Operating Largest Genomic Analysis Project in the Cloud

CASE STUDY:

Baylor College of Medicine, Human Genome Sequencing Center & the CHARGE Consortium

> *"The use of cloud computing and collaboration with DNAnexus is allowing us to achieve our goals faster and in a more cost-effective manner."*
>
> **Eric Boerwinkle, Ph.D.**
> *Director at Human Genome Sequencing Center, Baylor College of Medicine*

THE CUSTOMER

The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium was formed to better understand how human genetics contributes to heart disease and aging. The discoveries CHARGE makes will be instrumental in understanding disease and aging in mechanistic detail, enabling the development of new medical interventions and analysis tools.

Baylor College of Medicine in Houston, Texas is home to The Human Genome Sequencing Center (HGSC), which is on the forefront of technical innovation in genomics-based medicine. As the organization chosen to sequence and analyze the genomes of more than 14,000 CHARGE participants, HGSC found that their local compute and storage infrastructure was not equipped for the additional computational load.

SUMMARY

INSTITUTIONS

Baylor College of Medicine, Human Genome Sequencing Center

The CHARGE Consortium

WEBSITES

www.hgsc.bcm.edu
www.chargeconsortium.com

INDUSTRY

Translational Research

CHALLENGE

Scale up ultra large-scale genomic analysis project quickly & efficiently

SOLUTION

The DNAnexus Platform

RESULTS

- Processed 3,751 whole genomes & 10,940 exomes using 3.3 M core-hours generating 430 TB of results
- Enabled HGSC & 300 investigators to upload, analyze, and collaborate on results quickly
- Completed job 5.7x faster than onpremise cluster
- Provided peak capacity of 20,800 CPU cores on-demand

The CHARGE Consortium involves more than 300 researchers across five institutions around the world who are collecting genetic samples from global studies and running them through HGSC's Mercury variant-calling pipeline in order to identify genetic variants that may contribute to heart disease and aging.

Mercury is a semi-automated and modular set of tools for the analysis of next-generation sequencing data in clinically focused studies. HGSC designed the pipeline to identify genomic mutations and determine whether they have serious disease-causing effects.

THE CHALLENGE

The goal of HGSC was to quickly analyze the sequence data of more than 14,000 genomes and make the results available to CHARGE investigators. Already impacted by internal production of 25 TB of sequence data per month, HGSC's options for genome data analysis included quadrupling their current compute core capacity with new hardware for this short-term project, or jamming the cluster, which would force the suspension of the Center's other research projects for 3-4 weeks. The second major challenge to be overcome was the delivery of more than 430 TB of genomic data to more than 300 CHARGE investigators, which would be cost and time prohibitive.

THE SOLUTION

Cloud-based solutions provide a unique environment where infrastructure management systems can be built on top of them for specific applications. To address their need for greater computational capacity, HGSC partnered with DNAnexus to design a cloud-based computational infrastructure to analyze the CHARGE data in a secure, cost efficient manner while enabling worldwide collaboration without the delays inherent in setting up a physical infrastructure.

In the DNAnexus platform implementation for HGSC, each element in the Mercury pipeline was imported as an "app." Each app could then run independently in the cloud, utilizing different CPU, RMA, disk and bandwidth resources.

THE RESULTS

As part of its participation in the CHARGE Consortium, HGSC utilized the DNAnexus platform to analyze the genomes of over 14,000 people, encompassing 3,751 whole genomes and 10,940 exomes, using the bioinformatics building blocks of their own Mercury pipeline. Over the course of a four-week period approximately 3.3 million core-hours of computational time were used, generating 430 TB of results and nearly 1 PB of data storage hosted for further analysis. The job was completed 5.7x faster than could have been accomplished using the local cluster.

At the project's peak, HGSC was able to spin up 20,800 cores on-demand. During this period, HGSC was running one of the largest genomic analysis clusters in the world, without any capital investment or sacrifice of their local cluster capacity.

DNAnexus

DNAnexus combines expertise in cloud computing and bioinformatics to create the global network for genomic medicine. DNAnexus provides security, scalability, and collaboration for enterprises and organizations that are pursuing genomic-based approaches to health in order to accelerate medical discovery. DNAnexus is supporting customers around the world that are tackling some of the most challenging and exciting opportunites in human health.