Final Report

Genomic resources in lake sturgeon (*Acipenser fulvescens*): A seed proposal

1 June 2010 through 31 May 2011

J. Andrew DeWoody

Purdue University

I. Overall project summary

This IISG seed grant provided 4.5 months of graduate stipend (at one-quarter time effort) for Ph.D. candidate Nick Marra to analyze our lake sturgeon DNA sequence data for genetic markers. The original sequence data were generated with support from the Indiana Department of Natural Resources and the Great Lakes Fishery Trust. These raw data, generated using so-called "next-generation" sequencing methods, consist of 125,000,000 nucleotides of DNA spanning ~473,000 sequence reads. From these sequences, we identified two types of genetic markers—single nucleotide polymorphisms and microsatellites.

A. Single nucleotide polymorphisms

Our bioinformatic analyses of lake sturgeon sequences resulted in the identification of 4075 single nucleotide polymorphisms (SNPs). SNPs are simply variable nucleotide sites; one chromosome might have an "A" at a given site in a gene, whereas another chromosome might have a "G" (i.e., there would be an A allele and a G allele). They are usually biallelic markers (only 2 alleles per locus); the power of this marker comes from their sheer numbers. SNPs have the potential to be developed into informative markers that can be used to map the genetic basis of phenotypic traits such as egg production, growth rates, disease susceptibility, etc. We have not mapped such traits, but we have considered the inheritance of these SNPs.

Lake sturgeon are polyploid animals, meaning that each juvenile inherits multiple genome copies from its mother and multiple genome copies from its father. However, large segments of their genome have been lost through a process termed "diploidization", whereby polyploid animals gradually reduce the number of genome copies to 2 (diploidy). We have used the lake sturgeon SNP data to provide estimates of gene copy number in lake sturgeon (Hale et al. attached). This is a complicated endeavor, and we are under no illusion that our estimates will prove definitive. They will be refined as more and more markers are evaluated, as genome sequencing becomes simpler, and as the scientific community better understands copy number variation. However, our efforts represent the first genome-wide assessment of ploidy in lake sturgeon. We estimate that roughly 10% of lake sturgeon SNPs are inherited in a disomic (2n) fashion, 50% tetrasomically (4n), 25-30% octosomically (8n), and ~10% appear to be present in copy numbers >8n (including dodecasomic 12n and perhaps hexadecasomic 16n loci).

B. Microsatellites

On a per marker basis, SNPs contain relatively little information content. This is because a single locus can have only four possible allelic states (G, A, T, or C). This is in sharp contrast to length polymorphisms at microsatellite loci, where alleles can vary from zero to hundreds of repeats. For example, one chromosome might have a 13 copies of an AAT repeat $(AAT)_{13}$ whereas an alternate allele might have 22 copies $(AAT)_{22}$. Thus, microsatellites typically harbor

far more allelic variation than do SNPs, and mean microsatellite heterozygosity is notable higher than mean SNP heterozygosity. This makes microsatellites much more efficient markers on a per locus basis, and also means they can be used for fine-scale discrimination of parentage, gene flow, and individual identification.

With IISG support, we studied lake sturgeon microsatellites in light of those from rodents and salamanders. We uncovered >2000 microsatellites expressed in lake sturgeon genes, mostly trinucleotide repeats (Doyle et al., attached). Furthermore, we found that lake sturgeon microsatellites are generally longer than those in rodents or salamanders and thus more likely to be polymorphic.

The lake sturgeon microsatellite data will be archived at www.datadryad.org.

II. Students supported

We received IISG funds to support a graduate student to conduct the bioinformatic and laboratory assays required to develop sturgeon markers. Nick Marra conducted the assays; his dissertation research is on the use of next-generation sequencing in ecology and evolutionary biology. Marra will begin his fourth year as a graduate (PhD) student in the Fall of 2011. In the spring of 2011, Marra was awarded a $15,000 Doctoral Disseratation Improvement Grant by the National Science Foundation (NSF).

Beyond Marra's NSF grant, this IISG Seed Grant provided us with the impetus to submit a $400,000 NOAA Sea Grant Aquaculture Research Program proposal in 2010 ("***Integrating genomic data and pedigree information to enhance aquaculture production"***). Unfortunately, that application was unsuccessful.

III. List of publications (see attached drafts*)

A. *Microsatellite analyses across transcriptomes of lake sturgeon, tiger salamanders, and kangaroo rats*, by Jacqueline M Doyle, Gregor Siegmund, Joseph D Ruhl, Soo Hyung Eo, Matthew C Hale, Nicholas J Marra, Peter M Waser, and J Andrew DeWoody.

B. *Transcriptome sequences from polyploid lake sturgeon (*Acipenser fulvescens*) reveal a complex evolutionary history of genome duplication and subsequent diploidization*, by Matthew C Hale, Jeffrey R. Lucas, and J Andrew DeWoody.

***These manuscripts have not yet been published, and thus we respectfully request that they not be distributed.**