Abstract

The family Percidae contains over 200 species, most of which are within the subfamily Etheostomatinae. This subfamily (the darters) represents a species rich radiation of freshwater fishes in North America. Evolutionary relationships between darter species have been derived from morphological, mitochondrial DNA sequence and limited nuclear DNA sequence comparisons. However, a thorough understanding of the evolution of the darter species will require comparisons at the whole genome level. As a first step, the genome of the Tallapoosa darter (Etheostoma tallapoosae) has been sequenced utilizing two Illumina MiSeq 250-PE runs generating 52 million reads. This provided an average 12 fold coverage of the estimated 1 billion nucleotide genome. The sequences were assembled with Minia into contigs and these were assembled into scaffolds with SPSPACE. A BLAST server has been set up to allow for the annotation of Tallapoosa darter scaffolds homologous to sequences of interest. The scaffolds were also imported into an instance of WebApollo along with gene track data generated by fgenesh. A set of scripts were developed to facilitate the formatting and import of these tracks and scaffold sequences into WebApollo that will make it simple for labs to set up WebApollo instances for their own genome data without extensive computer system experience. A web site has been developed that gives access to both the BLAST and WebApollo servers to the public to spur interest in darter genomics and to enable annotation of the Tallapoosa darter genome by a community of darter researchers.

Introduction

So far, the study of darter evolution has utilized morphological, behavioral and limited DNA sequence analysis. Various subfamilies of the darter family have been elucidated from these studies, there are still many unresolved questions. For example, to what extent do related species share alleles and to what extent does lineage sorting or hybridization during evolution. What are the actual adaptive genetic changes that define darter species and to what extent, if any, do allopatrically distributed and genetically differentiated populations of the same species show adaptive genetic differentiation? A complete understanding of darter evolution must utilize the analysis of complete genomes. While this approach was not financially feasible in the past, the cost of genomic analysis is about to cross a threshold where sequencing of darter genomes of individual species and, soon, populations within species will become affordable.

As a starting point, it will be necessary to have a fully annotated reference darter genome sequence to which the genomes of all other darter species can be compared. As a first step in this direction we have recently obtained the genomic sequence of the Tallapoosa darter (E. tallapoosae).

Sequencing

This sequence was obtained as a result of two 250 nucleotide PE runs on an Illumina MiSeq. A total of 13 billion nucleotides of sequence was obtained from 52 million such 250 nucleotide sequence reads. This represents, on average, about a 12 fold coverage of the darter genome.

Alignment of reads to previously cloned genetic fragments shows that coverage ranges from 2 to 3 fold to as high as 28 fold.

Assembly

The 250 PE sequences were assembled into contigs utilizing the Minia assembler. This assembler was chosen because of its low memory requirements. The sequences were assembled with most of the contigs having kmer = 31 to kmer = 80 settings and minimum abundance = 2 or 3 settings. The best assembly in terms of total number of nucleotides assembled and the maximum contig length was achieved with the settings: kmer = 73 and minimum abundance = 2.

Total number of contigs = 539616

Sum (bp) = 66984269

Total number of N's = 0

Sum (bp) no N's = 66984269

Min contig size = 222

Average contig size = 1224

N50 = 2185

Because of the paired end nature of the reads, it was possible to further assemble some of these contigs into scaffolds with SPSPACE. The following results were obtained:

Total number of scaffolds = 470492

Sum (bp) = 66664090

Max scaffold size = 57949

Min scaffold size = 222

Scaffold N50 size = 1404

N50 = 2913

Utility of Assembly for Annotation

In general, the lengths of the scaffolds are relatively short. While a subsequent phase of this genomic sequencing effort will address the issue of scaffold length, can this current version of the Tallapoosa darter genome assembly be utilized to begin an annotation of the genome? Obviously, those scaffolds that are above 5,000 nucleotides in length likely contain a gene or a significant part of a gene.

To check accuracy of assembly, the scaffolds were aligned to adjacent cloned Tallapoosa darter gene models. In all cases the scaffolds aligned precisely to those genomic sequences.

To further check accuracy of assembly and the utility of scaffolds for annotation, the scaffolds were searched by blastn with several full length Perca flavescens mRNA sequences (closest related species to darters).

The examples below show two instances where genes were identified within the scaffolds.

In the first example, the Urate Oxidase gene was found to be contained within one scaffold.

In the second example, the nephrin (NEP1) gene actually spanned many scaffolds. These scaffolds were identified by a high degree of homology to different portions of the NEP1 mRNA sequence. These scaffolds were then concatenated in the order corresponding to NEP1 mRNA homology.

While the current assembly of the Tallapoosa darter genome based on a 12 fold coverage of PE250 reads produced scaffolds that are relatively short, it appears that the assembly is of sufficiently high quality to facilitate the start of darter genome annotation. WebApollo was chosen as the tool to carry out this annotation process of the Tallapoosa darter genome assembly.

Setting up WebApollo

Initial attempts to annotate some of the Tallapoosa darter scaffolds were carried out using the red line workflow on the DNA Subcommittee website (dnasubway.plantscience.org). The annotation workflow that proved highly successful was to begin with fgenesh derived gene models in Apollo, determine with blastp against the GenBank or database if the gene model codes for a known protein and if a homolog exists, use the homologous protein as input for fgenesh+ determination of the exon/intron structure of the gene within the scaffold. The gene model was then adjusted according to the fgenesh derived model. It was decided, therefore, to implement this workflow in WebApollo.

WebApollo is a fairly complex server side application to set up. However, a virtual machine implementation of WebApollo has been made available that is preconfigured and was easily incorporated into a VirtualBox running on a MacMini server. This makes it relatively easy to implement WebApollo by research groups lacking server administration expertise.

Once the WebApollo instance was installed, the fgenesh derived darter scaffolds were imported into WebApollo along with fgenesh derived gene models as evidence tracks. The WebApollo virtual machine includes a script (setup_webapollo.sh) that makes it simple for individuals with little computer system experience to create a database of scaffolds and to then add individual scaffolds with evidence tracks one at a time. A modified version of this script was utilized in setting up the Tallapoosa darter WebApollo instance along with additional scripts that were written to make the necessary file format conversions, and enable an unattended import of all the desired scaffolds and evidence tracks into WebApollo. The following diagram shows the workflow that was implemented.

The program fgenesh (SoftBerry) was used to generate all incisive gene models file from a file containing the Tallapoosa darter scaffolds in fasta format.

The script fgenesh-splitter.sh was used to extract the fgenesh derived gene models for each of the scaffolds into a separate text file. These gene models are in a gff3 file format.

The script fgenesh-converter.sh was used to convert each of the fgenesh formatted gene model text files into a gff3 formatted file and this script also appends the relevant fasta formatted scaffold sequence into each of the gff3 files.

The standard script setup_webapollo.sh was used to repeatedly call a modified version of the setup_webapollo.sh script to sequentially add each of the gff3 files from within a specific directory. For this step the setup_webapollo.sh script was modified so that indexing and server restarts were omitted.

The standard script setup_webapollo.sh as supplied in the WebApollo virtual machine was used to add the last gff3 file to the Tallapoosa darter scaffold database thus ensuring the indexing of the database as well as restarting of the web server.

It is anticipated that as other research groups sequence the genomes of other darter species that these research groups will want to set up their own instances of WebApollo. Since many such research groups likely will not have the necessary server administration and unix expertise, a number of scripts were written that make it possible for individuals with very minimal unix experience to import scaffolds and fgenesh generated evidence tracks into WebApollo. These scripts are:

fgenesh-splitter.sh
fgenesh-converter.sh
add_to_webapollo.sh

The purpose and use of these scripts is summarized in the previous workflow diagram. Of course, the use of these scripts and the associated workflow is not limited to setting up WebApollo instances of just darter genomes. These may also be of utility to other groups setting up WebApollo instances for annotation of genomes of other species.

Tallapoosa Darter Genome Annotation with WebApollo

To begin annotation of a scaffold, a scaffold is selected from a list.

Once the scaffold opens in the viewer, the fgenesh derived gene model(s) is/are displayed in an evidence track.

If a blastp search of GenBank shows a homologous protein, that protein sequence along with the scaffold DNA sequence is subject to fgenesh+ (SoftBerry) gene prediction analysis

and the gene model in WebApollo is adjusted accordingly and with additional manual adjustments as necessary.

Acknowledgements

This work was partially supported by NIFA award 2011-67044-19254.