Transforming Large Scale Genomics and Collaborative Research

PRESENTED BY

intel

Bio·IT World

INTRODUCTION

For years, many aspects of life sciences work have required the fast analysis of large datasets. And to accomplish this, most organizations installed high performance computer and storage infrastructures.

However, the greater volumes of data being produced with newer sequencing equipment, the need to run that data through more complex analysis, and the demand for faster time to results are straining the IT infrastructures in place in most organizations today. These problems are more pronounced as organizations make growing use of translational and precision medicine applications that use genomic analysis in research and clinical settings.

To achieve the vision of precision medicine and advance science, researchers and the genomics community at large face many challenges. As sequencing data begins to explode in volume, the ability to transform this data into biomedical insights will require a cost-effective, open standards based computational infrastructure that is built for extreme scalability. Infrastructure and tools will need to be optimized accordingly to execute, store and process massively large genomic workflows.

> "Sequencing just one cancer patient's genome can create up to 1 terabyte of data."

CHALLENGES ABOUND

Since the first sequencing of the human genome, the way to handle the analysis and storage issues related to sequencing and data analysis in the life sciences was to simply throw raw compute and storage capacity at the problem. But that approach no longer works. Besides dealing with capacity challenges, life sciences organizations must also deal with performance and data management issues when it comes to their choice in IT infrastructures.

A big factor in the field relates to large amounts of data that is generated per patient. As the cost of DNA sequencing continues to drop and molecular imaging becomes more commonplace, public and private databases with associated molecular and clinical outcome data continue to grow. These individual data sets can become extremely large; sequencing just one cancer patient's genome can create up to 1 terabyte of data.[1]

Today, all of the data associated with a patient is being collected and stored in silos. Ninety six percent of cancer patient data is locked away at individual hospitals and cancer institutions, with only four percent of cancer patients participating in clinical trials, where data is collected systematically and made available for research.[2]

Data copying is a prerequisite for even sharing that four percent of clinical trial data today. As a result, the greater genomics community must shoulder the cost of storing multiple copies of massive datasets. While there is no shortage of overall data, pooling, accessing and analyzing all of these large, siloed data sets across research institutions has not been feasible. The large size of the data, as well as privacy concerns, makes it very difficult for data to be pooled together for joint analysis. A simple example illustrates the challenges faced. A single patient's genomic data often exceeds one terabyte and its use must meet stringent federal HIPAA requirements.

In addition to providing secure sharing and access of data, dramatic advances in analytics will also be needed to help clinicians and researchers sift through enormous volumes of genomic, imaging and clinical data in order to understand the root causes of disease like cancer and create effective treatment plans that comprehend the unique characteristics of each patient's individual form of the disease.

Furthermore, translational and precision medicine work being done at leading life sciences organizations is taking analysis to new levels. Going beyond the requirements of traditional genetic analysis, some organizations are increasingly adopting Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) to look for discrete genetic errors at the root of a patient's disorder across the 23,000 genes in the human genome.

In all of these efforts, the volumes of data keep growing, while the time to perform things like WES and whole genome analysis must be reduced.

> "Translational and precision medicine are taking analysis to new levels."

WHAT'S NEEDED?

To advance the understanding of the genetic causes of diseases, leverage genomic analysis in clinical settings, and develop personalize patient treatments, researchers need new tools so large genomic workflows can run at cloud scale in very fast times.

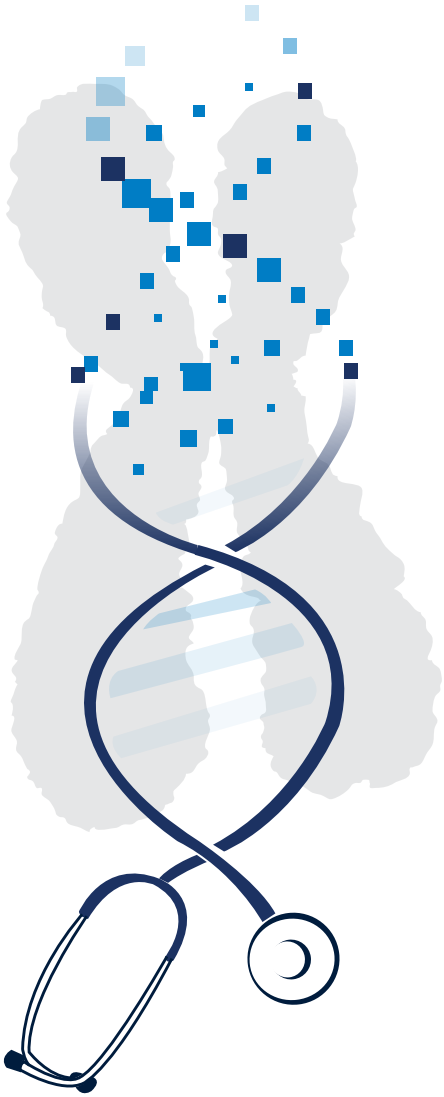
Additionally, due to the multidisciplinary nature of life sciences research, data and analysis tools must be made available to a variety of geographically dispersed researchers. Rather than copying the large datasets and setting up the tools in every location for every researcher, a more collaborative access approach is needed.

One way to address these challenges is to use cloud technology. That is exactly what Intel, in collaboration with leading life sciences institutes, is doing. To that end, Intel is committed to building a platform consisting of tools that help physicians and researchers solve challenges in the precision medicine field and lower the barriers for innovation.

An example of bringing this platform to life is Intel's work with Oregon Health & Science University (OHSU) in building a computational foundation with the goal of accelerating precision medicine for cancer. Announced last year, the Collaborative Cancer Cloud (CCC) is an advanced, federated analytics solution that allows scientists and physicians to securely share and study large amounts of data across distributed sites – all while protecting patient privacy and preserving an institution's intellectual property that may result from future discoveries.[3]

CCC combines next generation Intel technologies and bio-science advancements to enable solutions that make it easier, faster, and more affordable for developers, researchers, and clinicians to understand any disease that has a genetic component, starting with cancer. The end goal is to empower researchers and doctors to help patients receive a diagnosis based on their genome and potentially arm clinicians with the data needed for a targeted treatment plan. By 2020, the organizations envision this happening in 24 hours -- All in One Day.

The CCC's approach uniquely addresses four technical challenges to help accelerate collaborative analytics. It keeps data local at a clinical site or research institution. It accelerates the speed of imaging and omics analytics.



It allows researchers to perform secure joint computations across multiple data sites without compromising local privacy. And CCC scales up to meet the storage requirements of processing very large patient data sets.

The CCC delivers the performance, storage, security and scalability to access, analyze and share large volumes of patient genomic, imaging and clinical data in a federated model, ushering in a new era of secure, collaborative analytics to accelerate potentially lifesaving discoveries.

OHSU and Intel have expanded participation in the CCC to include Dana-Farber Cancer Institute and Ontario Institute for Cancer Research. These cancer institutions will use Intel's technology to securely share and analyze their collectively large amounts of data in a distributed way, while still preserving the privacy and security of the patient data at each site.

The collaboration will initially focus on developing genomic pilot projects based on leading industry standard tools. They will identify novel analytics approaches using machine learning techniques against a collective set of molecular and imaging data in order to support big data analytics in a federated, aligned environment.

Dana-Farber Cancer Institute and the Ontario Institute for Cancer Research bring rich data to the table to advance precision medicine and research. Dana-Farber's Profile project includes one of the world's largest databases for genetic abnormalities that drive cancer, with over 15,000 genetic profiles of patients' tumors, adding about 400 each month to the database. The Ontario Institute of Cancer Research offers the largest repository of cancer genomes in Canada. The institute hosts the Secretariat and Data Coordination Center for the International Cancer Genome Consortium that comprises of molecular data from over 14,700 donors and close to 37 million simple somatic mutations from 66 cancer projects.

The long-term goals are to open this federated, secure CCC platform to dozens of other institutions, accelerating the ability of clinicians and researchers around the world to understand the root causes of cancer and develop targeted, molecular treatments. In the future, these underlying technologies may be applied to cardiovascular diseases and neurological disorders, among others, to accelerate scientific discovery, yield new insights and inform treatment plans. Ultimately, the collaboration should reduce "big data" barriers and help make precision medicine widely available to patients.

> "CCC combines next generation Intel technologies and bio-science advancements."

ADVANCED ANALYSIS TOOLS A MUST

Intel is also working with other institutes to develop tools focused on large-scale genome sequencing to help accelerate personalized care.

Specifically, Intel is working with the Broad Institute, one of the world's leaders in biomedical research, on the co-development of such tools and fundamental capabilities. In a joint effort, the organizations are revolutionizing how genomic workflows get executed, stored and processed at cloud scale.

The first set of tools consists of **Cromwell and GenomicsDB**.

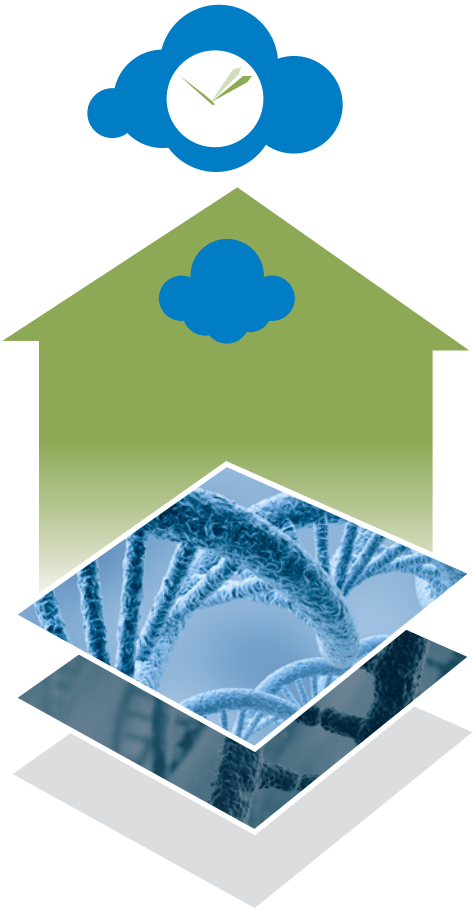
Orchestrating genomic workflows at cloud scale is complex. Broad Institute's workflow execution engine called **Cromwell** is an open, standards based integrated workflow orchestration engine built and optimized for distributing/federating tasks across private, public and hybrid clouds. It enables complex computational workflows to be dispatched and coordinated across different

sites. Complex workflows are simplified, more portable, reproducible, and can be run across any cloud environment. Broad is working with Intel to extend Cromwell's capabilities so that the workflow orchestration engine supports multiple input languages and execute on multiple back-ends simultaneously, enabling researchers to run jobs anywhere. This integrated workflow engine knows how to find the best way to execute the tasks, and what hardware to run it on. If the workload has been run before, it will be cached so it runs faster.

Efficiently storing and computing on massive amounts of genomic data is not possible today. It is extremely difficult to store and process massive amounts of genomic data efficiently when data is stored as flat files. What researchers need is a genomic database that can efficiently store variant-level information (VCF data). **GenomicsDB** is a revolutionary way to store and process massive genomic data in private, public and hybrid clouds. This novel array-oriented database enables genomic analytics to be performed directly on the variant data in the database, without having to use large files. This database technology is an open, standards based database engine that is built and optimized for the storing and processing of sparse matrix data perfect for 'omics data such as genomic, metabolomic and microbiomic data. GenomicsDB is now being used in the Broad's production pipeline to perform joint genotyping. Engineers at the broad running GATK's variant discovery analysis on ~100 whole genomes saw dramatic reductions in the time it takes to run this, from 8 days to 18 hours.[4]

Both Cromwell and the GenomicsDB have been integrated and optimized to work with the BROAD Institutes Genomics Analysis Toolkit (GATK), enabling portability, reproducibility and cloud integration of this genomic pipeline along with the ability to upload this pipeline directly into GenomicsDB for efficient storage and in-database computing. GATK is today's industry standard open software package built for high throughput sequencing data. The toolkit offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine and high-performance computing features make it capable of taking on projects of any size.

> "Researchers need new tools so large genomic workflows can run at cloud scale in very fast times."



SUMMARY

Precision medicine – taking into account individual differences in people's genes, environments and lifestyles – is one of the biggest of the big data problems and is on the cusp of a remarkable transformation in medicine.

However, deriving results requires a secure way to share and provide access to the large volumes of genomic data about each patient. Additionally, getting results in a time frame suitable to take action requires analysis tools that take advantage of high performance compute capabilities.

These are areas where Intel is working with leading life sciences organizations to advance the field. Intel's efforts with its partners will one day allow any researcher and clinician to conduct automated analysis of very large datasets spread out over multiple centers.

To learn more about these efforts and to start using these tools, please visit http://github.com/intel-hls.

REFERENCES

1 For optimum results Illumina recommends a minimum coverage of 30x for normal tissue and 60x coverage for tumor samples. http://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/has/v2-0/hiseq-analysis-software-v2-0-user-guide-15070536-a.pdf. A FASTQ file containing all the normal tissue reads for whole genome sequencing at 30x coverage er is approximately 2600GB of data, and at 60x coverage for the tumor sample it would be approximately 400gb. At that size, if the patient is sequenced at least 2 times per year to capture changes in mutations, they would generate at least 1.2TB per year. https://medium.com/precision-medicine/how-big-is-the-human-genome-e90caa3409b0#.77vgcdvg6

2 According to an article in the Wall Street Journal, citing Google Ventures, less than 4% of US Cancer patients are in clinical trials where data is made available for research, meaning that information about treatments and patient outcomes for 96% of patients who receive standard of care is not easily accessible. http://blogs.wsj.com/venturecapital/2014/05/07/google-ventures-leads-130m-round-for-big-data-medical-software-company-flatiron-health/

3 Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

4 http://genomicinfo.broadinstitute.org/acton/media/13431/broad-intel-collaboration

> "Precision medicine is on the cusp of a remarkable transformation in medicine."