"If I need to move a huge quantity of data over to that storage, I know it's not going to bring the system to its knees. It's designed precisely for that purpose."

— *David Kinnamon, Ph.D., Director of Human Genetics Research Informatics*

# Data Management

## Kinnamon creates platform to handle troves of genetic info

The more genetics researchers learn about the building blocks of life, the more data they produce. This is a great problem to have—the more they know, and the more detail in which they know it, the better we can treat diseases at the individual level, streamline screening processes and create targeted pharmaceuticals. However, for researchers and genetic counselors, wading through the myriad of genotypes and phenotypic information to get to a specific set of relevant genes or variants is an unnecessary hindrance to the life-changing work they could be doing. Daniel Kinnamon, Ph.D., director of Human Genetics Research Informatics in the Division of Human Genetics at The Ohio State University's Wexner Medical Center, is head of a project that will make their lives easier.

Kinnamon worked directly with staff at the Ohio Supercomputer Center to create a human genetics data management platform that takes advantage of the Center's high performance storage and compute capacity. Most end users of the medical center data only need to view information for a few genes or variants at a time. Searching through a number of large files, many more than 10 gigabytes in size, for the needed information is both infeasible and inefficient.



PHOTO TOP // David Kinnamon, Ph.D., Director of Human Genetics Research Informatics in the Division of Human Genetics at The Ohio State University's Wexner Medical Center

"Alignments are stored in huge files that are not humanly readable," Kinnamon said. "There's an open-source browser for these types of files. We were able to take that software and integrate it into our platform's web interface. Because OSC has high performance storage that we use for our platform's genomic data, we could actually have the web interface submit a request to get just the relevant section of the file off of storage and respond quickly. That allows our end users to interact with these data that otherwise would have been totally inaccessible to them."

The platform comprises two OSC-hosted virtual servers residing on a single physical server. An application server provides a front-end data management and analysis option for users while another server stores and manages data. Small analysis jobs can be performed locally on the application server while larger jobs, e.g., sequence alignment or genome-wide association studies, will be executed on the OSC Oakley or Owens clusters. OSC has also recently provisioned a third virtual server to enable secure access to protected data via the web interface by end users external to the Wexner Medical Center.

Currently, Kinnamon's platform supports 28 different research protocols, including a $12.4 million NHLBI- and NHGRI-funded study on the genetics of dilated cardiomyopathy (DCM), a heart condition in which the left ventricle becomes enlarged and weakened, decreasing the heart's ability to pump blood. Researchers on the project want to identify and characterize genes that cause or predispose an individual to dilated cardiomyopathy, which is a leading cause of heart failure. The project currently has phenotypic data on hundreds of families and exome sequences on hundreds of individuals with DCM within these families. The team is narrowing copious amounts of data down to 35 genes related to DCM. Genetic counselors and molecular geneticists involved in the study need to use information on potentially relevant variants that may exist in several public databases to figure out if they might contribute to disease. Through the use of Kinnamon's platform, they can view this information and curate variants quickly through a user-friendly web interface without having to know cryptic command line instructions.

"We have exome sequences back on 387 people, and we're expecting to get about 1,500 by early 2020 for just this study, so you can get an idea of how rapidly our storage needs are going to scale," Kinnamon said. "If I need to move a huge quantity of data over to that storage, I know it's not going to bring the system to its knees. It's designed precisely for that purpose. Scalability is the reason we decided to build this entire system at OSC from the get-go." Kinnamon adds that the scalability provided by OSC will be crucial for the next DCM genetics study currently in the planning phase, which will be collecting data on 10,000 patients and their families. •

PROJECT LEAD // Daniel Kinnamon, Ph.D., The Ohio State University Wexner Medical Center RESEARCH TITLE // Division of Human Genetics Data Management Platform FUNDING SOURCES // The Ohio State University Wexner Medical Center and Comprehensive Cancer Center, NHLBI, NHGRI, NCI, Pelotonia WEBSITE // wexnermedical.osu.edu/departments/internal-medicine/genetics/team/daniel-kinnamon-phd; dcmproject.com

PHOTO TOP // Kinnamon's research team views programming and graphics related to their data management platform.