PINPOINTING MUTATIONS FOR PRECISION MEDICINE

UNDERSTANDING THE HUMAN BODY AT THE GENE LEVEL HAS THE POTENTIAL TO REVOLUTIONIZE HEALTHCARE, ALLOWING DOCTORS TO BETTER UNDERSTAND DISEASES AND MORE ACCURATELY APPLY TREATMENTS.

BACKGROUND: TARGETED THERAPIES — Using genomics to choose the right therapies for the right patients.

In the simplest terms, genomics is the study of genes and their functions. Through genomics, scientists learn how genes interrelate and influence the development of an organism. But genomics starts with the genome — the complete set of genetic instructions for building, running and maintaining an organism.

The human genome is made up of 3.2 billion bases of DNA and contains between 22,000 and 25,000 genes. It's an immense amount of information and makes understanding the genetic basis of a disease a highly data-intensive task. To put it in context, if all 3.2 billion letters in a human genome were printed out, they would fill a stack of paperbacks 200 feet high.

The process starts with sequencing each patient's genome to determine the exact order of the bases in their DNA. With today's technology, this task is relatively fast and cheap to perform. But once a genome is sequenced, scientists must still figure out what it all means.

For example, why might a drug work on one person but not another?

The complete data analysis process is highly complex and involves multiple steps and tools to handle the large amounts of data. Variant calling is an important step in this process.

Identifying mutations is the first step to understanding what they mean and how they might affect a patient.

As the name suggests, variant calling identifies variants — or mutations — in DNA.

Identifying mutations is the first step to understanding what they mean and how they might affect a patient.

CLEARING COMPUTING ROADBLOCKS FREES SCIENTISTS TO FOCUS ON SCIENCE … AND BEGIN TO IDENTIFY AND UNDERSTAND GENE MUTATIONS.

USE CASE: Genetic Variant Analysis on the Urika®-GX Platform

In one recent project, a research team from the Broad Institute and Cray using a Cray® Urika®-GX agile analytics platform was able to significantly speed up workflows in Hail, an open-source, distributed compute framework for analyzing genetic data.

Hail is a scalable framework for exploring and analyzing genetic data at massive scale. Built on top of Apache Spark™, Hail can analyze terabyte-scale genetic data. Still under active development, Hail is used in medical and population genomics at the Broad for a variety of diseases. It also serves as the core analysis platform for the Genome Aggregation Database (gnomAD) — the largest publicly available collection of human DNA sequencing data, and a critical resource for the interpretation of disease-causing genetic changes.

Having collaborated successfully in the past, Cray and the Broad teamed up to test Hail on the Urika-GX agile analytics platform.

The Urika-GX platform fuses supercomputing with an open, enterprise framework giving it the speed and agility to handle a variety of workloads. Among its abilities, the Urika-GX system runs Hadoop®, Spark, graph and HPC analytics workloads concurrently.

SOLUTION

Cray® Urika®-GX agile analytics platform

SYSTEM DETAILS

Nodes: 32

Processor Cores: 1,024

Memory: 8 TB

SSD: 22 TB

Local Disk: 128 TB

Network: Aries™ Interconnect

As for speed and parallelization, both are built into the architecture with a combination of up to 35 TB of PCIe SSD on-node memory and 22 TB of DRAM for deep local memory hierarchy. For the Hail team, the Aries™ interconnect in particular was the "secret sauce." Aries enables large-scale, memory-intensive workloads. Additionally, the system features an open framework for easy customization and simple, standards-based tools for straightforward management.

The 1,024-core test system gave the Hail team 700 cores — quadruple their previous number. Runtimes improved significantly as a result; for example, whereas a gene association analysis on a 150-core Spark cluster ran in 43 minutes, the same analysis finished in only 9 minutes on the Urika-GX system.

The team also reported scale and productivity gains, with growth in scale of data analytics on genomics projects.

The gnomAD genomics efforts also benefited from the Urika-GX system's performance. Running Hail on the Urika-GX platform, the gnomAD effort analyzed 126,216 and 15,135 whole genomes, and has identified more than 270 million variants, of which over 160 million are novel.

Because of the power and flexibility of the Urika-GX platform, the Broad's success with their Hail workflows can be easily generalized to other software tools, such as GATK and other large datasets. The Urika-GX platform's analytics power and unmatched graph computing capability can be used to further analyze and contextualize the results of any workflow, speeding time-to-insight and paving the way for better understanding of disease and improved patient outcomes.

ABOUT CRAY

Cray provides systems and solutions that help solve the most difficult computing, storage and data analytics challenges. Our portfolio includes powerful supercomputers, optimized cluster systems, advanced storage systems and data analytics and discovery platforms.

SAMPLE RESULTS: HAIL

Reduced runtime from 43 minutes to 9 minutes

5x application speedup

Growth in scale of data analytics on genomics projects

Software development team more productive

CRAY® URIKA®-GX BENEFITS FOR GENOMICS

- Shortened path to better decisions and patient outcomes
- Improved responsiveness and agility
- Increased power and flexibility
- Reduced infrastructure TCO
- Stronger cybersecurity and surveillance

Cray Inc.
901 Fifth Avenue, Suite 1000
Seattle, WA 98164
Tel: 206.701.2000
Fax: 206.701.2500
WWW.CRAY.COM