A comprehensive non-redundant reference transcriptome for the Atlantic silverside *Menidia menidia*

Nina Overgaard Therkildsen[a,1,*] and Hannes Baumann[b]

[a]Department of Biology, Hopkins Marine Station, Stanford University, 120 Ocean View Blvd, CA-93950 Pacific Grove, USA

[b]Department of Marine Sciences, University of Connecticut, 1080 Shennecossett Road, CT-06340 Groton, USA

[1]Current address: Department of Natural Resources, Cornell University, 208 Fernow Hall, NY-14853 Ithaca, USA

*Correspondence to: Nina Overgaard Therkildsen (nt246@cornell.edu)
Department of Natural Resources, Cornell University, 208 Fernow Hall, NY-14853 Ithaca, USA

Declarations of interest: none

ABSTRACT

The Atlantic silverside (*Menidia menidia*) has been the focus of extensive research efforts in ecology, evolutionary biology, and physiology over the past three decades, but lack of genomic resources has so far hindered examination of the molecular basis underlying the remarkable patterns of phenotypic variation described in this species. We here present the first reference transcriptome for *M. menidia*. We sought to capture a single representative sequence from as many genes as possible by first using a combination of Trinity and the CLC Genomics Workbench to *de novo* assemble contigs based on RNA-seq data from multiple individuals, tissue types, and life stages. To reduce redundancy, we passed the combined raw assemblies through a stringent filtering pipeline based both on sequence similarity to related species and computational predictions of transcript quality, condensing an initial set of >480,000 contigs to a final set of 20,998 representative contigs, amounting to a total length of 53.3 Mb. In this final assembly, 91% of the contigs were functionally annotated with putative gene function and gene ontology (GO) terms and/or InterProScan identifiers. The assembly contains complete or nearly complete copies of >95% of 248 highly conserved core genes present in low copy number across higher eukaryotes, and partial copies of another 3.8%, suggesting that our assembly provides relatively comprehensive coverage of the *M. menidia* transcriptome. The assembly provided here will be an important resource for future research.

Keywords: Transcriptome, *de novo* assembly, RNA-seq, Atlantic silverside, *Menidia menidia*

INTRODUCTION

The Atlantic silverside *Menidia menidia* (Atherinidae) is an abundant forage fish that inhabits nearshore environments along the east coast of North America, from northern Florida, USA to the Gulf of St. Lawrence, Canada (Hice *et al.* 2012). Its broad distribution along one of the steepest latitudinal temperature gradients in the world, combined with its ecological importance, its semelparous annual life cycle, and the relative ease with which it can be reared in the laboratory has made the Atlantic silverside a valuable model species for ecological and evolutionary research over the past three decades. Extensive laboratory and field studies have, for example, shown that the Atlantic silverside shows a remarkable degree of either co-gradient or counter-gradient variation in a suite of traits across latitudes, including growth rates, fecundity, metabolic rates, vertebral counts, swimming performance, and predator avoidance (reviewed in Conover *et al.* 2005). Common garden experiments have established that these trait differences have a clear genetic basis and often vary between locations less than 100 km apart—spatial scales across which silverside populations mix extensively (Hice *et al.* 2012). This detailed demonstration of pronounced genetic trait differences maintained despite strong gene flow has played an important role in shifting earlier perceptions about local adaptation being rare or absent in highly connected marine environments (Conover *et al.* 2005).

The Atlantic silverside has also been important for studies of rapid adaptation, providing crucial experimental evidence for fishing causing rapid evolution in the exploited populations (Conover & Munch 2002). It also provided the first discovery of temperature-dependent sex determination in fishes (reviewed in Conover *et al.* 2005). More recently, it has been an important model for quantifying novel effects of climate stressors such as ocean warming, acidification and reduced oxygen levels (e.g. Murray *et al.* 2017; Baumann *et al.* 2018) and for examining the geographic distribution of environmental contaminants (Baumann *et al.* 2016). Its close relative, the inland silverside (*M. beryllina*) is also frequently used in ecotoxicology studies (e.g. Jeffries et al. 2015). The Atlantic silverside is therefore a central species for diverse research programs, yet genomic resources have not been available for exploring the molecular basis underlying the many fascinating evolutionary, ecological, and physiological patterns it exhibits.

This paper outlines how we generated and annotated the first comprehensive, non-redundant *de novo* reference transcriptome for the Atlantic silverside. We needed this reference for 'in silico exome capture' (Therkildsen & Palumbi 2017, Therkildsen et al. 2019) that would let us survey genome-wide patterns of variation underlying local adaptation and rapid fisheries-induced evolution in this species in a cost-effective way. *De novo* transcriptome assemblies often contain considerable redundancy with different allelic variants, transcript splice variants, or overlapping fragments of the same transcript being represented by separate contigs. This redundancy provides important information for some types of analysis, but because we wanted to use the transcriptome as a reference for population genomic analysis, our goal was to identify just a single representative complete transcript for each gene, so that genomic sequencing reads could map to unique positions. By pooling RNA samples from multiple individuals, tissue types, and life stages, assembling contigs with two different *de novo* assembly algorithms, and passing the resulting raw assemblies through a stringent filtering pipeline based both on sequence similarity to related species and computational predictions of transcript quality, we successfully minimized redundancy while capturing and retaining maximal diversity of transcripts in our final assembly. The *de novo* Atlantic silverside transcriptome presented here with associated functional annotation will be an important resource for future research.

DATA DESCRIPTION

SAMPLES, LIBRARY PREPARATION, AND SEQUENCING

To capture a broad diversity of transcripts expressed at different life stages and in different tissue types, we based our RNA sequencing on five larval and three adult *M. menidia* (Table 1). The adults were collected directly from the wild at Poquot Beach (NY, N 40.9475, W 73.1025) in June 2013. At the same time, we also collected three pairs of parent fish that were strip-spawned to produce three groups of full-sib larvae for rearing in the laboratory following the procedure described by Murray et al. (2014). Twelve days post hatching, the larvae were sacrificed and all samples were stored in RNAlater. All animal handling was in accordance with NIH guidelines and approved under Institutional Animal Care and Use Committee (IACUC) protocol 2010-1842-F at Stony Brook University.

We extracted total RNA from all samples with the Qiagen RNeasy Plus Universal Tissue Mini Kit (Qiagen GmbH, Hilden, Germany). For each of five larvae, we used the entire animal in a single extraction. For each adult, we did separate extractions for different tissue types (including brain, heart, liver, gonad, muscle, gill, skin, spinal cord, fin, eye) and pooled even quantities of these extracts for each individual. We then prepared a single individually indexed cDNA library for each fish (pooled extracts from all tissue types in a single library) with Illumina's TruSeq RNA sample prep kit v2 (Illumina Inc., San Diego, CA, USA). All eight libraries were sequenced in a single lane of 100 bp paired-end reads on an Illumina HiSeq 2000 at the University of Utah's Bioinformatics Core Facility, yielding a total of 170 million raw sequence read pairs (between 16 and 26 million read pairs per individual, amounting to a total of 34 Gb sequence). Our workflow for processing the raw reads are shown in Fig. 1

DATA QUALITY FILTERING

After removing exact duplicate read pairs (~13% of the total) with the program Fastuniq v1.1 (Xu *et al.* 2012), we used Trimmomatic v0.32 (Bolger *et al.* 2014) to trim off adapter sequence and the first base of each read (because of a highly inflated C-content at this position). We also used Trimmomatic's sliding window approach to trim off the rest of the read if the average sequence quality over any four bases fell below 20, and discarded reads shorter than 50 bp after this filtering (~7% of reads). We conservatively discarded a further 2.6% of reads because they mapped to potential contamination databases (human, bacterial, viral, rRNA, and Artemia (feed for the larvae)) with bowtie2 v2.2.3 (Langmead & Salzberg 2012) in 'sensitive' preset mode. Finally, we used the program FLASH v. 1.2.9 (Magoč & Salzberg 2011) with default settings to merge overlapping read ends into single consensus sequences (merging 68% of the remaining pairs), resulting in a filtered data set of 92 million merged reads (length 50-240 bp) and 43 million read pairs amounting in total to 21.1 Gb sequence.

DE NOVO ASSEMBLY AND REDUNDANCY REDUCTION

Because different assembly algorithms and parameter settings may recover different transcripts, we used two different programs to *de novo* assemble the pooled set of filtered RNA-seq reads from all libraries. First, we generated two assemblies with the CLC

Genomic Workbench v6.0.2 (CLC Bio), one using the automatically optimized parameters (word size 25, bubble size 50) and one using a larger word size (k-mer) of 40 to facilitate more contiguous and accurate assembly of highly expressed transcripts. For both assemblies, we mapped reads back to the initial contigs to update the consensus sequence, and we broke up scaffolded sequence with no read support, only maintaining contigs >200 bp. In parallel, we assembled the reads with Trinity v. r20131110 (Grabherr *et al.* 2011) using the default settings (including a fixed k-mer size of 25). The Trinity output explicitly clusters related 'isoforms', and since we were only interested in retaining a single representative transcript for each gene, we mapped all reads back to the assembly and extracted only the isoform with the highest mapped read depth within each subcomponent following the procedure by Yang and Smith (2013). The three *de novo* assemblies contained between 135,931 and 193,079 contigs each, for a total of 483,424 contigs in the combined set (Table 2).

A comparison of the three assemblies (CLC (k-mer 25), CLC (k-mer 40), and Trinity (single isoform per cluster)) with blastn v2.2.29+ (Camacho *et al.* 2009) revealed that only 78-87% of the contigs in each assembly had significant hits (e-value <10^{-3}) to the other assemblies, indicating that each assembly contained a set of unique transcripts. To maintain maximal transcript diversity, we therefore proceeded with a merged set of all three assemblies. The merged assemblies contain substantial redundancy, so to collapse the contig set into the longest representative for each unique sequence we used cd-hit-est v4.5.4 (Li & Godzik 2006) to remove shorter contigs that showed >95% sequence similarity to other contigs. Due to assembly challenges, some genes could also be presented by multiple different fragments rather than a transcript of complete length, so to join partial assemblies (fragments) of the same transcript, we used CAP3 v12/21/07 (Huang 1999) to meta-assemble contigs with >95% similarity over at least 100 bp (an approach shown to improve the quality of transcriptome assemblies e.g. by Melchior et al. (2014)). Since both the *de novo* assembly processes and the meta-assembly may introduce chimeric contigs, we used the method by Yang and Smith (2013) to break up likely chimeras (observed in 0.8% of transcripts) based on separate blastx comparisons to the peptide sets for three reference fish species (see below). The resulting redundancy-reduced contig set contained 177,877 contigs (Table 2).

SELECTING PUTATIVE GENE ORTHOLOGS

Because we wanted to reduce our contig set to only include a single representative transcript for each gene, we used a reciprocal best hit blast approach to extract non-redundant putative orthologs to the gene sets in the three most closely related species for which annotated genome assemblies were available at the time: platyfish (*Xiphophorus maculatus*), medaka (*Oryzias latipes*), and Nile tilapia (*Oreochromis niloticus*). We compared our contig set against the full peptide set for each reference species (downloaded from Ensembl release 75 (Zerbino *et al.* 2018)) with blastx, and then compared the peptide sequences for each species to our contig set with tblastn. For each reference species, we recorded reciprocal best hits (RBHs) when a contig and a protein had a best match to each other (e-value<10^{-4}). We then used a sequential approach to select a combined set of 19,349 contigs in our *Menidia* assembly that were RBHs (and therefore putative orthologs) to a unique peptide sequence in at least one of the reference species (Supplementary Note).

Because all our reference species diverged from the silverside >75 million years ago (Near *et al.* 2012; Campanella *et al.* 2015), the RBH contig selection procedure will fail to identify recently diverged genes. To recover additional high quality non-redundant transcripts, we used TransDecoder v. r20131110 (Haas *et al.* 2013) to predict coding regions in our redundancy-reduced contig set on the basis of nucleotide composition, open reading frame (ORF) length and Pfam domain content. Transdecoder predicted candidate coding sequence of at least 100 amino acids in 39,604 contigs and of the 15,222 that contained complete length ORFs, we retained 1,961 which did not have a significant (e-value<10^{-2}) blastn hit to the RBH contig set (and therefore are non-redundant). To minimize potential contamination in our final assembly, we compared the joined contig set (RBH contigs and non-redundant contigs with complete ORFs) to the NCBI non-redundant protein database (NR) (downloaded on July 14 2014) with blastx and used the program MEGAN v. 5.7.1 (Huson *et al.* 2011) to identify and remove 311 contigs with best hits to non-chordate taxa or to human sequence, ending up with a final reference transcriptome contig set of 20,998 contigs.

FUNCTIONAL ANNOTATION

The final contig set was functionally annotated with the Blast2GO v3.1.2 suite (Conesa *et al.* 2005). For each sequence, we imported significant hits (e-value < 10^{-6}) from blastx searches against the UniProt Swiss-Prot and the NCBI non-redundant (NR) protein databases and used Blast2GO's Blast Description Annotator tool to select the most informative and relevant descriptor before mapping GO (Gene Ontology) terms to the matches and applying the built-in annotation rule with the default parameters and evidence code weights. We also imported GO-terms associated with the reciprocal-best-hit genes in the reference fish species, and merged the combined sets of assigned annotations with GO-terms inferred from InterProScan analysis of each sequence. As a final step, we used the Blast2GO Validate Annotations tool to ensure that no parent-child redundancy was present in the assigned GO-terms, and we applied the Annex tool to augment the annotation based on inference of biological processes from commonly associated molecular functions and cellular components. This way we obtained a total of 490,807 GO-terms annotated to 19,117 of the contigs (91% of all contigs; the median number of GO-terms per contig was 19, Table S1).

EVALUATION OF COMPLETENESS AND UNIQUENESS

The final non-redundant transcriptome assembly had significant blastx hits to 84% of gene models in the platyfish genome (81% of these were reciprocal best hits), and for 74% of these genes the top high-scoring segment pair covered >90% of the total length of the reference peptide sequence, indicating complete or nearly complete transcripts. CEGMA v2.5 (Parra *et al.* 2007) also detected complete or nearly complete copies of >95% of 248 highly conserved core genes present in low copy number across higher eukaryotes (and partial copies of another 3.8%). Similarly, BUSCO (Simão et al. 2015) analysis flagged only 3.1% of 4,584 highly conserved genes in Actinopterygii (ray-finned fish) species as missing from the assembly (90.5% of these reference genes were detected as complete copies, 6.4% as fragmented), further suggesting that the assembly provides a relatively comprehensive coverage of the *M. menidia* transcriptome. The extensive transcript diversity is likely caused by the inclusion of many different tissue types across two life stages in our RNA-seq libraries.

Exposure to a variety of stressors prior to RNA harvesting may have increased transcript diversity further and could be pursued in future work targeting specific response pathways.

The number of contigs in our final assembly is much closer to the number of coding genes found in related species (21,437 - 23,774 for medaka, platyfish, and tilapia, (Zerbino *et al.* 2018)) than any of the larger assemblies. Yet, the strict redundancy reduction did result in a small loss of transcripts diversity as the complete Trinity assembly and the merged set of all raw assemblies actually included up to 100% of the CEGMA genes. However, this loss of diversity was compensated for by much better mapping specificity. For the full Trinity assembly and the total merged assembly, >93% of the cleaned RNA-seq reads mapped back to the *de novo* reference with bowtie2 v2.2.3 (Langmead & Salzberg 2012) in the 'sensitive' preset mode, but only 61% (Trinity) or 9% (total merged) of these mapped to a unique position (the remaining reads mapping to multiple contigs, Table 2). In contrast, almost all (98%) of the 74% of RNA-seq reads that mapped to the final assembly mapped only to a single position, suggesting that most genes are only represented by a single contig and that this assembly therefore will be useful reference for mapping genomic reads, as further demonstrated in Therkildsen and Palumbi (2017). In addition to the highly non-redundant assembly that will be useful for population genomics and many other purposes, we are also making each of our intermediate larger assemblies (see Fig. 1 and Table 2) available as supplementary data files (File S2 and S3) for other types of analysis that specifically targets redundancy among similar transcripts, e.g. analysis of splice variation or variation within closely related gene families. With recent technological advances, *de novo* assembly of the entire genome is an increasingly attainable goal for many non-model organisms. Yet, the cost and effort involved in assembling only the transcriptome generally is still much lower, so an important role remains for reference transcriptomes - especially for studies focusing on functional genomic variation.

Acknowledgements

This work was funded by a National Science Foundation (USA) grant to Stephen R. Palumbi [OCE-1434325] and a Villum Foundation (Denmark) postdoctoral fellowship to Nina Overgaard Therkildsen.

References

Baumann H, Cross EL, Murray CS (2018) Robust quantification of fish early life CO_2 sensitivities via serial experimentation. *Biology Letters*, **14**, 20180408.

Baumann Z, Mason RP, Conover DO, Balcom P, Chen CY, Buckman KL, Fisher NS, Baumann H (2016) Mercury bioaccumulation increases with latitude in a coastal marine fish (Atlantic silverside, *Menidia menidia*). *Canadian Journal of Fisheries and Aquatic Sciences*, **74**, 1009–1015.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Camacho C, Coulouris G, Avagyan V, Ma N (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Campanella D, Hughes LC, Unmack PJ, Bloom DD, Piller KR, Orti G (2015) Multi-locus fossil-calibrated phylogeny of Atheriniformes (Teleostei, Ovalentaria). *Molecular Phylogenetics and Evolution*, **86**, 8–23.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Conover DO, Munch SB (2002) Sustaining fisheries yields over evolutionary time scales. *Science*, **297**, 94–96.

Conover DO, Arnott SA, Walsh MR, Munch SB (2005) Darwinian fishery science: lessons from the Atlantic silverside (*Menidia menidia*). *Canadian Journal of Fisheries and Aquatic Sciences*, **62**, 730–737.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Hice LA, Duffy TA, Munch SB, Conover DO (2012) Spatial scale and divergent patterns of variation in adapted traits in the ocean. *Ecology Letters*, **15**, 568–575.

Huang X (1999) CAP3: A DNA Sequence Assembly Program. *Genome Research*, **9**, 868–877.

Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, **21**, 1552–1560.

Jeffries KM, Brander SM, Britton MT, Fangue NA, Connon RE (2015) Chronic exposures to low and high concentrations of ibuprofen elicit different gene response patterns in a euryhaline fish. *Environmental Science and Pollution Research*, **22**, 17397–17413.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*.

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Melicher D, Torson AS, Dworkin I, Bowsher JH (2014) A pipeline for the de novo assembly of the *Themira biloba* (Sepsidae: Diptera) transcriptome using a multiple k-mer length approach. *BMC Genomics*, **15**, 1–13.

Murray CS, Fuiman LA, Baumann H (2017) Consequences of elevated CO_2 exposure across multiple life stages in a coastal forage fish. *ICES Journal of Marine Science*, **74**, 1051–1061.

Murray CS, Malvezzi A, Gobler CJ, Baumann H (2014) Offspring sensitivity to ocean acidification changes seasonally in a coastal marine fish. *Marine Ecology Progress Series*, **504**, 1–11.

Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, Wainwright, P.C., Friedman M, Smith WL (2012) Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 13698–13703.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–3212.

Therkildsen NO, Palumbi SR (2017) Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, **17**, 194–208.

Therkildsen NO, Wilder AP, Conover DO, Munch SB, Baumann H, Palumbi SR (2019) Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. *Science*, **365**, 487–490.

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S (2012) FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS ONE*, **7**, e52249.

Yang Y, Smith SA (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, **14**, 328.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR *et al.* (2018) Ensembl 2018. *Nucleic Acids Research*, **46**, D754–D761.

Data accessibility

The raw sequence data are deposited in the NCBI Sequence Read Archive (SRA) with accession numbers SRR3990241- SRR3990248 associated with BioProject PRJNA330848. The final assembly of the Atlantic silverside transcriptome (20,998 contigs) is deposited in the NCBI GenBank Transcriptome Shotgun Assembly Sequence Database

(TSA) under Accession no. GEVY00000000. The transcriptome annotation table is provided as Supplementary Table S1, and the full merged and redundancy-reduced assemblies are provided in fasta format as Supplementary Files S2 and S3.

List of supplementary files:

Table S1. Annotation table for the final transcriptome assembly:
MenidiaTranscriptome_AnnotationTable_GO_Interpro.csv

File S1: Supplementary Note

File S2: Fasta file with the total combined contig set (483,424 contigs):
MenidiaTranscriptome_Complete_Merged_Contig_Set.fa

File S3: Fasta file with the redundancy-reduced contig set (177,877 contigs):
MenidiaTranscriptome_RedundancyReduced_Merged_Contig_Set.fa

Figure legends

Fig. 1. Diagram showing the sequence of steps in our bioinformatic workflow for cleaning the RNA-seq read data, *de novo* assembly, and redundancy reduction. Yellow boxes

represent RNA-seq read data, blue boxes represent data processing steps, and red boxes represent transcriptome assemblies. Statistics such as the total number of contigs, the total assembled length and the proportion of conserved core genes found in each of the intermediate assemblies are provided in Table 2.