**Shared Memory Gene Assembly**

Because of its shared-memory design, PSC’s Blacklight has become an invaluable tool in assembly and analysis of next-generation sequencing data, including projects in behavioral genomics and metagenomics.

In genomics, the next generation is now. This relatively new branch of the life sciences has in the last few years, due to new technologies, exploded with possibility and data. “Next generation” sequencing tools have taken genomics well beyond the Human Genome Project to studies of nearly every kind of organism, from ants and bumblebees to Patagonian tucutucos (more on that below) among many others, by deciphering the order of nucleotide bases — A, G, C and T (adenine, guanine, cytosine and thymine) — at unprecedented speed.

The essential difference is long versus short reads. Previous sequencers did reads of about 300 to 500 and sometimes up to 1000 bases. The technologies gain their advantage by doing much shorter reads, 50 to 150 bases — at greatly reduced cost per base — and can generate in a week as much sequence data as would require a year for the traditional sequencers. Consequently, genomics has shifted into data-intensive overdrive, with many opportunities to do important research. While it’s an unprecedented blessing for the life sciences, it’s also an unprecedented challenge for data processing and analysis.

Once a sequencing instrument has produced millions or, as the case may be, billions of reads from an organism’s DNA (or RNA), researchers face the task of assembling them. To add to the degree of difficulty, short reads amplify the computational challenge — many more pieces of data must be fit together based on shorter overlaps. How do you assemble all those sequence fragments into complete and accurate genomic strands? Imagine a jigsaw puzzle piece with 100 big pieces versus the same picture with 2,000 little pieces. It’s a potentially mind-boggling problem handled by very sophisticated computational algorithms, requiring many runs, careful checking and, some say, as much art as science.

Since Blacklight came online in October 2010, with two partitions of 16 terabytes of shared memory, the largest shared-memory system in the world, it’s become a powerful tool in meeting this challenge. An article in GenomeWeb (February 1, 2012) highlighted Blacklight’s advantages — very large shared memory making it possible to contain entire base-pair datasets in random-access memory (RAM) — dramatically improving workflow and throughput time, as compared to non-shared memory clusters, for genomics assembly and analysis.

Beyond that, observes PSC scientist and XSEDE consultant Phil Blood, many large genomics assemblies simply couldn’t be done without large shared-memory, such as Blacklight provides. To enhance Blacklight’s genomics capabilities for XSEDE researchers, Blood has made nearly all genomics software tools available for easy use. “Blacklight has an extensive collection of pre-compiled modules for the analyses of next-generation sequence data,” says Matthew MacManes, of the University of California, Berkeley. MacManes attempted genomics analysis with a number of high-end computing systems before coming to Blacklight. He used nearly 20 different programs in his assembly and analysis at PSC. “Having these programs installed and maintained by PSC staff is extremely helpful.”

MacManes’ project with Blacklight involved assembly and analysis of RNA from tucutucos, a burrowing rodent from Patagonia — of particular interest to those tucos live in social groups while others of the same species are anti-social and live alone. MacManes identified a number of genes that are expressed or not depending on whether the tucos live alone or in a colony. “For my research,” says MacManes, “the use of Blacklight has been absolutely revolutionary, allowing me to complete analyses that link specific patterns of gene expression with mammalian social behavior.”

In science, new tools often make it possible to look at new questions, and the availability of next-generation sequencing has led to studies in “metagenomics” — analysis of genes from many organisms that coexist in the same place — unimaginable a few years ago. In metagenomics, unlike the traditional approach of analyzing the genome of one organism, you can pick an environment and take a sample,” says Blood. “It could be Old Faithful, or what’s inside human intestines, wherever you might find interesting microbial communities.”

In a metagenomics study with Blacklight, Blood helped researchers from Oklahoma State assemble sequencing data from soil that came from a sugarcane plantation in Brazil. The goal is to find enzymes that can efficiently break down non-feed plants, such as sugarcane.
“Animals to be found, and this variation makes tucos compelling subjects for study,” says Matthew MacManes, a sponsored post-doctoral fellow. The colonial tuco-tuco (Ctenomys sociabilis) — so-called for a clicking sound it makes — is the subject of MacManes’ interest in behavior. MacManes, along with Prade and Brian Couger, has taken an innovative metagenomics approach. To find enzymes that can do the heavy-lifting of biodigesting, they started with a soil sample, gathered from a Brazilian sugarcane field.

“Currently there is simply no better resource for this type of work.”

To grapple with assembly and analysis of this huge amount of sequence data, beginning with the task of building complete “transcriptomes” — full strands of RNA — MacManes first turned to large distributed-memory machines at several sites, but eventually came to PSC’s Blacklight. Using 80 cores of Blacklight, 640 gigabytes of RAM, he completed the assembly in 14 days of computing, with subsequent analysis extending for months.

The work identified a number of genes that are differentially expressed depending on tuco social behavior, and MacManes and Lacey have a manuscript in preparation reporting their findings. “Blacklight is a key resource for my analyses of next-generation sequence data,” says MacManes. “Without it, I would simply have been unable to complete the requisite analyses. I feel so strongly about Blacklight that I have referred colleagues and collaborators. Currently there is simply no better resource for this type of work.”

In Brazil,” says Rolf Prade, a professor of microbiology at Oklahoma State University and a Brazilian, “biofuel is standard gasoline for cars.” The world’s second largest producer of biofuel, Brazil gets its ethanol from sugarcane, uniquely available there due to enormous amounts of arable land and suitable climate. Like corn in the United States, however, sugarcane has major disadvantages as a biofuel source. The amount of energy available per amount of input crop is much greater in non-feed stock plants, with denser fibrous structure (lignocellulosic plants), such as switchgrass and wheat straw, but these non-food plants are expensive and difficult to degrade into biofuel.

Considerable research worldwide is focused on finding inexpensive means to overcome this obstacle. A team at Oklahoma State led by Mostafa Elshahed, along with Prade and Brian Couger, has taken an innovative metagenomics approach. To find enzymes that can do the heavy-lifting of biodigestion, they started with a soil sample, gathered from a Brazilian sugarcane field.

“Making biofuels from lignocellulosics doesn’t work because we don’t know how to decompose the biomass. It’s because we don’t have all the proteins, work because we don’t have all the enzymes,” says Prade. “This is like a protein discovery platform,” says Couger. “It wouldn’t have been possible for us to do this on any computational cluster other than Blacklight.”

The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. “This is the largest metagenomics assembly ever done,” says Couger, “and it would have been intractable on any computational cluster other than Blacklight.” To make use of this assembled data as a means to identify enzymes, the researchers also did a series of protein separations from the soil samples and applied mass spectrometry. From precise molecular weights, they inferred amino-acid sequences. Matching these sequences with the assembly data, the researchers identified more than 8,000 gene candidates related to glycoside hydrolase, a category of enzymes that can degrade plant cell walls.

“The colonial tuco-tuco (Ctenomys sociabilis) — so-called for a clicking sound it makes — is the subject of MacManes’ interest in behavior. MacManes, along with Prade and Brian Couger, has taken an innovative metagenomics approach. To find enzymes that can do the heavy-lifting of biodigesting, they started with a soil sample, gathered from a Brazilian sugarcane field.

“Making biofuels from lignocellulosics doesn’t work because we don’t know how to decompose the biomass. It’s because we don’t have all the enzymes,” says Prade. “This is like a protein discovery platform,” says Couger. “It wouldn’t have been possible for us to do this on any computational cluster other than Blacklight.”

The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. “This is the largest metagenomics assembly ever done,” says Couger, “and it would have been intractable on any computational cluster other than Blacklight.”

The 1979 Brazilian Fiat 147 was the first modern automobile capable of running only on ethanol, which Brazil produces from sugarcane, shown ready for harvest at a plantation in São Paulo State.

“The colonial tuco-tuco (Ctenomys sociabilis) — so-called for a clicking sound it makes — is the subject of MacManes’ interest in behavior. MacManes, along with Prade and Brian Couger, has taken an innovative metagenomics approach. To find enzymes that can do the heavy-lifting of biodigesting, they started with a soil sample, gathered from a Brazilian sugarcane field.

“Making biofuels from lignocellulosics doesn’t work because we don’t know how to decompose the biomass. It’s because we don’t have all the enzymes,” says Prade. “This is like a protein discovery platform,” says Couger. “It wouldn’t have been possible for us to do this on any computational cluster other than Blacklight.”

The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. “This is the largest metagenomics assembly ever done,” says Couger, “and it would have been intractable on any computational cluster other than Blacklight.”

The 1979 Brazilian Fiat 147 was the first modern automobile capable of running only on ethanol, which Brazil produces from sugarcane, shown ready for harvest at a plantation in São Paulo State.

“Making biofuels from lignocellulosics doesn’t work because we don’t know how to decompose the biomass. It’s because we don’t have all the enzymes,” says Prade. “This is like a protein discovery platform,” says Couger. “It wouldn’t have been possible for us to do this on any computational cluster other than Blacklight.”

The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. “This is the largest metagenomics assembly ever done,” says Couger, “and it would have been intractable on any computational cluster other than Blacklight.”

The 1979 Brazilian Fiat 147 was the first modern automobile capable of running only on ethanol, which Brazil produces from sugarcane, shown ready for harvest at a plantation in São Paulo State.

“Making biofuels from lignocellulosics doesn’t work because we don’t know how to decompose the biomass. It’s because we don’t have all the enzymes,” says Prade. “This is like a protein discovery platform,” says Couger. “It wouldn’t have been possible for us to do this on any computational cluster other than Blacklight.”

The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. “This is the largest metagenomics assembly ever done,” says Couger, “and it would have been intractable on any computational cluster other than Blacklight.”

The 1979 Brazilian Fiat 147 was the first modern automobile capable of running only on ethanol, which Brazil produces from sugarcane, shown ready for harvest at a plantation in São Paulo State.

“Making biofuels from lignocellulosics doesn’t work because we don’t know how to decompose the biomass. It’s because we don’t have all the enzymes,” says Prade. “This is like a protein discovery platform,” says Couger. “It wouldn’t have been possible for us to do this on any computational cluster other than Blacklight.”

The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. “This is the largest metagenomics assembly ever done,” says Couger, “and it would have been intractable on any computational cluster other than Blacklight.”

The 1979 Brazilian Fiat 147 was the first modern automobile capable of running only on ethanol, which Brazil produces from sugarcane, shown ready for harvest at a plantation in São Paulo State.

“Making biofuels from lignocellulosics doesn’t work because we don’t know how to decompose the biomass. It’s because we don’t have all the enzymes,” says Prade. “This is like a protein discovery platform,” says Couger. “It wouldn’t have been possible for us to do this on any computational cluster other than Blacklight.”

The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. “This is the largest metagenomics assembly ever done,” says Couger, “and it would have been intractable on any computational cluster other than Blacklight.”

The 1979 Brazilian Fiat 147 was the first modern automobile capable of running only on ethanol, which Brazil produces from sugarcane, shown ready for harvest at a plantation in São Paulo State.

“Making biofuels from lignocellulosics doesn’t work because we don’t know how to decompose the biomass. It’s because we don’t have all the enzymes,” says Prade. “This is like a protein discovery platform,” says Couger. “It wouldn’t have been possible for us to do this on any computational cluster other than Blacklight.”

The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. “This is the largest metagenomics assembly ever done,” says Couger, “and it would have been intractable on any computational cluster other than Blacklight.”

The 1979 Brazilian Fiat 147 was the first modern automobile capable of running only on ethanol, which Brazil produces from sugarcane, shown ready for harvest at a plantation in São Paulo State.

“Making biofuels from lignocellulosics doesn’t work because we don’t know how to decompose the biomass. It’s because we don’t have all the enzymes,” says Prade. “This is like a protein discovery platform,” says Couger. “It wouldn’t have been possible for us to do this on any computational cluster other than Blacklight.”

The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. “This is the largest metagenomics assembly ever done,” says Couger, “and it would have been intractable on any computational cluster other than Blacklight.”

The 1979 Brazilian Fiat 147 was the first modern automobile capable of running only on ethanol, which Brazil produces from sugarcane, shown ready for harvest at a plantation in São Paulo State.