UC San Diego
UC San Diego News Center

March 20, 2014 | By Jan Zverina

# SDSC Assists in Whole-Genome Sequencing Analysis Under Collaboration with Janssen

## Collaboration uses large flash memory-based cluster to further study rheumatoid arthritis

A recent whole-genome sequencing (WGS) analysis project supported by the San Diego Supercomputer Center (SDSC) at the University of California, San Diego has demonstrated the effectiveness of innovative applications of “flash” memory technology to rapidly process large data sets that are pervasive throughout human genomics research.

Janssen Research and Development, LLC (Janssen), in collaboration with SDSC and the Scripps Translational Science Institute (STSI), recently launched a project to conduct whole-genome sequencing of 438 patients with rheumatoid arthritis to better understand the disease, as well as explore genetic factors of patient response to a biologic therapy discovered, developed, and currently marketed by Janssen in the United States.

The analysis began with 50 terabytes of “read” data generated by DNA sequencers from samples originally obtained from each of the study participants. These source data were fed into a 14-step processing “pipeline” using open source software tools. Key components of the analysis were mapping the DNA read sequences from each patient against a reference genome and calling to identify the variants between the two.

The read mapping and variant calling were done by Kristopher Standish, a UC San Diego graduate student working under Nicholas Schork, formerly with STSI and now with the J. Craig Venter Institute. SDSC provided high-performance computing and storage resources, as well as expertise to set up and optimize the computational pipeline.

“The need to conduct analysis of 438 full human genomes in a relatively short timeframe necessitated a thorough understanding not only of the computational workload, but of the memory, storage, and input/output requirements,” said Wayne Pfeiffer, an SDSC Distinguished Scientist and the Center’s lead researcher in the collaboration. “The emergence of ‘big data’ challenges such as those in human genomics has brought to the fore situations where computer analyses are more likely memory-and I/O (input/output)-bound than compute-bound, meaning that while the actual computer processors may have plenty of capacity, the ability to store and/or move around large amounts of data becomes the limiting factor in throughput."

In the case of the Janssen collaboration, one step in particular – the "sort" step of the read mapping stage – was particularly challenging, requiring a relatively small number of processor cores, but rapid access to several terabytes of data, more than can be kept in the supercomputer's high performance main memory. The conventional approach of storing data on hard disk drives during the sort step resulted in a severely I/O-bound situation, dramatically limiting throughput.

"The solution was to take advantage of *Gordon's* flash memory, which provides much higher speed than conventional disk drives for the random access I/O operations of the sort step," said Pfeiffer. "Several terabytes of flash were aggregated into what we call "BigFlash" nodes, which significantly reduced the I/O bottleneck in this step and contributed to helping researchers meet the project's timelines."

"The bulk of the analysis was completed in six weeks (including learning time on *Gordon*) using more than 300,000 core hours of computer time," said Glenn K. Lockwood, a user services consultant at SDSC. "That analysis would have taken more than four years of 24/7 compute time on an 8-core workstation."

The collaboration also demonstrated the need for large-scale, high-performance computing resources when analyzing hundreds of human genomes in constrained timeframes. With 340 teraflops of computing power, 64 terabytes of main memory, and 300 terabytes of flash memory, *Gordon* ranked among the 50 fastest supercomputers in the world when it debuted in late 2011, according to the Top500 list.

According to Lockwood, at the project's peak throughput, the WGS pipeline was using 350 terabytes of storage on SDSC's high-performance storage system and 5,000 processor cores representing 30 percent of the system capacity.

"The Janssen collaboration validated our vision for the *Gordon* system," said Michael Norman, SDSC's director and principal investigator for the *Gordon* project. "We saw that emerging big data challenges such as human genomics would dictate new supercomputer architectures where memory and IOPS (I/O operations per second) would be more important than raw computing power, so we designed the system accordingly."

The *Gordon* supercomputer and other SDSC computational and storage systems are available to industrial collaborators on a space-available basis for conducting research and development. Interested parties should contact Ron Hawkins, director of industry relations.

## MEDIA CONTACT

**Jan Zverina**, 858-534-5111, jzverina@sdsc.edu
**Warren R. Froelich**, 858-822-3622, froelich@sdsc.edu