SHORT COMMUNICATION

doi: 10.1111/age.12211

Transcriptome annotation and marker discovery in white bass (*Morone chrysops*) and striped bass (*Morone saxatilis*)

Chao Li*, Benjamin H. Beck†, S. Adam Fuller† and Eric Peatman*

*School of Fisheries, Aquaculture and Aquatic Sciences, Auburn University, Auburn, AL 36849, USA. †United States Department of Agriculture, Agricultural Research Service, Stuttgart National Aquaculture Research Center, Stuttgart, AR 72160, USA.

Summary

Striped bass (*Morone saxatilis*) and white bass (*Morone chrysops*) are the parental species of the hybrid striped bass, a major U.S. aquaculture species. Currently, genomic resources for striped bass, white bass, and their hybrid lag behind those of other aquaculture species. Current resources consist of a medium-density genetic linkage map and a well-annotated ovarian transcriptome. A well-annotated transcriptome from across striped bass and white bass tissues is needed to advance both broad-based RNA-seq studies of gene expression as well as aid in more targeted studies of important genes and pathways critical for reproductive physiology and immunity. Here, we carried out Illumina-based transcriptome sequencing and annotation in both species utilizing the TRINITY and TRINOTATE packages. The assembled Moronid reference transcriptomes and identified SSRs and SNPs should advance ongoing studies of reproduction, physiology, and immunology in these species and provide markers for broodstock management and selection.

Keywords RNA-seq, SNP, Morone, gene assembly

The white bass (*Morone chrysops*) and the striped bass (*M. saxatilis*) are temperate basses with high ecological, recreational, and commercial value in North America. Their hybrid (*M. chrysops* × *M. saxatilis*) is a major U.S. aquaculture species, whereas the parental species serve as important teleost models for reproductive physiology (Beck *et al.* 2012; Zmora *et al.* 2014). Genetic information has been restricted to a single-tissue (ovary) transcriptome and a microsatellite linkage map from striped bass (Liu *et al.* 2012; Reading *et al.* 2012), limiting gene discovery and expression and functional studies in the two species and their hybrid. Here, we set out to produce well-annotated transcriptomes for both species to advance future broad-based RNA-seq studies of gene expression as well as to aid in more targeted studies of important genes and pathways.

Major tissues and organs (brain, liver, spleen, trunk kidney, ovary, testes, gill, and intestine) from 10 individuals from white bass and 10 individuals from striped bass were harvested, and equal amounts of tissue from each system were pooled prior to RNA extraction. The result was two master pools of RNA, one for each species. Each pool was used for library construction and sequencing in a lane of Illumina HiSeq2000 with 100-bp paired-end sequencing at HudsonAlpha Institute. Reference transcriptomes for each species were generated using the TRINITY *de novo* assembly software with the CuffFly option to get the fewest isoforms per gene that are well supported by the reads (Grabherr *et al.* 2011). Reference contigs were annotated by BLAST against the UniProtKB/Swiss-Prot database and the non-redundant (nr) protein database using the BLASTX program. To achieve a comprehensive annotation, the transcriptome functional annotation and analysis software TRINOTATE was utilized (Haas *et al.* 2013; http://trinotate.sourceforge.net). The pipeline includes protein domain identification (HMMER/PFAM), protein signal prediction (SignalP/tmHMM), and comparison to currently curated annotation databases (EMBL/Uniprot/eggNOG/GO Pathways databases). All functional annotation data derived from the analysis of transcripts were integrated into a SQLite database, which allows for fast, efficient searching for terms with specific qualities related to a desired scientific hypothesis or a means to create a whole annotation report for a transcriptome. In each species, we implemented the microsatellite-marker mining process in MSATFINDER with a repeat threshold of eight dinucleotide repeats or five tri-, tetra-, penta-, or hexa-nucleotide repeats. Transcriptome-wide species-specific SNPs were also identified between white bass and striped

Address for correspondence

E. Peatman, School of Fisheries, Aquaculture and Aquatic Sciences, Auburn University, Auburn, AL 36849, USA.
E-mail: peatmer@auburn.edu

Accepted for publication 9 July 2014

Table 1 Summary of TRINITY *de novo* assembly results of Illumina RNA-seq data from striped bass and white bass.

	Striped bass	White bass
Contigs	203 587	185 531
Largest contig (bp)	21 100	28 262
Large contigs (≥1000 bp)	68 395	66 891
Large contigs (≥500 bp)	98 864	94 485
N50 (bp)	2915	3132
Average contig length (bp)	1263	1371

bass using the program POPOOLATION2 (Kofler *et al.* 2011) using the default parameters of minor allele frequency of 0.1 and minor allele count of 6.

A total of 262×10^6 total high-quality reads were obtained with 135×10^6 reads from striped bass and 127×10^6 reads from white bass. Using the TRINITY *de novo* assembly software, reads were assembled into 203 587 striped bass unique contigs and 185 531 white bass unique contigs. N50 and average contig sizes were 2915 and 1263 bp respectively for striped bass, and 3132 and 1371 bp respectively for white bass (Table 1). These included 166 867 and 185 351 transcripts that were identified for the first time in striped bass and white bass respectively. Annotation was carried out by BLAST against the UniProt and NR (NCBI non-redundant) databases for both species. At an *E*-value ≤1*e*−5, 21 186 and 29 624, unigene matches were obtained against the UniProt and NR databases respectively in striped bass, and 21 001 and 28 906 matches were returned in white bass against the same databases. Of these NR matches, 25 902 (87.4%) in striped bass and 25 484 (88.2%) in white bass were predicted to have full-length transcript coverage based on TRINITY analysis. Using more stringent criteria, similar results were obtained from both species, with 18 630 UniProt and 23 605 NR annotated unigenes in striped bass and 18 584 UniProt and 22 354 NR annotated unigenes in white bass (score ≥100, *E*-value ≤1*e*−20; Table 2). As part of TRINOTATE, Gene Ontology (GO) processes generated 154 390 GO terms based on UniProt annotation for striped bass and 157 966 GO terms for white bass, with similar distributions of terms between species across the biological process, molecular function, and cellular component categories (Fig. S1). TRINOTATE contig annotations from the striped bass and white bass transcriptomes are available in Tables S1 and S2. The striped bass and white bass transcriptome contigs are available on NCBI's Transcriptome Shotgun Assembly (TSA) database with Accessions GAZY00000000 (white bass) and GBAA00000000 (striped bass).

In both species, the transcriptomes yielded microsatellite and SNP markers valuable in future downstream analyses (Tables S3 and S4). In striped bass, from a total of 32 111 microsatellites identified by MSATFINDER, 36.05% (n = 11 577) had sufficient flanking regions to allow design of primers. These 11 577 microsatellites were distributed across 10 055 contigs (Table S3). Similarly, in white bass, from a total of 30 408 microsatellites, 34.53% (n = 10 500) had sufficient flanking regions to allow design of primers. These 10 500 microsatellites were distributed across 9054 contigs. A SNP analysis comparing between species yielded 2220 markers polymorphic in one species but not the other, including 1661 SNPs associated with genes. These markers may prove useful in population genetics analyses aimed at assessing hybridization of Moronid basses in the wild. Additionally, in the future, the reference transcriptomes will serve as an important sequence anchor for short-read genotyping studies using techniques such as RAD-seq or GBS (Davey *et al.* 2011).

Table 2 Summary of gene identification and annotation of assembled striped bass and white bass contigs based on BLAST homology searches against various protein databases (UniProt and NR). Putative gene matches were at *E*-value ≤1*e*−5. Hypothetical gene matches denote those BLAST hits with uninformative annotation. Quality unigene hits denote more stringent parameters, including score ≥100, *E*-value ≤1*e*−20.

	Striped bass		White bass	
	UniProt	NR	UniProt	NR
Contigs with putative gene matches	69 134	79 062	68 312	76 884
Annotated contigs ≥1000 bp	54 487	58 316	54 430	57 839
Annotated contigs ≥500 bp	62 602	68 876	62 158	67 682
Unigene matches	21 186	29 624	21 001	28 906
Hypothetical gene matches	0	1901	0	1858
Quality unigene matches	18 630	23 605	18 584	22 354

The TRINITY-based assembly of the white bass and striped bass transcriptomes generated high-quality, gene-length transcripts, which will be of great utility in future expression and functional studies in Moronid species. Microsatellite and SNP markers identified at the same time are expected to aid in aquaculture, conservation, and sportfish genetic management and improvement.

Acknowledgements

The authors wish to thank the National Animal Genome Resource Support Program (NRSP-8) for funding support for this project.

Data accessibility

Raw Illumina reads: NCBI SRA: SRP039910. Transcriptome Assembly: TSA Accessions GAZY00000000 (white bass) and GBAA00000000 (striped bass).

References

Beck B.H., Fuller S.A., Peatman E., McEntire M.E., Darwish A. & Freeman D.W. (2012) Chronic exogenous kisspeptin administration accelerates gonadal development in basses of the genus

Morone. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* **162**, 265–73.

Davey J.W., Hohenlohe P.A., Etter P.D., Boone J.Q., Catchen J.M. & Blaxter M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499–510.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R. & Zeng Q. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–52.

Haas B.J., Papanicolaou A., Yassour M., Grabherr M., Blood P.D., Bowden J., Couger M.B., Eccles D., Li B. & Lieber M. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the TRINITY platform for reference generation and analysis. *Nature Protocols* **8**, 1494–512.

Kofler R., Pandey R.V. & Schlötterer C. (2011) POPOOLATION2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**, 3435–6.

Liu S., Rexroad C.E. III, Couch C.R., Cordes J.F., Reece K.S. & Sullivan C.V. (2012) A microsatellite linkage map of striped bass (*Morone saxatilis*) reveals conserved synteny with the three-spined stickleback (*Gasterosteus aculeatus*). *Marine Biotechnology* **14**, 237–44.

Reading B.J., Chapman R.W., Schaff J.E., Scholl E.H., Opperman C.H. & Sullivan C.V. (2012) An ovary transcriptome for all maturational stages of the striped bass (*Morone saxatilis*), a highly advanced perciform fish. *BMC Research Notes* **5**, 111.

Zmora N., Stubblefield J., Golan M., Servili A., Levavi-Sivan B. & Zohar Y. (2014) The medio-basal hypothalamus as a dynamic and plastic reproduction related kisspeptin-gnrh-pituitary center in fish. *Endocrinology* **155**, 1874–86.

Supporting information

Additional supporting information may be found in the online version of this article.

Figure S1 Gene ontology (GO) term categorization and distribution of striped bass and white bass transcriptome. GO terms were processed by BLAST2GO and categorized at level 2 under three main categories (cellular component, molecular function, and biological process).

Table S1a Transcriptome annotation of striped bass using TRINOTATE analysis with all, hits and unique contig tabs as annotated based on BLASTP hits to the UniProt database.

Table S1b Transcriptome annotation of striped bass using TRINOTATE analysis with all, hits and unique contig tabs as annotated based on BLASTP hits to the UniProt database (continued).

Table S2a Transcriptome annotation of white bass using TRINOTATE analysis with all, hits and unique contig tabs as annotated based on BLASTP hits to the UniProt database.

Table S2b Transcriptome annotation of white bass using TRINOTATE analysis with all, hits and unique contig tabs as annotated based on BLASTP hits to the UniProt database (continued).

Table S3 Statistics of simple sequence repeats (SSRs) and SNPs identified from the striped and white bass transcriptomes.

Table S4 Contigs containing SSR and SNP markers in the Moronid bass transcriptomes.