Professional Development Grant Report

Striped scorpion genome sequencing

Tsunemi Yamashita

Department of Biological Sciences

Grant award

Spring 2016

B. Problem researched

The proposed project will assist in a genomic analysis of the striped scorpion, *Centruroides vittatus*. We plan to collect DNA from a sample scorpion, verify its quality, and send the DNA to a DNA core facility for genomic sequencing. The data from the genomic sequence will be analyzed to identify, catalog, and characterize the toxin genes associated with venom toxicity in this scorpion. This research is significant as scorpion genomics, especially focused on venom toxin genes, is poorly understood.

C. Research Activity Review

Total genomic DNA was extracted from two female scorpions collected in Pope County, AR with the Qiagen genomic-tip and genomic DNA buffer set (Qiagen, Inc.). The genomic DNA quality was analyzed through 0.9% agarose gel electrophoresis and UV spectroscopy. The genomic DNA was sent to the National Center for Genome Resources (NCGR, NM) for PacBio 20K library generation and 10 SMRTcell sequencing for each individual genome. The de novo assembly was conducted at the High Performance Computing Center at the University of AR-Fayetteville. We ran the raw sequencher data through the Canu Pipeline Assembler (v1.3) with quality assessment through Quast (Quality Assessment Tool, ver. 4.0) and quality control check through FastQC (0.11.5). The mtDNA and a *Mycoplasma* genome was removed from the assemblies. One assembly (Q1133) was BLASTed against several databases (UniProt and NCBI Protein & nucleotide databases-RefSeq) for an initial annotation and identification of toxin genes and proteins. Transcript and protein predictions were completed with SNAP and BLASTp analysis to annotate.

D: Summary of findings

Two scorpion genomic libraries were constructed followed by two 10 SMRT cell sequencing runs for genome assembly. The Quast program produced the following Table for the two PacBio genomic sequences (Table 1) after assemblies with genome size estimates of 170 and 350 Mb. The assemblies required 80hrs on 2.6 GHz Trestle Core with 64GB RAM with sequence coverage estimated as 40X. An undescribed *Mycoplasma* genome was identified in the genome.

Gene Annotation

The two assemblies produced different estimates of unique genes and both estimates were distinct from an earlier MiSeq assembly.

QUAST predicted unique genes: 18,250 (Q1133) & 5,655 (Q1171)
SNAP predicted genes: 23,361

BLASTp annotation:

Metasellus occidentalis RefSeq	2.4%
Insect UniProt	20.0%
NCBI NonRedundant	21.1%
None	56.6%

Toxin gene and protein identification:
A Protein Blast identified 28 scorpion toxin genes; four with significant identities to *Centruroides* toxins. The Nucleotide Blast with >80 toxin genes identified two contigs with significant matches to scorpion toxin genes.

E. Conclusions

- The polished Q1133 assembly appears to show a more robust assembly than a second PacBio and MiSeq assemblies.

- The nucleotide and protein BLAST suggest up to five sodium toxin genes for these individuals.

These results were presented in a poster at the 2017 Plant and Animal genomics meeting in San Diego, CA.

The PDG funds were supplemented through a NIH-INBRE DNA Sequencing Core Voucher of $5,000.00.