thus, it is critical that SARS-CoV-2 genome samples continue to be analyzed and the effects of the mutation on vaccines' effectiveness be studied. TRON's initial research has been released as a bioRxiv pre-print.

Result

"The most interesting finding from the research," added Löwer, "was seeing the many variants of the spike protein. Secondly, was the ability to go back over the last year and a half and track exactly how it evolved. We are able to detect even small changes early in its evolution. We can see from how it starts in a single patient and mutates within the patient to several variants and across populations and geographical regions. And because the virus continues to travel around the world, we see how mutations move across the globe over time."

TRON's study was only the beginning of an effort to first understand and then help scientists monitor and analyze the evolution of the virus. With their new cluster, TRON developed a platform for mutation detection and a web-based dashboard for scientists to navigate the database of spike protein variants.

"As a project," stated Thomas Bukur of TRON, a scientist managing the CoVigator project, "without the new servers, we would have only been able to complete the initial study. But we would not be able to provide an ongoing study and service that analyzes millions of samples and identifies spike protein variants. With this platform, we are able to keep the work going."

The TRON CoVigator service delivers a comprehensive overview of temporal and spatial distribution of mutations of the SARS-CoV-2 virus spike protein. The pipelines are run repeatedly to update the database. The CoVigator offers the research community a decisive tool to reveal the virus' evolution.

Understanding mutations in new and known viruses is critical to be able to address and continually manage the therapeutic response to widespread and dangerous illnesses. There will certainly be other pandemics. And, with climate change, scientists are watching the migration of dangerous tropical diseases out of equatorial regions into more temperate zones. These threats will create new challenges for healthcare. The workflows and pipelines TRON scientists created can be rapidly adapted to new viruses and strains of viruses, offering new tools for collaborative immuno-biology research and response.

Solution Summary

TRON applied its expertise in cancer immuno-biology research to the Covid-19 pandemic by analyzing hundreds of thousands of SARS-CoV-2 genome samples for variants in its spike protein. Their initial research was published in a paper preprinted by bioRxviv. With a new cluster of servers built on 2nd Gen Intel Xeon Scalable processors, they continue to analyze existing and new samples for viral mutations. Identifying and revealing the many variants gives researchers data on how the virus changes and the impact these changes might have on vaccines. TRON makes both their analytical pipelines available (on GitHub) and the database of variants searchable through its TRON CoVigator web service.

Where to Get More Information

Learn more about TRON:

Access the TRON pipelines at https://github.com/TRON-Bioinformatics.

Access CoVigator at covigator.tron-mainz.de

Find out more about primeLine Solutions and Intel servers.

Explore the capabilities of the 2nd and 3rd Generation Intel Xeon Scalable processors with integrated Intel® Deep Learning Boost capabilities for accelerated AI inferencing.

Solution Ingredients

- 10 Intel Server System R1208WFTYSR computational nodes
- 2nd Gen Intel Xeon Gold 6240R processors (480 cores and 960 threads)
- 2nd Gen Intel Xeon Silver 4208 processors for log-in node
- Intel SSD S4610 480 GB drives

intel.