UC Santa Cruz builds national data center for cancer genome research

May 1 2012

The emerging field of "personalized" or "precision" medicine holds great promise in the fight against cancer. If scientists can identify the genetic changes that drive each patient's cancer cells, they can use that information to develop targeted treatments. But achieving this goal will require massive amounts of genomic and clinical data and a sophisticated infrastructure to manage and analyze the data.

The University of California, Santa Cruz, has now completed a first step in building this infrastructure, said UC Santa Cruz bioinformatics expert David Haussler. Haussler's team has established the Cancer Genomics Hub (CGHub), a large-scale data repository and user portal for the National Cancer Institute's cancer genome research programs. CGHub's initial "beta" release is providing cancer researchers with efficient access to a large and rapidly growing store of valuable biomedical data. The project is funded by the National Cancer Institute (NCI) through a $10.3 million subcontract with SAIC-Frederick Inc., the prime contractor for the Frederick National Laboratory for Cancer Research.

"By providing researchers with comprehensive catalogs of the key genomic changes in many major types and subtypes of cancer, these efforts will support the development of more effective ways to diagnose and treat cancer," said Haussler, a distinguished professor of biomolecular engineering in the Baskin School of Engineering at UCSC and a Howard Hughes Medical Institute investigator.

In personalized care, doctors design treatments to target specific genetic

changes found in a patient's cancer cells. Researchers are trying to catalog all the genetic abnormalities found in different types of cancers and find connections between specific genetic changes and how patients respond to different treatments. The scale and complexity of the information being gathered creates a critical challenge in the area of data management.

Although recent studies using genetically targeted treatments have shown promising results, much more research is needed to enable their widespread use, Haussler said. "There won't be one magic bullet, because cancer is not one disease, or even 100 diseases. Every instance of cancer is different. We have to improve our understanding of the molecular biology of cancer and develop computer algorithms so that we can analyze the genetic changes in each individual patient. It will take time. But with cancer genomics, we will eventually know our enemy completely."

Haussler's team assembled the first draft of the human genome sequence in 2000 and created and maintains the UCSC Genome Browser, a web-based tool that is used extensively in biomedical research and serves as the platform for several large-scale genomics projects. His group's contributions to cancer genomics research include creation of a Cancer Genomics Browser for analyzing data from large-scale cancer studies.

Haussler's group built CGHub to support all three major NCI cancer genome sequencing programs: The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and the Cancer Genome Characterization Initiative (CGCI). TCGA is a collaborative effort led by NCI and the National Human Genome Research Institute to map the genomic changes that occur in at least 20 major types and subtypes of adult cancer. The TARGET program is a related effort focusing on the five most common childhood cancers, and the CGCI makes available genomic data from HIV-

associated cancers and certain lymphoid and childhood cancers.

These programs are laying the foundation for personalized cancer care by creating a database that scientists around the world can use to connect specific genomic changes with clinical outcomes. Haussler's group has been closely involved in data analysis for TCGA.

"TCGA is allowing us for the first time to look at cancer in full molecular detail," Haussler said. "Cancer is a disease caused by disruption of DNA molecules within the cell. When life starts, every cell in the body has the same DNA. In the course of a person's lifetime, however, some cells may accumulate changes in their DNA that cause them to go rogue and multiply without control, creating the disease we call cancer. For the first time now, we are able to look into an individual patient's cancer cells and see all the genetic disruptions, among which are the molecular drivers of that person's cancer."

There are currently only a few situations in which doctors can prescribe a treatment plan based on the specific genetic mutations in a patient's cancer cells. That is expected to change as projects like TCGA, TARGET, and CGCI yield a comprehensive catalog that researchers can use to find new targets for medicines and discover clues to improve patient outcomes. But there is an urgent need for an efficient and user-friendly portal to give researchers access to the data. The NCI genome projects are producing staggering amounts of data.

"The scale of this is far beyond anything faced in medical research before," Haussler said. "Each genome file, the DNA record from a tumor or normal tissue, is 300 billion bytes. And for every case there are two of these files, the cancer genome and the normal genome. Add to this RNA sequence data, and the prospect of deeper sequencing in the future, and we must plan for up to a terabyte (1,000 billion bytes) for each case."

TCGA currently generates about 10 terabytes of data each month. For comparison, the Hubble Space Telescope amassed about 45 terabytes of data in its first 20 years of operation. TCGA's output will increase tenfold or more over the next two years. Over the next four years, if the project produces a terabyte of DNA and RNA data from each of more than 10,000 patients, it will have produced 10 petabytes of data (a petabyte is 1,000 terabytes). And TCGA is just the beginning of the data deluge, Haussler said, noting that 10,000 cases is a small fraction of the 1.5 million new cancer cases diagnosed every year in the United States alone.

New data compression schemes are expected to reduce the total storage space needed, so the CGHub repository is designed initially to hold 5 petabytes and to allow further growth as needed. That is still a massive amount of data, and CGHub will need to accommodate transfers of extremely large data files.

Managed by the UCSC team, the CGHub computer system is located at the San Diego Supercomputer Center. It is connected by high-performance national research networks to major centers nationwide that are participating in these projects, including UCSC. Haussler's team designed and oversees the storage and computing infrastructure for the repository, which has an automated query and download interface for large-scale, high-speed use. It will eventually also include an interactive web-based interface to allow researchers to browse and query the system and download custom datasets.

It may take years for cancer genomics research to bring about major changes in cancer care. The first step, and the focus of the NCI cancer genomics programs, is to determine which genomic changes are involved in each type of cancer and to understand the molecular and clinical effects of those changes. Then biomedical researchers must identify or develop treatments to block those effects.

"Right now, cancer research needs something on a very large scale, like the Large Hadron Collider in physics," Haussler said. "Instead of bringing subatomic particles together in high-energy collisions and computing their behavior, we're bringing cancer genomes together in a common database and computing the disease drivers."

CGHub program director is Robert Zimmerman, and project team members include technical director Mark Diekhans; operations manager Linda Rosewood; hardware systems lead Erich Weiler; engineering lead Chris Wilks; engineering consultant Brian Craft; and networking consultants Brad Smith and Jim Warner. The core code, including GT software for downloading data, was licensed from Annai Systems. The cancer genomics group at UCSC also includes co-principal investigator Joshua Stuart, an associate professor of biomolecular engineering at UCSC; assistant research scientist Jing Zhu; engineers Kyle Ellrott, Teresa Swatloski, and Singer Ma; user testing engineer Mary Goldman; postdoctoral scholars Adam Ewing, Benedict Paten, and Daniel Zerbino; research associate Charlie Vaske; and graduate students Tracy Ballinger, Steve Benz, Daniel Carlin, James Durbin, Ted Goldstein, Mia Grifford, Sam Ng, Amie Radenbaugh, Zack Sanborn, and Chris Szeto.

Provided by University of California - Santa Cruz