Marine Genomics 15 (2014) 13–15

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Genomics/Technical resources

Sequencing and characterization of the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome

Samuel E. Fox [a], Mark R. Christie [b,*], Melanie Marine [b], Henry D. Priest [a,c,d], Todd C. Mockler [a,c,d], Michael S. Blouin [b]

[a] Department of Botany and Plant Pathology, Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR, USA
[b] Department of Zoology, Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR, USA
[c] Donald Danforth Plant Science Center, Saint Louis, MO 63132, USA
[d] Division of Biology and Biomedical Sciences, Washington University, St. Louis, MO 63110, USA

ARTICLE INFO

Article history:
Received 20 November 2013
Accepted 2 December 2013
Available online 17 January 2014

Keywords:
Aquaculture
Fisheries
Hatcheries
RNA-Seq
Transcriptome

ABSTRACT

Identifying the traits that differ between hatchery and wild fish may allow for pragmatic changes to hatchery practice. To meet those ends, we sequenced, assembled, and characterized the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome. Using the Illumina sequencing platform, we sequenced nearly 41 million 76-mer reads representing 3.1 Gbp of steelhead transcriptome. Upon final assembly, this sequence data yielded 86,402 transcript scaffolds, of which, 66,530 (77%) displayed homology to proteins of the non-redundant NCBI database. Gene descriptions and gene ontology terms were used to annotate the transcriptome resulting in 4030 unique gene ontology (GO) annotations attributed to the assembled sequences. We also conducted a comparative analysis that identified homologous genes within four other fish species including zebrafish (*Danio rerio*), stickleback (*Gasterosteus aculeatus*), and two pufferfish species (*Tetraodon nigroviridis* and *Takifugu rubripes*). Comparing our steelhead reference assembly directly to the transcriptome for rainbow trout (the fresh water life-history variant of the same species) revealed that while the steelhead and rainbow trout transcriptomes are complementary, the steelhead data will be useful for investigating questions related to anadromous (ocean-going) fishes. These sequence data and web tools provide a useful set of resources for salmonid researchers and the broader genomics community (available at http://salmon.cgrb.oregonstate.edu).

1. Introduction

Despite the global economic and environmental importance of salmon, genomic resources for the study of these anadromous fishes are limited. Here we use RNA-Seq to characterize the transcriptome of steelhead (ocean-going *Oncorhynchus mykiss*). The use of next-generation platforms for de novo sequencing of transcriptomes has been repeatedly demonstrated to be suitable for marker and gene discovery, comparative analysis, and gene expression analysis. For example, high throughput sequencing has been used for transcriptome assembly and annotation in several fishes including sea bream, guppy, Atlantic cod, mud loach, and rainbow trout (Calduch-Giner et al., 2013; Fraser et al., 2011; Johansen et al., 2011; Long et al., 2013; Salem et al., 2010). Rainbow trout and steelhead are different life-history forms of the same species (*O. mykiss*), however, the freshwater-resident rainbow trout and ocean-going steelhead differ behaviorally, phenotypically, and physiologically (Hale et al., 2013; Hayes et al., 2012). In 2010, a 454-based transcriptome was published for rainbow trout (Salem et al., 2010), but no transcriptome data are currently available for steelhead. The aim of this study was to assemble, annotate, and analyze a high quality reference transcriptome that will enable researchers to assess gene expression levels, conduct comparative analyses, and identify and utilize molecular markers in the anadromous *O. mykiss* steelhead.

2. Data description

2.1. Salmon collection and genotyping

The steelhead for this study were collected from the Hood River, in Oregon. During the winter-runs of 2008 and 2010, Oregon Department of Fish and Wildlife (ODFW) personnel collected wild-born (W) and first-generation hatchery-born (H) steelhead. WxW and HxH crosses were completed by ODFW personnel at Parkdale Hatchery, OR, and the fertilized eggs were raised using standard hatchery protocol at Oak Springs Hatchery, OR (53 °F). As soon as yolk sack absorption was complete, the fry were frozen in liquid nitrogen and stored at −80 °C. DNA was extracted using a standard protocol for the DNeasy Blood & Tissue Kit (Qiagen). The sex of the fry was established by genotyping all fish with a sex-specific marker OmyY1 primer at an annealing

* Corresponding author at: Department of Zoology, 3029 Cordley Hall, Oregon State University, Corvallis, Oregon, USA. Tel.: +1 541 231 0719.
E-mail address: christim@science.oregonstate.edu (M.R. Christie).

1874-7787/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.margen.2013.12.001

Table 1
Summary of the Illumina sequencing of *O. mykiss*. This table shows the samples and numbers of reads generated for use in transcriptome assembly and expression analysis. Greater than 40 million 76mer Illumina reads, totaling approximately 3.1 gigabases of sequence were used to assemble the reference transcriptome.

	Sequence reads	Bases (bp)
All prefiltered Illumina reads	68,445,070	5,475,605,600
Hatchery male paired end reads	8,540,846	649,104,296
Hatchery female paired end reads	9,807,973	745,405,948
Wild male paired end reads	6,190,274	470,460,824
Wild female paired end reads	6,681,589	507,800,764
Total paired end reads	31,220,682	2,372,771,832
Hatchery male single end reads	3,025,505	229,938,380
Hatchery female single end reads	2,840,043	215,843,268
Wild male single end reads	1,675,012	127,300,912
Wild female single end reads	2,213,258	168,207,608
All single end reads	9,753,818	741,290,168
Total used for transcriptome assembly	40,974,500	3,114,062,000

temperature of 60 °C (Brunelli et al., 2008). For the transcriptome assembly we used paired-end sequences from one male and one female WxW fish, and from one male and one female HxH fish from the 2008 crosses, supplemented with single-end 80 bp reads from 4 male and 5 female HxH fish, and 4 male and 2 female WxW fish from the 2010 crosses.

2.2. Sample preparation and Illumina sequencing

Total RNA was isolated using a modified protocol described in detail elsewhere (Fox et al., 2009). RNA was extracted using Trizol Reagent (Invitrogen). Total RNA was treated for 10 min at 65 °C with RNAsecure reagent (Ambion). To eliminate genomic DNA amplification, all RNA preparations were treated for 15 min at 37 °C with RNase-free Turbo DNase (Ambion). Total RNA was further purified using RNAeasy Mini RNA kit (Qiagen) according to the manufacturer's protocol. Isolation of mRNA essentially free of ribosomal and other non-polyadenylated RNAs was critical for generation of non-biased randomly primed (RP) libraries. For the creation of RP cDNA libraries, poly(A) mRNA was isolated by two consecutive purification cycles on oligo d(T) cellulose using a Micro-PolyA-Purist kit (Ambion). Concentration, integrity and extent of contamination by ribosomal RNA were assessed using ND-1000 spectrophotometer (Thermo Fisher Scientific) and Bioanalyzer 2100 (Agilent Technologies).

For RP cDNA libraries, first strand cDNA was synthesized using 1 μg of poly(A) mRNA. Random hexamer primers (300 ng per μg of RNA), and Superscript III reverse transcriptase (Invitrogen) were added to the reaction and incubated at 75 °C for 5 min. Second strand cDNA was synthesized by combining 20 μL of the 1st strand reaction, 8 μL of 10× Klenow Buffer (New England Biolabs; NEB), 1 unit of RNase H (Invitrogen), 68.8 μL of water and 30 units of DNA polymerase I/Klenow fragment (NEB). The reaction was incubated for 90 min at 15 °C and cDNA was purified using a QIAquick MinElute PCR Purification Kit (Qiagen). Preparation of cDNA for Illumina Genome Analyzer is described further in the Supplementary Methods.

2.3. Transcriptome assembly

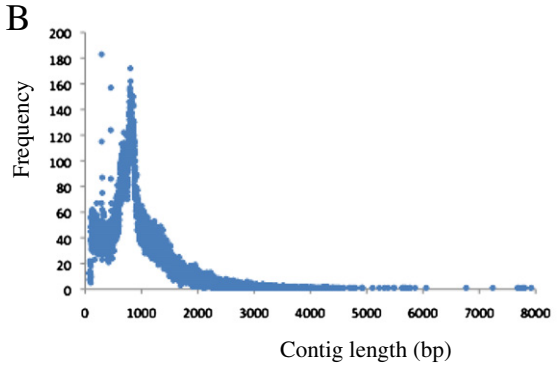
The 68,445,070 raw Illumina reads were processed by removing N's, adaptor sequences and parsed for barcode sequences. A total of 27,470,570 reads were removed and the remaining 40,974,500 high-quality reads were used for assembling the reference transcriptome (Table 1). An additional 322,920 *O. mykiss* ESTs and 90,019 transcript consensus units were obtained from the *O. mykiss* TGI database located at Dana-Farber Cancer Institute (http://compbio.dfci.harvard.edu/tgi/). The first-pass assembly was performed with Velvet and ABySS to assemble short contigs using both our Illumina reads and available ESTs (Simpson et al., 2009; Zerbino and Birney, 2008). Second-pass assembly was completed with the MIRA assembler to combine Velvet and ABySS contigs, steelhead Illumina reads, and tentative consensus transcripts into longer transcript scaffolds (Chevreux et al., 1999).

A

		Average	Median
All Contigs	86,402	899	807
100 bp and larger	86,157	902	
1 kb and larger	27,095	1509	
Largest contig	7928		

C

BLAST parameter	Total contigs with BLASTx hits	Unique BLASTx 'best-hit'	Unique GO annotation
10^{-2}	52,165	25,729	4,265
10^{-5}	49,149	24,624	4,030
10^{-10}	27,354	16,192	3,795



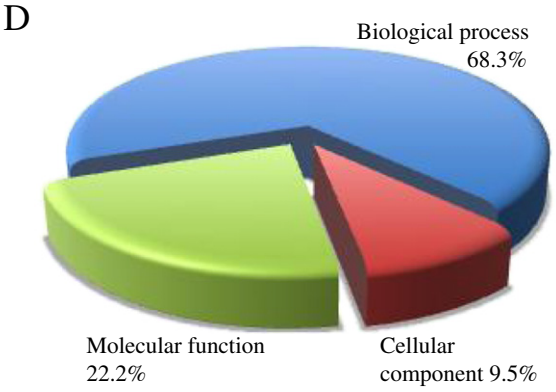


Fig. 1. Summary of steelhead trout transcriptome assembly. We assembled 86,402 transcript scaffolds with an average size of 899 bp and median of 807 bp (A). The frequency distribution of the contig lengths shows that most contigs are near the median length between 500 and 1200 bp with the longest nearly 8 kb (B). Using an *E*-value of 10^{-5}, 49,149 steelhead transcript scaffolds shared homology with a protein of the NCBI non-redundant database. 24,624 unique 'gene-hits' were identified using BLASTx of which 4030 unique GO annotations were attributed to them (C). Of the 24,624 genes with GO annotations, 68.3% belonged to biological processes while 22.2% and 9.5% were annotated as molecular function and cellular components respectively (D).

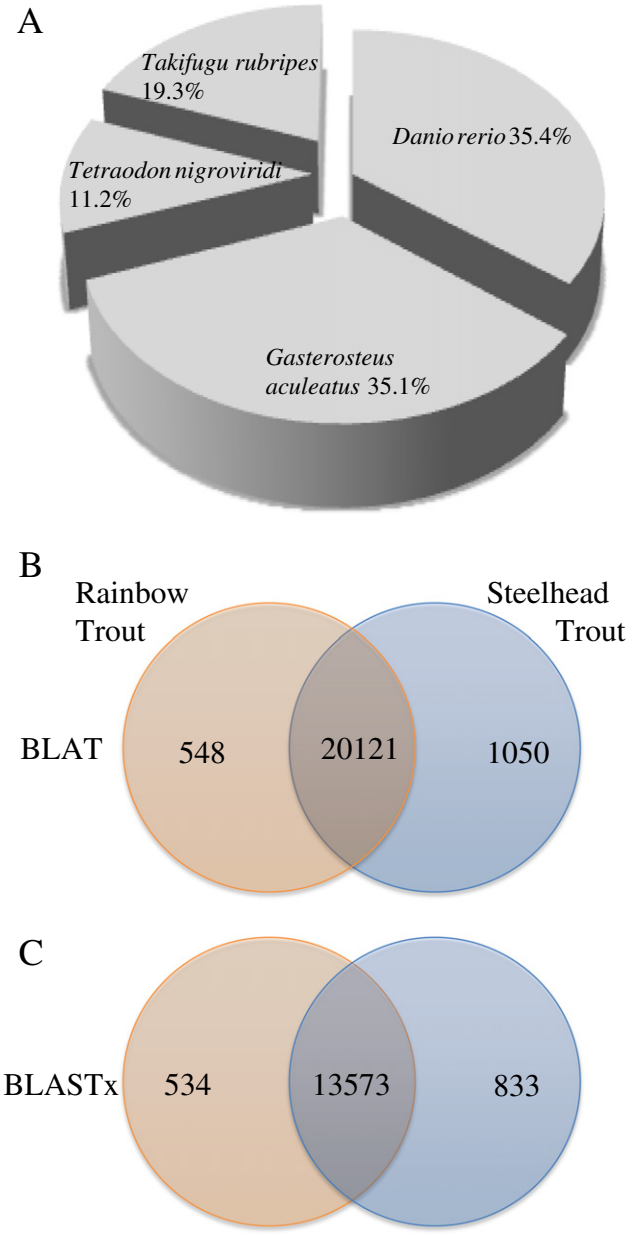


Fig. 2. Comparison of steelhead assembly to other fish species. Using BLASTx with an *E*-value cutoff of 10^{-5} we compared the assembly to protein datasets that have been structurally and functionally annotated; *Danio rerio* (zebrafish), *Gasterosteus aculeatus* (stickleback), *Takifugu rubripes* (pufferfish), and *Tetraodon nigroviridis* (pufferfish) (B). Approximately 70% of the genes with homology to one of the four fish species had a 'best match' to either zebrafish or stickleback proteins (15,874 gene hits, 35.4% and 15,249 gene hits, 35.1% respectively). Venn diagrams showing the number of sequence alignments of rainbow trout and steelhead trout sequences against Zebrafish peptide sequences using the BLAT (B) and BLASTx (C) alignment algorithms.

2.4. Functional annotation

A total of 86,157 transcript scaffolds were greater than 100 bp with an average length of 902 bp (Fig. 1; Supplementary file 1). We assembled 27,095 transcript scaffolds greater than 1 kb averaging 1509 bp. Of the 86,402 transcript scaffolds, 49,149 homologous genes were identified using *E*-value cutoff of 10^{-5}. We assigned GO terms to each sequence using the Blast2GO tools (Supplementary file 2; Gotz et al., 2008). A total of 4030 gene ontology definitions were identified among the group of 24,624 transcript scaffolds. The biological processes class was the most highly represented (68.3%), followed by molecular function (22.2%) and cellular component (9.5%; Fig. 1). See Supplementary Methods for more details regarding functional annotation.

The top BLASTx hits to other fishes included zebrafish (*Danio rerio*), Atlantic salmon (*Salmo salar*), and pufferfish (*Tetraodon nigroviridis and Takifugu rubripes*) (Fig. 2). In addition, we conducted comparisons between the rainbow trout/steelhead assemblies and zebrafish proteins (Fig. 2; See Supplementary Methods for details).

2.5. O. mykiss sequence resources

The *O. mykiss* raw sequence data from this study is accessioned in the NCBI Sequence Read Archive (SRA accession SRP009644.1). We have also made a searchable BLAST database to facilitate research using steelhead trout (http://salmon.cgrb.oregonstate.edu/). All transcriptome reference contigs, predicted proteins, and associated GO terms are available via FTP on the Salmon Resource website.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.margen.2013.12.001.

Acknowledgments

We thank the Oregon Department of Fish and Wildlife for their assistance in collecting the hatchery and wild steelhead trout, and staff of the Parkdale and Oak Springs hatchery for crossing and raising the fish, in particular, Jim Gidley and Lyle Curtis. We are grateful to Anne-Marie Girard and Caprice Rosato for the qualitative assessment of RNA and cDNA and Mark Dasenko (Center for Gene Research and Biocomputing, Oregon State University) for Illumina cluster generation and sequencing. We thank Matthew Peterson and especially Chris Sullivan (Center for Gene Research and Biocomputing, Oregon State University) for computational support.

References

Brunelli, J.P., Wertzler, K.J., Sundin, K., Thorgaard, G.H., 2008. Y-specific sequences and polymorphisms in rainbow trout and Chinook salmon. Genome 51, 739–748.
Calduch-Giner, J.A., Bermejo-Nogales, A., Benedito-Palos, L., Estensoro, I., Ballester-Lozano, G., Sitja-Bobadilla, A., Perez-Sanchez, J., 2013. Deep sequencing for *de novo* construction of a marine fish (*Sparus aurata*) transcriptome database with a large coverage of protein-coding transcripts. BMC Genomics 14, 178–189.
Chevreux, B., Wetter, T., Suhai, S., 1999. Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics, 99, pp. 45–56.
Fox, S., Filichkin, S., Mockler, T.C., 2009. Applications of ultra-high-throughput sequencing. Methods Mol. Biol. 553, 79–108.
Fraser, B.A., Weadick, C.J., Janowitzm, I., Roddm, F.H., Hughesm, K.A., 2011. Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome. BMC Genomics 12, 202–216.
Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J., Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36, 3420–3435.
Hale, M.C., Thrower, F.P., Berntson, E.A., Miller, M.R., Nichols, K., 2013. Evaluating adaptive divergence between migratory and non-migratory ecotypes of a salmonid fish, *Oncorhynchus mykiss*. G3 (Bethesda) 2 (9), 1113–1127.
Hayes, S.A., Hanson, C.V., Pearse, D.E., Bond, M.H., Garza, J.C., MacFarlane, B., 2012. Should I stay or should I go? The influence of genetic origin on emigration behavior and physiology of resident and anadromous juvenile *Oncorhynchus mykiss*. N. Am. J. Fish Manag. 32 (4), 772–780.
Johansen, S.D., Karlsen, B.O., Furmanek, T., Andreassen, M., Jorgensen, T.E., Bizuayehu, T.T., Breines, R., Emblem, A., Kettunen, P., Luukko, K., et al., 2011. RNA deep sequencing of the Atlantic cod transcriptome. Comp. Biochem. Physiol. Part D Genomics Proteomics 6, 18–22.
Long, Y., Li, Q., Zhou, B., Song, G., Li, T., Cui, Z., 2013. *De novo* assembly of mud loach (*Misgurnus anguillicaudatus*) skin transcriptome to identify putative genes involved in immunity and epidermal mucs secretion. PLoS ONE 8, e56998.
Salem, M., Rexroad, C.E., Wang, J., Thorgaard, G.H., Yao, J., 2010. Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches. BMC Genomics 11, 564–574.
Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.
Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.