REPRINTED FROM

genomeweb | BioInform
The Integrated Informatics News Source

Feb. 13, 2012

# A Blacklight for Bioinformatics

*by Matthew Dublin*

**WHEN IT COMES** to high-performance computing systems suited to genomics, researchers might be hard pressed to do better than Pittsburgh Supercomputing Center's Blacklight supercomputer system. Under its hood are 256 SGI blades that comprise a total of 4,096 processing cores at 2.27 gigahertz. This means a total peak performance of 37 teraflops — powerful, yes, but nothing to write home about considering the number one-ranked system in the world currently operates at 10.5 petaflops. What does make Blacklight stand out, however, is that the 16 cores on each blade share 128 gigabytes of local memory, which means it has a total memory capacity of 32 terabytes. That memory is then divided into two partitions, with each partition allotted 16 terabytes of shared coherent memory, making it the largest coherent shared memory system in the world.

## Enter Blacklight

Blacklight came online in October 2010 and has already shown that large shared-memory architectures might be the way forward for -genomic data analysis on high-performance computing systems.

Blacklight is currently being used to analyze genomic data sets with the same bioinformatics software researchers already use in their labs. One example is the Little Skate Genome Sequencing and Annotation Project — a multi-institutional effort supported by the North East Cyber-infrastructure Consortium, consisting of Mount Desert Island Biological Laboratory in Maine, the University of Delaware, Dartmouth College in New Hampshire, the University of Rhode Island, and the University of Vermont. The aim of the project is to sequence and analyze the genome of the little skate, a chondrichthyan fish that evolved about 450 million years ago, and is one of the most primitive vertebrates to have both a jaw and paired appendages. The little skate is also one of 11 non-mammal model organisms thought to have the potential to provide data on fundamental vertebrate characteristics such as jaws, teeth, an adaptive immune system, and a pressurized circulatory system. Investigators participating in the project hope that the little skate's genome can help them better understand these aspects of human physiology.

To assemble the little skate genome, bioinformaticians used ABySS, a *de novo,* parallel, paired-end sequence assembler, on standard commodity computer clusters at institutions within the NECC. According to the University of Vermont's James Vincent, the lead bioinformaticist on the project, certain aspects of the ABySS assembly process require lots of parallel computations — which can be completed on a traditional Beowulf cluster — while others require a single processing core, but lots of memory.

"You do this first part on a traditional cluster, then you stop and find another machine that has a large amount of memory and move all of your stuff there. And typically single machines with a huge amount of memory only have a handful of cores. So this is where Blacklight comes in," Vincent says. "It's both kinds of machines — you can run on many thousands of cores, like a traditional cluster, and then within that same job switch over to running as if you are running on a single machine with a single memory space and access all of the memory space within that

*continued on next page*

REPRINTED FROM
genomeweb BioInform

machine. So it does both of those things together which makes life easy for someone like me."

Before reaching out to the Pittsburgh Supercomputing Center, Vincent says that he toiled for months -running ABySS on billions of 100-base reads from little skate sequence data sets, using the available commodity clusters in the NECC network. But by taking advantage of Blacklight's unique shared memory design, he and his team were able to complete a *de novo* assembly of the little skate genome in just weeks.

## Divide and conquer

University of Pittsburgh researchers are also using Blacklight to study congenital heart disease using transgenic mice with the goal of developing a diagnostic chip that could one day be used to identify heart disease in humans.

Michael Barmada, director of the Center for Computational Genetics at Pittsburgh, together with Cecilia Lo, a professor at the medical school there, are screening more than 100,000 mutant mouse fetuses to look for heart defects. Once a defect is identified, the genome of that mouse is sequenced and compared to a reference genome from a healthy mouse.

With a seemingly never ending conveyer belt of mouse genomes requiring assembly, Barmada and Lo needed a quick and easy way to break up their sequence data and process it concurrently. "We had tried to do the alignments of sequence data on a multi-processor workstation and it was taking days and days, that's what pushed us to try out Blacklight," Barmada says. "It really helped us optimize a lot of the parameters for the alignment steps and all of the next-generation sequence analysis. We could set up our pipeline and not have to worry about it being optimized, run it on Blacklight, see how much memory it took, and how to break it up into smaller pieces that would then run efficiently on other computing systems, not just on Blacklight."

Before utilizing Blacklight, assembling the genome of one mutant mouse could take almost two weeks on a 24-core machine. Barmada and Lo are now able to assemble the genome of one mutant mouse in less than eight hours.

More genomics projects are on the way for the new system, says Phil Blood, a senior scientific specialist at PSC and Blacklight user-support consultant. Blood is working with bio-informatics software developers from the Broad Institute's Computational Research and Development group on some of their software solutions, including the short-read genome assembler ALLPATHS-LG.

"We're starting to get involved in some joint efforts. In particular, we've been having discussions with the Broad Institute about two of their recent software development projects that are now going through their paces on Blacklight," Blood says. "We have people running genome assemblies on ALLPATHS-LG, and we've also had some discussion with their developers on the ALLPATHS-LG team and we're interested in getting it to work optimally on Blacklight. Right now, we are helping to work out some of those wrinkles, not only from the code on the machine for the first run, but we have also tested the code with full research-level assemblies that are more complex than the test case we used initially to get it running."

Blood and his team also made the Trinity software package — a collection of tools for efficient and robust *de novo* reconstruction of transcriptomes from RNA-seq data — available on the supercomputer in November. Although Blood says transferring software tools originally designed to run on a workstation or small cluster can be challenging, that most bioinformatics tools are open source simplifies the process.

"There are other packages that people want to use that are not open source … which means you have to go work with developers in the commercial setting to sort that out," he says. "Those are the main challenges with that project."

Blacklight is accessible via an application process through the National Science Foundation's Extreme Science and Engineering Discovery Environment, a collection of integrated digital resources and services from around the world that can be accessed as single virtual system. Any researcher affiliated with an academic institution in the US, or who has a collaborator there, can access Blacklight for free by submitting a CV and an abstract.

While the usual barrier to entry into the world of supercomputing for most bench biologists or bioinformaticians is the difficulty of porting commonly used software analysis tools to a large high-performance computing, Blacklight has a great track record when it comes to making this process not only possible, but relatively easy.

"People like Phil Blood are very responsive and helpful for working out the code, bugs, getting things fixed, and these are all extremely time consuming things for someone like me to do," Vincent says. "So the unique architecture of Blacklight combine with the very professional staff of the PSC is a hard combination to beat."