Gene prediction in the fathead minnow [*Pimephales promelas*] genome

Saari, TW//Villeneuve, Daniel L//Ankley, Gerald T//Burns, F//Cogburn, A//Deschamps, S//Jackson, R//Hoke, RA//Schroeder, Anthony L

The fathead minnow is a well-established model organism which has been widely used for regulatory ecotoxicity testing and research for over half century. While much information has been gathered on the organism over the years, the fathead minnow genome, a critical source of information for modern research, has been largely uncharacterized, to date. To address this limitation, the fathead minnow genome was recently sequenced using high-throughput Illumina sequencing with paired end reads of various insert sizes. Two draft assemblies were generated using SOAPdenovo and String Graph Assembler (SGA), respectively. Both methods yielded contigs totaling ~810MB, with a GC content of ~38%. When the assemblies were tested against a set of highly conserved eukaryotic genes, the SOAPdenovo assembly yielded the best results, with 91% of genes mapping to the SOAPdenovo assembly, 74% of which were mapped completely on one scaffold. The present study is focused on further annotation of the draft assemblies utilizing the AUGUSTUS gene finding program in a well-established eukaryotic genome annotation pipeline. RepeatMasker was used to filter out tandem repeat satellite sequences and retrotransposable element sequences using the newly-available zebrafish repeat library from RepBase. The AUGUSTUS program was then trained for *ab initio* gene finding with spliced alignments of randomly selected zebrafish protein sequences. *Ab initio* gene finding will provide the first iteration of gene prediction, followed by additional iterations which will incorporate fathead minnow expressed sequence tag data to improve accuracy. The completion of this genome annotation will provide a much-needed resource for future design of ecotoxicology studies using the fathead minnow.