UC San Diego
UC San Diego News Center

March 20, 2012 | By Jan Zverina

SDSC's "Big Data" Expertise Aiding Genomics Research

Focus on Genomics Medicine is Growing, says SDSC's Norman

ctor Michael Norman Photo: Ben Tolo

The San Diego Supercomputer Center (SDSC) at the University of California, San Diego, has in the last three years undergone a major reboot, remaking itself into a center of expertise on all aspects of "big data" research including genomics, one of the fastest growing areas of scientific study.

"We have in recent years become a lot more than a supercomputer center," SDSC Director Michael Norman told attendees earlier this month at the third annual X-Gen Congress & Expo, a four-day event focused on exploring the potential of established and emerging genomic technologies. "Our real expertise is now in all aspects of 'big data', which includes data integration, performance modeling, data mining, software development, workflow automation, and more. We believe that data-enabled science is the beginning of a new scientific era."

SDSC's creation of a fully integrated "big data" environment already has led to several projects in the study of genes, and more are underway. "Our focus in genomic medicine is growing," said Norman.

"Next-generation sequencing of DNA and RNA are profoundly transforming biology and medicine, providing insight into our origins and diseases," according to Wayne Pfeiffer, a distinguished scientist at SDSC. "However, obtaining that insight from the sequencer data deluge requires complex software and increasingly powerful computers."

SDSC has an expanding repertoire of “big data” systems, the latest being *Gordon*, a unique flash memory-based supercomputer that is capable of storing 100,000 entire human genomes, while operating hundreds of times faster than conventional computers to study genetic data.

Genetic data creates many additional requirements regarding sharing and computing. The iDASH center (integrating Data for Analysis, Anonymization, and Sharing), under the leadership of Lucila Ohno-Machado, is the most recent National Center for Biomedical Computing funded under the National Institutes of Health (NIH) Roadmap for Bioinformatics and Computational Biology.

Conceived as a collaborative computational environment to improve access to health data and software, iDASH provides biomedical and behavior scientists with access to a sophisticated, secure privacy-preserving infrastructure to contribute, integrate, and analyze their data, as well as potentially reuse data from others (given permissions set up by data contributors) and leverage other research results.

“The iDASH center addresses fundamental challenges to research progress by providing a secure, privacy-preserving computational environment in which researchers can analyze molecular, clinical, and behavioral data,” said Ohno-Machado.

SDSC also has multiple collaborations with the Scripps Translational Science Institute (STSI), which has a dedicated 1gigabit-per-second (Gb/s) network connection to the center, along with 140 terabytes of online project storage. STSI has purchased time on SDSC’s *Triton Resource* to conduct research on a number of projects.

One such collaboration is called the Human Tumor Study, or HuTS, which is using SDSC’s *Triton Resource* to search for genome variants between blood and tumor tissue. Software used in this project includes the Genome Analysis Toolkit (GATK), the SOAPdenovo assembler, and various aligners such as ATAC, BLAT, and BWA.

Another collaboration involving SDSC, STSI, and others is called W115. In this project, Pfeiffer is using the Velvet and ABySS assemblers and the ATAC and BFAST aligners on the *Triton Resource* to study the full genome sequence of a 115 year-old woman to determine how many mutations occur in a long, healthy lifetime.

Further collaborations between SDSC and other genomic institutions including STSI are expected, said Norman, noting that *Gordon* and its data storage facilities have the bandwidth needed for such research. “The end goal here is to develop a rapid learning system for guiding individual therapies, and SDSC is now set up assist in reaching that goal,” said Norman.

Not Just Supercomputers

In addition to *Gordon*, which went into production earlier this year, SDSC operates *Trestles*, designed to enable modest-scale and gateway researchers to be as computationally productive as possible, and the *Triton Resource*, a medium-sized data-intensive compute cluster primarily for UC San Diego and UC researchers.

All three computer systems, for example, are integrated into four tiers of specialized data storage, which is crucial for genomics and other researchers who need to sift through massive amounts of data. SDSC's data storage facilities include:

- *Data Oasis*, a high-performance Lustre-based parallel file system with four petabytes of storage and a 100 gigabyte-per-second (GB/s) connection for scratch and medium-term storage.
- *Gordon's* 300 terabytes of flash-based solid state drive memory. Like *Data Oasis*, this is used for fast random access and fast sequential access.
- *Project Storage*, which provides academic and research partners a network-based storage service offering Common Internet File System (CIFS) and Network File System (NFS) storage to SDSC and UC San Diego systems. With transfer rates up to 1GB/s, Project Storage is an excellent option for interactive access and use as a traditional mounted file system.
- *SDSC Cloud*, a multi-platform, fully accessible and scalable disk-based cloud storage system with 5.5 petabytes of raw storage and more than two petabytes of formatted dual copy storage for archiving or sharing. One petabyte equals a quadrillion (1,000 trillion) bytes of information. The *SDSC Cloud* is believed to be the largest academic-based cloud storage system in the U.S.

SDSC's four tiers of specialized storage are all interconnected for use as needed. "One can access any storage from any of these systems and build workflows that hop from one system to another," said Norman.

SDSC's *Gordon* and *Trestles* systems and their storage systems are available for use to any researcher or educator at a U.S.-based institution and not-for-profit research through the National Science Foundation's (NSF) Extreme Science and Engineering Discovery Environment, or XSEDE program. Industry-based research time and storage is also available. Industry researchers interested in using SDSC's resources or expertise should contact Ron Hawkins at rhawkins@sdsc.edu or 858 534-5045.

The X-Gen Congress & Expo was held March 5-8 in San Diego.

MEDIA CONTACT

Warren R. Froelich, 858 822-3622, froelich@sdsc.edu

UC San Diego's Studio Ten 300 offers radio and television connections for media interviews with our faculty, which can be coordinated via studio@ucsd.edu. To connect with a UC San Diego faculty expert on relevant issues and trending news stories, visit https://ucsdnews.ucsd.edu/media-resources/faculty-experts.