PUTTING GENES TOGETHER

REALLY FAST

PSC's newest supercomputer, Blacklight, is helping to break open a potential bottleneck in processing and analysis of DNA sequence data

It didn't take long for Blacklight to show its mettle as a tool for genome sequencing. PSC's newest supercomputer, a resource of XSEDE (see p. 5), came online as a production system in October 2010, and due to its availability, two projects involving genomics — a science that has in the last few years shifted into data-intensive overdrive — made remarkable progress.

New sequencing instruments, hardware technologies that "read" sequences of DNA and decipher the order of nucleotide bases — A, G, C and T (adenine, guanine, cytosine and thymine) — have begun to produce data at unprecedented speed. "Within the last three to five years," says Cecilia Lo, chair of the University of Pittsburgh School of Medicine's Department of Developmental Biology, "new sequencers have come on line, carrying out sequencing that is referred to as 'next-generation sequencing.' What used to take years with capillary sequencing can now be accomplished in a matter of one or two weeks."

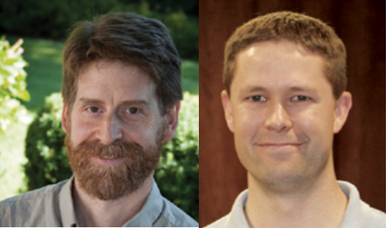
The essential difference is long versus short reads. Previous sequencers did reads of about 300 to 500 and sometimes up to 1000 bases. The new technologies do reads of 50 to 100 bases. "The result," says Lo, who has used Blacklight for her work on the genetic causes of congenital heart defects, "is that the cost of sequencing per base has gone down dramatically and the sequencing runs can be done much more quickly."

"To put it in perspective," says James Vincent of the University of Vermont, who directs the Bioinformatics Core of the Vermont Genetics Network, "it took about 13 years to complete the sequencing of the first human genome. The new instruments can sequence two human genomes in a single run." As part of a team of bioinformatics scientists collaborating through the Northeast Cyberinfrastructure Consortium (NECC), Vincent is using Blacklight to assemble the genome of the little skate (*Leucoraja erinacea*), a fish species of the northwestern Atlantic. "The amount of sequencing data that can be generated from a single instrument," he adds, "has for several years been doubling every four or five months."

While these skyrocketing quantities of sequence data are a blessing for biological science, they pose the problem and challenge of a potentially stifling analytical bottleneck. Once a sequencing instrument has produced millions or, as the case may be, billions of reads from an organism's DNA, researchers face the task of assembling them into a complete genome. Blacklight is helping to break this bottleneck.

For Lo's work, involving mouse genome data, her collaborators used Blacklight to process over 700 million reads and assemble them into a whole genome in eight hours. This compares to about two weeks on a laboratory-based cluster system she and her collaborators had been using.

For Vincent, the Blacklight advantage is perhaps even greater. With NECC's little skate project, he worked for several months on other computing systems before coming to PSC. "Within a week," says Vincent, "90-percent of my problems were solved." With billions of 100-base reads, he was able to complete a *de novo* assembly of the little skate genome in weeks, a large step toward a complete analyzed genome, progress that had eluded him on other systems for nearly a year.



THE RESEARCHERS

James Vincent (left)
UNIVERSITY OF VERMONT

Phil Blood
XSEDE ADVANCED USER SUPPORT CONSULTANT, PSC

THE LITTLE SKATE

As a sequencing project, the little skate isn't just any fish that hadn't yet had its genome sequenced. It's one of only 11 non-mammals selected by the NIH as a "model organism," organisms that have "the greatest potential to fill crucial gaps in human biomedical knowledge." Model organisms are often used as a reference for better understanding human disease conditions. The skate, for instance, shares characteristics with the human immune, circulatory and nervous systems.

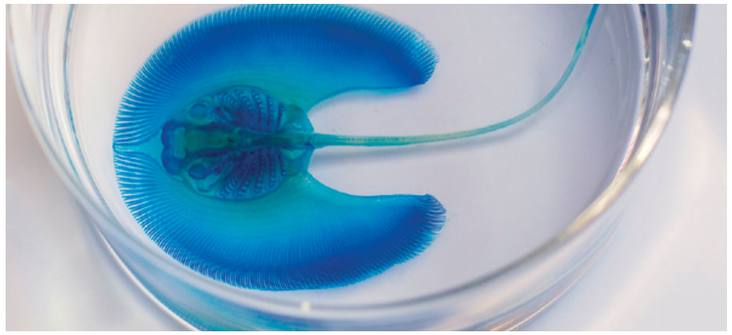
With an American Recovery and Reinvestment Act grant in 2009, NECC was formed to create a high-speed fiber-optic network in five northeast states. It also links five institutions with bioinformatics research programs — Mount Desert Island Biological Laboratory (MDIBL) in Maine, the University of Delaware, Dartmouth College in New Hampshire, the University of Rhode Island and the University of Vermont. When NECC was established, says Vincent, the plan to sequence the entire genome of the little skate, originally a project at MDIBL, gained momentum: "This is the kind of project we anticipated when building our network. It's both an excellent demonstration project for the use of the infrastructure, and at the same time, it's a superb scientific project."

A large part of the challenge is the lack of a reference genome. "It's called *de novo* assembly," says Vincent. "We have to create the little skate genome from scratch. This particular branch down the tree of life fits a niche that doesn't yet have sequenced genomes."

The skate genome is 3.4 billion bases, a little larger than the human genome, and the task was to take billions of 100-base reads — in which many of the bases, sometimes as many as 99, overlap with the next read — and match them with each other in the right order. "The sequencing instrument gives you billions of tiny pieces of a long continuous DNA string," says Vincent, "and to create a draft genome *de novo*, you have to put those little pieces back together."

Using software called ABySS, Vincent brought the project to PSC for NECC in 2011. ABySS's algorithm for genome assembly, he explains, builds a graph of the relationships from all the reads in memory. Some parts of the job require only one or a small number of processors, while others exploit massive parallelism — many processors at once. Because of this, says Vincent, it's often necessary to move back and forth between a massively parallel cluster and a single-processor machine with very large memory. "Blacklight's shared memory makes all this go away."

"Memory is shared across all the nodes, so you can treat it like a traditional cluster, or you can access all the memory you have allocated from a single node, which acts like a single, large-memory machine. You need both of these to complete the ABySS job, and you can do that all at once on Blacklight. This machine made running ABySS easy."



Little skate in a lab dish
Credit: Mount Desert Island Biological Laboratory

As a result, Vincent was able to complete a draft genome of the little skate, which he and his collaborators are now analyzing in comparison with another little skate draft genome — done by Ben King at MDIBL with different software.

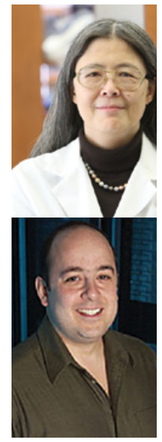
As much as Blacklight's shared-memory architecture, says Vincent, PSC staff made the project go.

As much as Blacklight's shared-memory architecture, Vincent credits PSC's consulting staff, in particular XSEDE advanced user support consultant Phil Blood. "I wouldn't have been able to do anything on Blacklight without PSC staff and Phil Blood in particular. They made the project go. Phil took a real interest and solved a lot of things that were hard for me. He found bugs in the software and got them resolved with the software authors. I'd worked for months and not made that progress. Without Phil's expertise, I might have given up and gone a different route."

TRACKING THE GENES FOR DEFECTIVE HEARTS

Cecilia Lo and her colleagues would like to identify the genes involved in human congenital heart disease. Using genetically modified mice, her research group aims to find gene mutations that cause problems in cardiac development that can lead to structural heart defects — such as holes in the walls of the heart or abnormal connection of the aorta or pulmonary artery — defects that affect almost one-percent of live births and can cause newborn infant death.

Ultimately, the goal is a "diagnostic chip" for human congenital heart disease. "This is the age of personalized medicine," says Lo, founding chair of her department, one of a handful of developmental biology departments nationwide, "and that's where medicine is headed. Such a chip would provide the possibility to retrieve sequence information on many if not all of the genes involved in structural heart disease."



THE RESEARCHERS

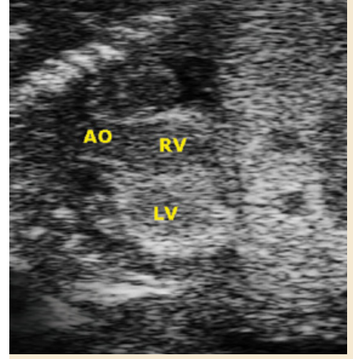
Cecilia Lo
UNIVERSITY OF PITTSBURGH SCHOOL OF MEDICINE (top)

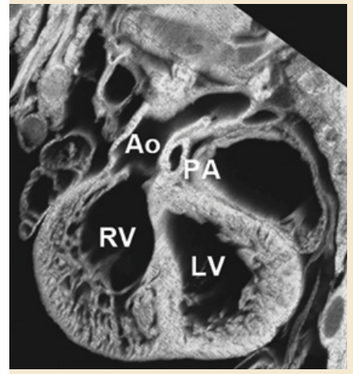
Michael Barmada
UNIVERSITY OF PITTSBURGH GRADUATE SCHOOL OF PUBLIC HEALTH

In the future, each person with congenital heart disease coming to the hospital for treatment, she explains, would have their blood drawn to obtain DNA. Sequencing analysis with this chip could determine if the patient has cardiac-related gene mutations, and a treatment plan could be customized to the patients genetic make up. "We want to understand how these genes contribute to structural heart disease, influence disease progression, and affect long-term outcome. Whether medication or a surgical approach, you would be able to optimize and personalize the medical care of each patient based on the individual's specific genotype."

To find the core set of heart-defect related genes, Lo and colleagues are screening over 100,000 mutant mouse fetuses over five years. When they see a heart defect, using noninvasive *in utero* ultrasound imaging, they follow up by sequencing the genome of that mouse — for comparison to the reference genome of a normal, healthy mouse. "In this way," says Lo, "we should be able to identify the mutations involved in the heart disease."

A next-generation sequencer can rapidly provide sequence data for the entire genome of each mutant mouse. To assemble the reads into a complete genome, one approach is to break the data into chunks, says Michael Barmada, one of Lo's collaborators, from the Department of Human Genetics at the University of Pittsburgh's Graduate School of Public Health. "We worked closely with PSC staff," he says, "who have been really helpful, in working this out on Blacklight."





CONGENITAL HEART DEFECTS
Ultrasound imaging (top) of a genetically-modified mouse *in utero* showed the aorta (AO) emerging from the right ventricle (RV). Susequent microscopic histopathology (bottom) showed a double-outlet right ventricle with pulmonary atresia — the aorta and a very small pulmonary artery (PA) emerge from the right ventricle.

After testing and benchmarking several approaches, Barmada split the sequence data into several independent chunks that ran concurrently on 1000 Blacklight cores, with the result that assembly of the genome for one mutant mouse took only eight hours. "This was taking at least a week-and-a-half," says Barmada, "on a 24-core machine in our lab."

"Blacklight is allowing us to do things that would be very difficult to do otherwise."

"Given the many mouse mutants awaiting analysis, we have a huge amount of sequencing data," says Lo, "that will need to be mapped back to the mouse reference genome. Blacklight is allowing us to do things that would be very difficult to do otherwise."

MORE INFO:
www.psc.edu/science/2011/sequencing