Fisheries and Aquaculture Journal

Zadlock, et al., Fish Aqua J 2017, 8:1
DOI: 10.4172/2150-3508.1000186

Research Article OMICS International

# De Novo Assembly and Analysis of the Testes Transcriptome from the Menhaden, *Bervoortia tyrannus*

**Frank J Zadlock IV[1*], Satshil B Rana[1], Zain A Alvi[1], Ziping Zhang[2], Wyatt Murphy[1] and Carolyn S Bentivegna[3]**

[1]*Department of Biological Science, Seton Hall University, South Orange, New Jersey, USA*

[2]*College of Animal Science, Fujian Agriculture and Forestry University, Fuzhou, China*

[3]*Department of Chemistry and Biochemistry, Seton Hall University, South Orange, New Jersey, USA*

[*]**Corresponding author:** Frank J Zadlock IV, Department of Biological Science, Seton Hall University, South Orange, New Jersey, USA, Tel: (973) 313-6146; E-mail: Frank.Zadlock@student.shu.edu

**Received date:** November 13, 2016; **Accepted date:** December 05, 2016; **Published date:** December 08, 2016

**Copyright:** © 2016 Zadlock FJ, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

## Abstract

**Background:** The menhaden, *Bervoortia tyrannus*, is one of the most important fish within the oceanic ecosystem and a crucial species supporting major fisheries along the Atlantic and Gulf coasts. However, little is known about menhaden from a genetic aspect. The objective of this project is to apply high throughput sequencing to the testes of menhaden to provide the genetic tools required to further study their population dynamics

**Result:** We applied Illumina Next Seq 500 technology to two different testes and used Velvet/Oases to perform the de novo assembly that resulted in the construction of 254,462 contigs. Applying BLASTX to annotate the contigs against the non-redundant protein database resulted in 46.89% matches. To validate the accuracy of the assembly, the reads were mapped back to transcripts (RMBT) with a percentage of 87.83%. To experimentally verify the assembly results, primers were designed based on the assembled transcriptome, and PCR products were verified by Sanger sequencing. To enhance the functional categorization of the annotated contigs, they were further classified using Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Clusters of Orthologous Groups (COG) databases.

**Conclusion:** This research is the first report of an annotated overview of the testes transcriptomes in *B. tyrannus*, resulting in the most comprehensive genetic resource available for menhaden to date. This work can provide a repository for future gene expression analysis, functional studies, and reproductive investigations in *B. tyrannus*. This will enhance the capabilities of population monitoring and can be used as a benchmark in comparative studies in other fish models. Overall, this research will open new opportunities and bring new insights for researchers studying *B. tyrannus*.

**Keywords:** Menhaden fish; De novo assembly; Assembly validation; Transcriptome analysis

## Introduction

Menhaden (Family Clupeidae, Genus Brevoortia) are high fecundity, filter-feeding marine teleost fish that are considered to be one of the most economically and ecologically important species in North America [1]. In oceans and estuaries, they contribute to ecosystem health by clearing the water of excess algal biomass and detritus [2,3]. They are also the main food source for a wide variety of predatory invertebrates (jellyfish, squids, etc.) fish, (striped bass, bluefish, etc.,) birds, (osprey and brown pelican), and marine mammals [4].

From an economic standpoint, there are two established commercial fisheries for menhaden. The first is the reduction fishery that turns the menhaden into fish oil omega-3 supplements for example, and into fishmeal for livestock and aquaculture consumption [4,5]. The second is the bait fishery that uses the menhaden as bait for bluefish, crab, and lobster [6-15]. They have been rarely studied at the genetic level most likely due to the lack of genomic and transcriptome data.

In this study, we applied Illumina Next Seq 500 technology to two different menhaden testes and performed de novo assemblies on the generated raw reads using Velvet/Oases. BLASTX was utilized to annotate the assembled contigs against the NCBI non-redundant protein database. To validate the accuracy of the assemblies in silico, the contigs were mapped back to transcripts (RMBT) [16]. To experimentally verify the assembly results, primers were designed based on the assembled transcriptome and PCR products were verified by Sanger sequencing. To enhance the functional categorization of the annotated contigs, they were further classified using Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Clusters of Orthologous Groups (COG) databases. This research also identified microsatellites, and various repetitive DNA elements between within the transcriptome.

To date, this research is the first report of an annotated overview for the testes transcriptome in *B. tyrannus*, resulting in the most comprehensive genetic resource available for the species. This work can provide a repository for future gene expression analysis, functional studies, and reproductive investigations in *B. tyrannus*. This will enhance the capabilities of population monitoring and can be used as a benchmark in comparative studies in other fish models. Overall, this

research will open new opportunities and bring new insights for researchers studying *B. tyrannus*.

## Materials and Methods

### Fish collection and nucleic acid isolation

Male fish identified as *Brevoortia tyrannus* were collected off the coast of New Jersey in November 2013 by using a trolling net. The authorities who issued the permission for the capturing of the menhaden were NOAA, National Marine Fisheries Service, Northeast Regional Office (Permit #410087) and the State of New Jersey, Department of Environmental Protection (Permit #1333). In all instances the fish were alive when captured, and capture methods followed approved animal handling protocols reviewed by the authorities who issued the field permits. The vertebrate work was approved by Virginia Institutes of Marine Science's Institutional Animal Care and Use Committee (IACUC-2011-02-04-7125-jxgart) and NEAMAP Inshore Trawl Survey. The fish were sacrificed using spinal cord dislocation and the gonads were dissected from two separate male menhaden. Tissue samples were stored in RNAlater (Qiagen) at -20°C prior to RNA extraction.

Total RNA was extracted from each testis using TRIzol® Kit (Invitrogen™) following the manufacturer's instructions. RNA samples were then digested by DNase I to remove potential genomic DNA. A BioAnalyzer 2100 (Agilent Technologies) was used to validate that the RNA integrity number (RIN) for each sample had a value >7.0.

### Illumina short-read library construction and sequencing

Waksman Genomics Facility at Rutgers University conducted all steps in transcriptome library preparation. Ambion MicroPoly A purist Kit was used to remove ribosomal RNA (rRNA) and recover high quality mRNA from the menhaden samples. Agilent Genomics BioAnalyzer mRNA Nano Kit was used to quantify the mRNA and confirm a successful rRNA depletion (<2% rRNA attained). The dUTP strand specific cDNA library preparation strategy was employed using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina. Illumina Tru-Seq adapters were used to barcode the libraries and were amplified for 12-15 cycles of PCR. Illumina NextSeq 500 High Output with 300 cycle kit was used to sequence the 155×155 bp pair end libraries. Agilent Technologies and ThermoFisher Scientific's Qubit DNA HS Assay Kit was used to quantify the completed RNA-seq libraries.

### Sequence data processing

The quality assessment and adapter trimming was performed using FastQC v0.10.1 and Trimmomatic v0.32 on the raw Illumina sequence reads [17,18]. All low quality reads with a Phred score value below 20 were removed. After trimming, FastQC analysis was performed again to verify the quality of the remaining raw sequence data. The data sets were further cleaned from contaminating sequences using Deconseq with the parameters set to 90% of the contig length with an identity of 94% [19].

### De novo transcriptome assembly and annotation

The high quality filtered reads were assembled using Velvet (v1.2.07), which assembles short reads using the de Bruijn algorithm [20] along with Oases (v0.2.08), which operates on the output of Velvet [21]. Firstly, to create the input files for Velvet, the paired-end fastq files were interleaved. Then a multiple k-mer assembly strategy was applied to assemble k-mer sizes 35, 41, 51, 61, 71, 81, and 91 [22]. The seven k-mer assemblies were merged with Oases followed by CD-HIT-EST (v 4.6.1) to further remove the redundancy and cluster the contigs for annotation [23]. Lastly, the two separate transcriptomes were merged together with CD-HIT-EST.

Annotation of the contigs was accomplished using local BLASTX (v2.2.29+). Homologous sequences were searched for against the NCBI non-redundant (Nr) protein database using an E-value <1e-5 [24]. Gene annotations were assigned based on the top BLASTX hit. Functional annotation was performed by merging the results from Blast2GO (v 2.7.2) with the results from InterProScan to expand the number of annotated sequences [25-27]. These annotations were used to assign putative functionalities, level-two GO terms, and KEGG (Kyoto Encyclopedia of Genes and Genomes) based metabolic pathways via BLAST2GO [28]. To further elucidate possible functions, the contig sequences were aligned to the Clusters of Orthologous Groups (COG) database using BLASTX with an E-value <1e-5 [29].

### Assembly and annotation assessment

In order to demonstrate that the de novo assembly was performed properly, an in silico method was used to match the assembled contigs to protein sequences of related species. To accomplish this, we utilized PhyloT (http://phylot.biobyte.de/contact.html) to generate a phylogenetic tree between menhaden (B. tyrannus) and the eleven publically available fish genomes (Poecilia formosa, Amazon molly; Astyanax mexicanus, Mexican tetra; Gadus morhua, Atlantic cod; Takifugu rubripes, Japanese pufferfish; Oryzias latipes, medaka; Xiphophorus maculatus, southern platyfish; Lepisosteus oculatus, spotted gar; Gasterosteus aculeatus, stickleback; Tetraodon nigroviridis, green spotted pufferfish; Oreochromis niloticus, Nile tilapia; and Danio rerio, zebrafish) on Ensembl. This program creates trees based on the NCBI taxonomy database, and it was visualized by the web based tool, Interactive Tree of Life (v2) [30]. Based on the analysis, zebrafish (D. rerio) and Mexican tetra (A. mexicanus) were shown to be the closest related. All assembled contigs were then compared to the Ensembl proteins of zebrafish and Mexican tetra using BLASTX with an E-value cut-off of 1e-5. Secondly, to determine the accuracy of the assembly, the percentage of raw reads that could be mapped back to transcripts (RMBT) was determined [16]. To accomplish this, indexes were generated using bowtie2-build followed by Bowtie2 (v 2.2.5) to map the reads against the assembly [31].

To experimentally verify the assembly results, primers were designed based on the assembled transcriptome and PCR was performed on 8 genes (CYP17a1, 3-β-HSD, β-actin, GAPDH, HIF-1a, StaR, ARNT, and EGR). From the isolated RNA, a cDNA library was constructed using oligo-dT primers (Applied Biosystems, Foster City, CA). The cDNA was used as a template to amplify the genes of interest. The PCR amplification was performed at 94°C for 30 sec, 60°C for 30 sec, and 72°C for 30 sec using the designed primers pairs for each gene. These primers were designed using Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) and are shown in Table 1. The presence of a unique PCR product of appropriate size was verified by agarose gel electrophoresis and the gene of interest was confirmed using commercially available Sanger sequencing (Genewiz Inc. Plainfield, NJ, USA) (data not shown).

## Repetitive element investigation and microsatellite identification

Assembled sequences were scanned with RepeatMasker (v 4.0.5) to identify all repetitive elements using zebrafish as a reference [32]. Microsatellite motifs were identified using the program Msatfinder (v 2.0.9) [33]. The repeat thresholds for di-, tri-, tetra-, penta-, hexa-nucleotide motifs were set as 8, 5, 5, 5 and 5. The mononucleotide repeats were removed by modifying the perl script. For future PCR validation, 80 microsatellite sequences that met the selection criteria of having flanking sequences longer than 50 bp on both sides have been designed as seen in Table 2.

## Results

To globally profile the two testes transcriptomes of menhaden, we employed Illumina NextSeq 500 technology to sequence the libraries generating 10,765,249 pair-end short reads encoding 1,560,309,410 bases for Menhaden 1 and 17,910,601 pair-end short reads encoding 2,491,172,965 bases for Menhaden 2 (Table 1). All the raw sequencing reads were deposited into the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI), and can be accessed under the accession numbers SRX892008 and SRX994942, respectively.

| | Menhaden 1 | Menhaden 2 |
|---|---|---|
| Number of nucleotide bases | 1,56,03,09,410 | 2,49,11,72,965 |
| Number of raw reads | 1,07,65,249 | 1,79,10,601 |
| Number of clean reads for assembly | 89,13,332 | 1,29,57,697 |
| Percent of used reads | 83% | 72% |

**Table 1:** Statistics of the raw reads after Illumina sequencing and processing.

To improve the accuracy of the assembly, the raw sequence reads were cleaned to remove Illumina adaptor sequences, low quality reads with a Phred score value less than 20, and contaminating sequences.

| | Menhaden Testes |
|---|---|
| Contig Number | 2,54,462 |
| N50 Length | 1,324 bp |
| Minimum Contig Length | 200 bp |
| Largest Contig Length | 30,617 bp |
| Average contig length | 831.68 bp |
| GC (%) | 46 |
| RMBT % | 87.83 |
| Nr Database Match | 46.89% |
| Zebrafish Genome Match | 42.56% |
| Mexican tetra Genome Match | 41.34% |

**Table 2:** Statistics of the Velvet/Oases Assembly and Annotation.

The filtering of the raw sequence reads resulted in 8,913,332 (83%) clean reads for Menhaden 1 and 12,957,697 (72%) clean reads for Menhaden 2 (Table 1). Velvet and Oases, common assembly programs successfully used in previous Illumina based de novo transcriptome studies, were employed to perform the assembly [7,10,14,21,22,34,35]. The paired-end sequence reads were assembled into 254,446 contigs with an N50 length of 1,324 bp and average contig length of 831.68 bp (Table 2). The contig length distribution ranged from 200 bp to more than 3,000 bps as shown in Figure 1.

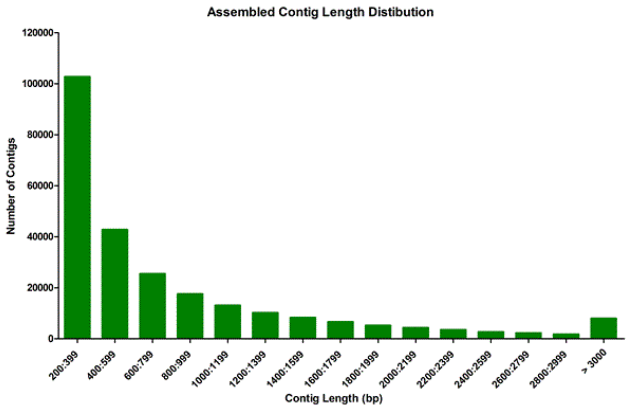


**Figure 1:** Contig length distribution results. The distribution of the contig lengths after the assembly.

## Assessment of transcriptome assembly

It is crucial to assess whether or not a de novo assemblies is reliable.

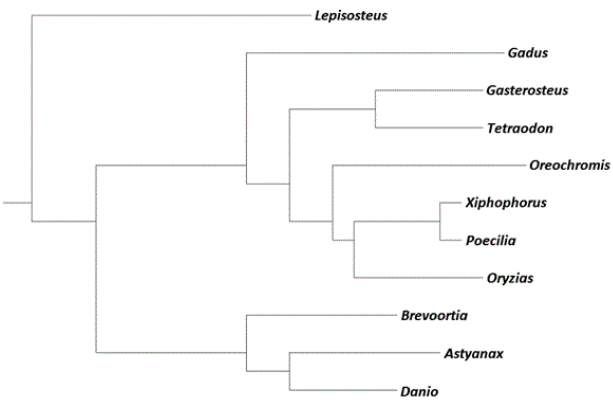


**Figure 2:** Phylogenetic tree analysis. A phylogenetic tree analysis of the 11 publically available fish genomes on Ensembl reveals that *D. rerio* (zebrafish) and *A. mexicanus* (Mexican tetra) are the closest relatives to *B. tyrannus*.

One of the options to gauge whether or not the de novo assembly was properly performed is to align the assembled contigs to a phylogenetically related fish genome when a reference genome, cDNA, or EST sequences are not available. Based on a phylogenetic tree comparison between menhaden and the eleven publically available fish

genomes on Ensembl, it was determined that zebrafish along with Mexican tetra were the closest related genera to menhaden (Figure 2). Therefore, to validate the accuracy of the Velvet/Oases assemblies, the contigs from each transcriptome were aligned to the zebrafish and Mexican tetra genomes independently using BLASTX with an E-value of <1e-5. The menhaden had 42.56% matches to the Zebrafish and 41.34% to the Mexican tetra database (Table 2). The RMBT percentage which infers the accuracy of the assembly was 87.83% (Table 2). A total of eight gene-specific primers were verified by the presence of a unique PCR product followed by Sanger sequencing to further legitimize the accuracy and annotation of the assembly (Figure 3 and 4).

## BLAST analyses

To annotate the menhaden testis transcriptomes.

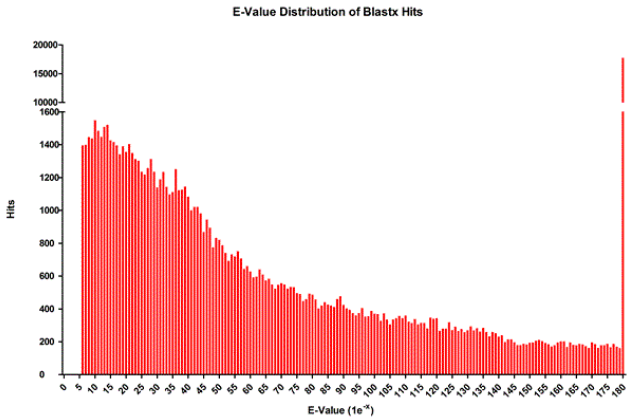


**Figure 3:** The E-value distribution of the BLASTX hits. The contig distribution based on BLASTX E-Values.

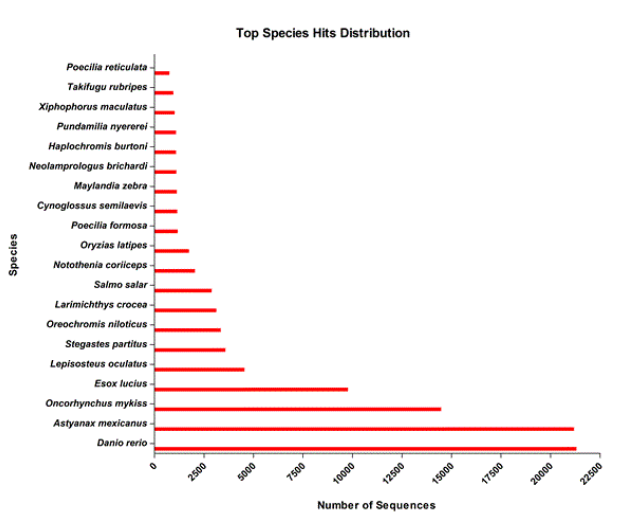


**Figure 4:** The top hit species distribution. The top species distribution based on the BLASTX results. The top 21 species shown all belong to various fish species.

The contigs were first searched against the NCBI non-redundant (Nr) protein database by using BLASTX with an E-value <1e-5 resulting in 46.89% matches as seen in Table 2. An E-value of <1e-5 was chosen because there is no available annotated menhaden genome nor a close relative that can provide a reference genome. Although we used an E-value threshold of <1e-5, the majority of the matches were below this threshold as graphically depicted in Figure 3. The species distribution of the top BLASTX hits against the Nr database for the transcriptome revealed that the top hits were contributed from other fish species as shown in Figure 3. The zebrafish and Mexican tetra provided the most matches with 21,598 (8.41%) and 21,541 (8.38%).

## Go ontology assignment and functional classification

Functional annotation is imperative to assign biological information to the transcriptomic data of non-model organisms. Therefore, Gene Ontology (GO) terms were subsequently assigned to the menhaden contigs based on their sequence matches to known protein sequences in the Nr database. By merging the Blast2GO annotations with the InterProScan results, 70,919 contigs out of 254,446 (27.87%) were assigned at least one GO term (Table 3). The majority of these GO assignments belonged to molecular function (101,730, 40.05%) followed by biological processes (100,885, 39.71%) and cellular component (51,378, 20.22%) (Table 2). The top level 2 GO terms for the molecular function category were binding, catalytic activity, transporter activity, signal transducer activity, molecular transducer activity, and molecular function regulator. For the biological process category, the top level 2 GO terms were cellular process, metabolic process, single-organism process, biological regulation, regulation of biological process, and response to stimulus. As for the cellular component category, cell, cell part, organelle, membrane, and macromolecular complex were the top level 2 GO terms (Figure 5).

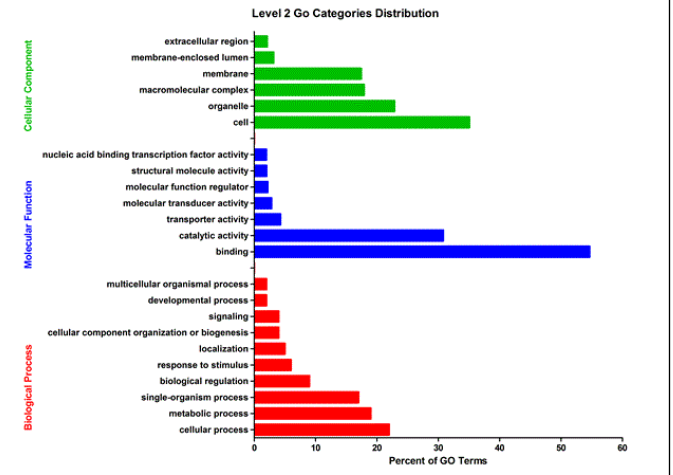


**Figure 5:** Functional classification based on the three categories of Gene Ontology. The GO-terms were categorized at level 2 into three major functional categories: biological process, cellular component, and molecular function. A detailed summary is listed in supplementary table.

In order to further resolve the functionality of the menhaden testis transcriptomes, the annotated contigs were categorized into different functional groups based on the Cluster of Orthologus Groups (COGs)

database. This database contains proteins generated by comparing the protein sequences of complete genomes from bacteria, algae, and eukaryotes where each cluster has proteins of paralogs from at least three lineages [36]. Of the 119,326 contigs that had BLASTX matches, 30,829 (25.83%) could be classified into one of the 26 COG categories (Figure 6). Among these categories, the majority of clusters were "General function prediction only" (26,264, 85.19%), "Signal transduction mechanisms" (17,150, 55.62%), "Transcription" (4,663, 15.12%), "Posttranslational modification, protein turnover, chaperones" (4,537, 14.71%), and "Translation, ribosomal structure and biogenesis" (4,219, 13.68%) respectively.

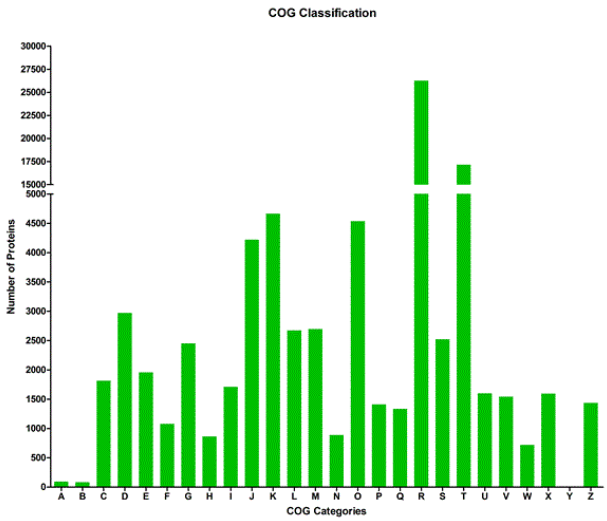


**Figure 6:** Histogram of the COG classification (A) RNA processing and modification; (B) Chromatin structure and dynamics; (C) Energy production and conversion; (D) Cell cycle control, cell division, chromosome partitioning; (E) Amino acid transport and metabolism; (F) Nucleotide transport and metabolism; (G) Carbohydrate transport and metabolism; (H) Coenzyme transport and metabolism; (I) Lipid transport and metabolism; (J) Translation, ribosomal structure and biogenesis; (K) Transcription; (L) Replication, recombination and repair; (M) Cell wall/membrane/envelope biogenesis; (N) Cell motility; (O) Posttranslational modification, protein turnover, chaperones; (P) Inorganic ion transport and metabolism; (Q) Secondary metabolites biosynthesis, transport and catabolism; (R) General function prediction only; (S) Function unknown; (T) Signal transduction mechanisms; (U) Intracellular trafficking, secretion, and vesicular transport; (V) Defense mechanisms; (Y) Nuclear structure; (Z) Cytoskeleton.

To further enhance the functional categorization of the testis in menhaden, we utilized the contigs with BLASTX matches to search through the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to discover the active metabolic pathways at the time of collection. Based on the results, 967 enzymes were mapped to 125 different metabolic pathways (Table 3). The top five pathways with KEGG annotations were purine metabolism (56, 17.2 5.79%), Arginine and proline metabolism (30, 3.1 2.99%), Amino sugar and nucleotide sugar metabolism (29, 2.99%), Pyrimidine metabolism (27, 2.79%), and Cysteine and methionine metabolism (25, 2.58%) as seen in Figure 7.

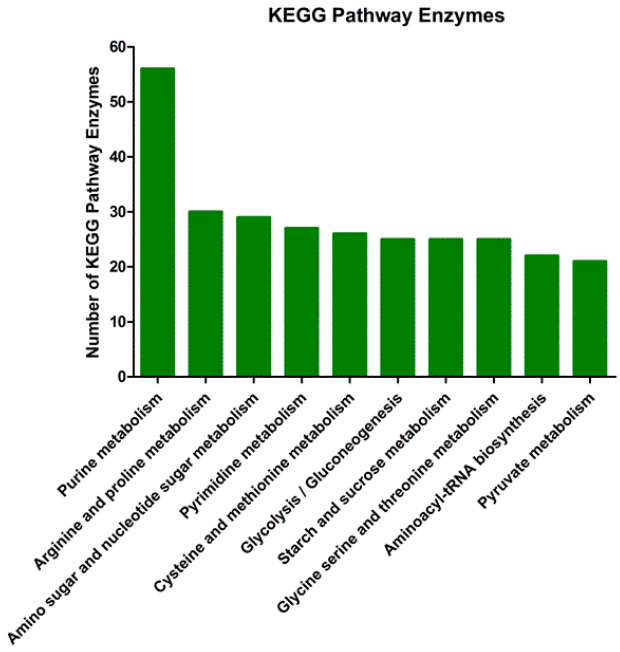


**Figure 7:** Summary of *B. tyrannus* KEGG pathways. Distribution of the top ten KEGG pathway enzymes related to *B. tyrannus*. A detailed summary is listed in Table 3.

## Repetitive Element Analysis and Microsatellite Identification

The extent of the repetitive elements in the menhaden's testes transcriptome was assessed by using Repeatmasker with the Zebrafish Repeat Database as a reference.

| **Menhaden Testes** | |
|---|---|
| Total number of sequences surveyed | 2,54,462 |
| Number of sequences containing repeats | 34,486 |
| Total number of bp searched | 21,16,32,176 |
| Total number of microsatellites found | 46,864 |
| Di-nucleotide repeats | 28,566 |
| Tri-nucleotide repeats | 13,373 |
| Tetra-nucleotide repeats | 4,079 |
| Penta-nucleotide repeats | 708 |
| Hexa-nucleotide repeats | 139 |

**Table 3:** Statistics of the microsatellite distribution within the transcriptome.

Searching through the 211,632,176 bps within the 254,446 contigs resulted in the detection of 11,017,712 bps (5.21%) of repeated sequences. The distribution and classification of the identified repetitive elements are shown in Table 3. The most abundant types of repetitive elements in the sequences were simple repeats (2.18%), retroelements (1.19 %), DNA transposons (0.93%), and small RNA (0.30%).

To determine the presence of microsatellites, the assembled contigs were scanned with the selection criteria of having sufficient flanking sequences of at least 50 bps for primer designing (Table 2). A total of 46,865 microsatellites within 34,489 contigs consisted of either di-,tri-, penta- or hexa-nucleotide repeats (Table 3). Eighty primer pairs have been designed as seen in Table 2).

## Discussion

Menhaden are an important species for the commercial fishing industries. They are also an ecologically important species within oceanic and estuarine ecosystems by providing a crucial link between phytoplankton and marine carnivores [1]. Despite the importance of this fish, detailed genetic information for it is currently lacking. Transcriptome analyses are a cost-effective method for the characterization of species that lack a fully sequenced genome [7,8]. The short reads from Illumina paired-end sequencing can provide a deep sequencing coverage for the accurate base calling in de-novo transcriptome assembly and has been utilized to characterize many non-model species [37-40]. However, no transcriptome sequencing data is available for menhaden.

In a recent study comparing 454 GS-FLX (Roche Diagnostics Corporation) to Illumina (Illumina, Inc.) for the utility of RNA-seq in a non-model bird, Illumina assemblies performed best for de novo transcriptome characterization in terms of contig length, transcriptome coverage, and complete assembly of gene transcripts [41]. Therefore, Illumina HiSeq 500 was utilized in this de novo study of this non-model fish. For the sequencing strategy, we elected to perform the paired-end sequencing approach because it facilitates the assembly process compared to the common one way sequencing approach. This is because the paired-end approach has the ability to produce longer contigs by filling gaps in the consensus sequence and longer contigs are better for mapping.

Due to the absence of a suitable reference genome for menhaden, the de novo assembly approach was performed. Velvet and Oases, common assembly programs in de novo transcriptomic studies using Illumina reads, were employed to assemble the processed reads [7,10,14,21,22,34,35,42,43]. It should be noted that the quality of the de novo transcriptome assembly is reliant on the user-defined k-mer value characterized as the length parameter defining the sequence overlap between two reads forming a contiguous sequence (referred as a contig) [44,45]. Low k-mer values will recover less abundant transcripts by producing a large number of contigs, but this will lead to the assembly of numerous and highly fragmented transcripts due to sequencing errors and lack of overlap [44,45]. Meanwhile, high k-mer values will theoretically result in a more contiguous assembly of high coverage transcripts and splice variants. However, this approach will produce fewer contigs and cause lower transcript representation due to capturing only highly represented reads [44,45]. Interestingly, the single k-mer approach is still a popular choice for the de novo assembly of transcriptomes even though this might result in the loss of relevant biological information due to the inadequate representation of various k-mer lengths [45]. Therefore, we employed a multiple k-mer approach in this study because clustering multiple single k-mer assemblies takes advantage of the characteristics of both the low and high k-mer lengths, resulting in a better transcript diversity of the assembly.

It is imperative to assess whether or not de novo assemblies are reliable. This can be investigated with either in silico or in vitro approaches. One in silico approach is to match the contigs to known cDNA or EST sequences to confirm the accuracy of the de novo assembly [8,46]. However, due to the scarcity of cDNA or EST sequences for menhaden, another alignment technique was performed for the validation. In this method, the contigs for the species of interest was aligned to its closest phylogenetic relative with a reference genome utilizing BLASTX [9]. Based on a phylogenetic tree analysis with the eleven publically available fish genomes on Ensembl, it was determined that zebrafish along with Mexican tetra were the closest related genera to menhaden. The BLASTX alignments for the menhaden against the zebrafish (42.56%) and Mexican tetra (41.34%) yielded analogous results with other non-model fish de novo transcriptome assembly studies signifying that the assemblies are reliable [9,12,13,47,48]. The low number of hits for both zebrafish and Mexican tetra could be due the presence of menhaden-specific contigs, as menhaden belong to separate clades in the phylogenetic tree (Figure 2). It also should be taken into consideration that the menhaden testis transcriptomes were compared to the entire genome of each phylogenetically related species instead their testis transcriptomes, which are not published. Additionally, some of the unannotated contigs may be short and therefore 1) not consist of well characterized protein domains, 2) predominately be 3' or 5' untranslated regions, or 3) be non-coding RNAs [49]. Another in silico approach to validate the accuracy of an assembly is the percentage of BLASTX matches against the Nr database. In this study, the percentage of matches for the menhaden was 46.23%. This percentage is comparable to other non-model fish de novo assemblies, further establishing the credibility of the assemblies [9,12,13,47,50]. It should be noted that the number of matches is affected by the amount of available sequence data, and there is a limited number of publically available fish genomes to contribute to the overall success of the annotation [49]. To validate that the BLASTX matches were relevant to the non-model organism of interest, the species distribution of the top BLASTX hits against the Nr database was performed using Blast2GO. Results showed that the top twenty one contributing species for the annotation were from other fish species, further strengthening the validity of the assembly. The top two contributors, zebrafish and Mexican tetra, correlated with the phylogenetic tree results that classified the reference genomes independently based on NCBI classification. To further legitimize the accuracy of the assembly along with its annotation, an in vitro approach of validating primers designed based off the assembly was performed. A total of eight gene specific primers were validated with Sanger sequencing (Table 1). All together, these successful in silico and in vitro validation steps established the reliability of the Velvet/Oases assemblies in both transcriptomes.

The GO project is an international collaborative effort to standardize and use ontologies to support biologically meaningful annotation of genes and their products in any organism [51]. GO ontologies provide the standardized description of attributes of genes along with their products in three key biological domains that are shared by all organisms: molecular function, biological process and cellular component [51]. By merging the Blast2GO annotations with the InterProScan results, 70,919 contigs out of 254,446 (27.87%) were

assigned at least one GO term (Table 3). These results are comparable to other annotation efforts in non-model fish studies that utilized the de novo assembly approach [9,13,47,50]. Again, it should be taken into consideration that the limited number of publically available fish genomes inhibits the overall successful rate of annotating non-model fish [49].

The GO analysis in this research was similar to the testes transcriptome analysis for the channel catfish by Sun et al., [52] and the yellow catfish by Lu et al., [53]. Four of the top five GO terms in the biological process category for the channel catfish (cellular process, metabolic process, biological regulation, and response to stimulus) and the top four GO terms for the yellow catfish (cellular process, single-organism process, metabolic process, and biological regulation) matched the top five GO terms in this study [52,53]. Within all three testis studies, the cellular process, metabolic process, and biological regulation GO terms were observed as being with in the top four contributors for the biological category. For the molecular function category, four of the top five GO terms for the channel catfish (binding, catalytic activity, molecular transducer activity, and transporter activity) and four of the top five GO terms for the yellow catfish (binding, catalytic activity, transporter activity, and molecular transducer activity) matched the top four GO terms in this study [52,53]. The binding, catalytic activity, transporter activity, and molecular transducer GO terms were represented within the top five GO terms of the molecular function category for all three testis transcriptome studies. For the cellular componentcategory, four of the top five GO terms for the channel catfish (cell, organelle, macromolecular complex, and extracellular region) and four of the top five GO terms for the yellow catfish (cell, organelle, membrane, and macromolecular complex) matched the top five GO terms in this study [52,53]. The cell, organelle, and macromolecular complex GO terms were observed within the top five GO terms in all three testis studies. Overall, the GO categorization results show similarities between three different fish species studies interested in profiling their testis transcriptomes.

The menhaden were caught during their reproductive season in November [4]. During this time of year, the sexually activated testes are in the process of growing and becoming enlarged [54]. Typically, environmental cues and metabolic signals control the reproductive state of fish by inducing the hypothalamus to release gonadotropin-releasing hormone (GnRH) [55]. This results in stimulating the pituitary gland to secrete gonadotrophic hormones (GTH) along with growth hormones (GH) [56-58]. The GH controls several complex processes including growth and metabolism of proteins, fats, along with carbohydrates [59]. These activities lead to an increase in cellular uptake of amino acids, lipolysis, and blood sugar that promote the exponential growth of the testes. The sexual activation also requires the biosynthesis of purines and pyrimidines for DNA synthesis that is essential for the growth of the testes along with spermatogenesis [60]. Therefore, it is not surprising that the top Go Ontology assignments, functional characterizations, and metabolic analysis all share a similar trend of growth promotion within the testes. This is represented by the general themes of metabolic processes, nucleotide biosynthesis, replication, translation, and energy results from the GO, COG and KEGG analysis.

## Conclusion

This research is the first to assign functional annotations to the testes transcriptome in menhaden, resulting in the most comprehensive biological information available for the species to date. This transcriptomic data can provide the ground work for studying the menhaden's population dynamics, biomonitoring, and reproductive health. Additionally, this data can be utilized in comparative studies in other fish models to further enhance breeding programs and evolutionary studies. Overall, this research will open new opportunities to use menhaden as a model organism for the oceanic and estuary ecosystems instead of relying on traditional fish model organisms that are not indigenous to the same environment.

## Acknowledgement

This project was sponsored by Funding provided by the Louisiana Department of Wildlife and Fisheries- PI, Ralph Portier (LSU), Co-PI, Carolyn S. Bentivegna.

## References

1. Franklin HB (2007) The Most Important Fish in the Sea: Menhaden and America. Washington: Island Press/Shearwater Books.
2. Deegan LA, Peterson BJ, Portier R (1990) Stable isotopes and cellulase activity as evidence for detritus as a food source for juvenile gulf menhaden. Estuaries 13: 14-19.
3. Durbin AG, Durbin EG (1998) Effects of menhaden predation of plankton populations in Narragansett Bay, Rhode Island. Estuaries 21: 449-465.
4. Ahrenholz DW (1991) Population biology and life history of the North American menhadens, Brevoortia spp. Marine Fisheries Review 53: 3-19.
5. Kristofersson D, Anderson JL (2006) Is there a relationship between fisheries and farming? Interdependence of fisheries, animal production and aquaculture. Marine Policy 30: 721-725.
6. NOAA, Chesapeake Bay Office.
7. Garg R, Patel RK, Tyagi AK, Jain M (2011) De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification. DNA Research 18: 53-63.
8. Fan H, Xiao Y, Yang Y, Xia W, Mason AS, et al. (2013) RNA-Seq Analysis of Cocos nucifera: Transcriptome Sequencing and Subsequent Functional Genomics Approaches. PLoS ONE 8: e59997.
9. Coppe A, Agostini C, Marino IAM, Zane L, Bargelloni L, et al. (2013) Genome Evolution in the Cold: Antarctic Icefish Muscle Transcriptome Reveals Selective Duplications Increasing Mitochondrial Function. Genome Biol Evol 5: 45-60.
10. Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M (2011) Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics 12: 317.
11. Finseth FR, Harrison RG (2014) A Comparison of Next Generation Sequencing Technologies for Transcriptome Assembly and Utility for RNA-Seq in a Non-Model Bird. PLoS ONE 9: e108550.
12. Ji P, Liu G, Xu J, Wang X, Li J, et al. (2012) Characterization of Common Carp Transcriptome: Sequencing, De Novo Assembly, Annotation and Comparative Genomics. PLoS ONE 7: e35152.
13. Huth TJ, Place SP (2013) De novo assembly and characterization of tissue specific transcriptomes in the emerald notothen, Trematomus bernacchii. BMC Genomics 14: 805.
14. Nandety RS, Kamita SG, Hammock BD, Falk BW (2013) Sequencing and De Novo Assembly of the Transcriptome of the Glassy-Winged Sharpshooter(Homalodisca vitripennis). PLoS ONE 8: e81681.
15. Van Belleghem SM, Roelofs D, Van Houdt J, Hendrickx F (2012) De novo Transcriptome Assembly and SNP Discovery in the Wing Polymorphic Salt Marsh Beetle Pogonus chalceus (Coleoptera, Carabidae). PLoS ONE 7: e42605.

16. Rana SB, Zadlock FJIV, Zhang Z, Murphy WR, Bentivegna CS (2016) Comparison of De Novo Transcriptome Assemblers and k-mer Strategies Using the Killifish, Fundulus heteroclitus. PLoS ONE 11: e0153104.
17. Andrews S (2010) FastQC: a quality control tool for high throughput sequence data.
18. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics 30: 2114-2120.
19. Schmieder R, Edwards R (2011) Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. PLoS ONE 6: e17288.
20. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18 : 821-829.
21. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 10861092.
22. Haznedaroglu BZ, Reeves D, Rismani-Yazdi H, Peccia J (2012) Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. BMC Bioinformatics 13: 170.
23. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.
24. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
25. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.
26. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36: 3420-35.
27. Quevillon ESV, Pillai S, Harte N, Mulder N, Apweiler R, et al. (2005) InterProScan: protein domains identifier. Nucl. Acids Res 33: W116-W120.
28. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.
29. Tatusov RL, Galperin M, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28: 33-36.
30. Letunic I, Bork P (2011) Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. Nucl Acids Res 39: 475-478.
31. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature methods 9: 357-359.
32. Smit AFA, Hubley R, Green P (2013-2015) RepeatMasker Open-4.0.
33. Thurston MI, Field D Msatfinder: Detection and characterization of microsatellites.
34. Barrero RA, Chapman B, Yang Y, Moolhuijzen P, Keeble-Gagnère G, et al. (2011) De novo assembly of Euphorbia fischeriana root transcriptome identifies prostratin pathway related genes. BMC Genomics 12: 600.
35. Fox SE, Geniza M, Hanumappa M, Naithani S, Sullivan C, et al. (2014) De Novo Transcriptome Assembly and Analyses of Gene Expression during Photomorphogenesis in Diploid Wheat Triticum monococcum. PLoS ONE 9: e96855.
36. Sharma N, Jung C-H, Bhalla PL, Singh MB (2014) RNA Sequencing Analysis of the Gametophyte Transcriptome from the Liverwort, Marchantia polymorpha. PLoS ONE 9: e97497.
37. Collins LJ, Biggs PJ, Voelckel C, Joly S (2008) An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome Inform 21: 3-14.
38. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. (2010) The sequence and de novo assembly of the giant panda genome. Nature 463: 311-317.
39. Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11: 400.
40. Wu T, Qin Z, Zhou X, Feng Z, Du Y (2010) Transcriptome profile analysis of floral sex determination in cucumber. J Plant Physiol 15: 905-13.
41. Finseth FR, Harrison RG (2014) A Comparison of Next Generation Sequencing Technologies for Transcriptome Assembly and Utility for RNA-Seq in a Non-Model Bird. PLoS ONE 9: e108550.
42. Gordo SM, Pinheiro DG, Moreira EC, Rodrigues SM, Poltronieri MC, et al. (2012) High-throughput sequencing of black pepper root transcriptome. BMC Plant Biology 12: 168.
43. Ashrafi H, Hill T, Stoffel K, Kozik A, Yao J, et al (2010) De novo assembly of the pepper transcriptome (Capsicum annuum): a benchmark for in silico discovery of SNPs, SSRs and candidate genes. BMC Genomics 13: 571.
44. Surget-Groba Y and Montoya-Burgos JI Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res 20: 1432-1440.
45. Chopra R, Burow G, Farmer A, Mudge J, Simpson CE, et al. (2014) Comparisons of De Novo Transcriptome Assemblers in Diploid and Polyploid Species Using Peanut (Arachis spp.) RNA-Seq Data. PLoS ONE 9: e115055.
46. Xia Z, Xu H, Zhai J, Li D, Luo H, et al. (2011) RNA-Seq analysis and de novo transcriptome assembly of Hevea brasiliensis. Plant Mol Biol 77: 299-308.
47. Shin SC, Kim SJ, Lee JK, Ahn DH, Kim MG, et al. (2012) Transcriptomics and Comparative Analysis of Three Antarctic Notothenioid Fishes. PLoS ONE 7: e43762.
48. Salisbury JP, Sîrbulescu RF, Moran BM, Auclair JR, Zupanc GKH, et al. (2015) The central nervous system transcriptome of the weakly electric brown ghost knifefish (Apteronotus leptorhynchus): de novo assembly, annotation, and proteomics validation. BMC Genomics 16: 166.
49. Gao J, Wang X, Zou Z, Jia X, Wang Y, et al. (2014) Transcriptome analysis of the differences in gene expression between testis and ovary in green mud crab (Scylla paramamosain). BMC Genomics 15: 585.
50. Windisch HS, Lucassen M, Frickenhaus S (2012) Evolutionary force in confamiliar marine vertebrates of different temperature realms: adaptive trends in zoarcid fish transcriptomes. BMC Genomics 13: 549.
51. The Gene Ontology Consortium (2008) The Gene Ontology project in 2008. Nucl Acid Res 36: D440-D444.
52. Sun F, Liu S, Gao X, Jiang Y, et al. (2014) Male-Biased Genes in Catfish as Revealed by RNA-Seq Analysis of the Testis Transcriptome. PLoS ONE 8: e68452.
53. Lu J, Luan P, Zhang X, Xue S, Peng L, et al. (2014) Gonadal transcriptomic analysis of yellow catfish (Pelteobagrus fulvidraco): identification of sex-related genes and genetic markers. Physiol Genomics 21: 798-807.
54. Maugars G, Schmitz M (2008) Expression of gonadotropin and gonadotropin receptor genes during early sexual maturation in male Atlantic salmon parr. Mol Reprod Dev 75: 403-413.
55. Tena-Sempere M (2006) KiSS-1 and reproduction: focus on its role in the metabolic regulation of fertility. Neuroendocrinology 83: 275-281.
56. Popa SM, Clifton DK, Steiner RA (2008) The Role of Kisspeptins and GPR54 in the Neuroendocrine Regulation of Reproduction. Annual Review of Physiology 70: 213-238.
57. Klausena C, Chang JP, Habibi HR (2001) The effect of gonadotropin-releasing hormone on growth hormone and gonadotropin subunit gene expression in the pituitary of goldfish, Carassius auratus. Comparative Biochemistry and Physiology 129: 511-516.
58. Li WS LH, Wong A (2002) Effects of gonadotropin-releasing hormone on growth hormone secretion and gene expression in common carp pituitary. Comparative Biochemistry and Physiology pp: 335-341.
59. Møller N, Jørgensen JO (2013) Effects of Growth Hormone on Glucose, Lipid, and Protein Metabolism in Human Subjects. Endocr Rev 30: 152-77.
60. Singh K, Deepika J (2009) One Carbon Metabolism, Spermatogenesis, and Male Infertility. Reprod Sci 20: 622-30.