Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

De novo characterization of skeletal muscle transcriptome of Atlantic herring (Clupea harengus) using Next Generation Sequencing (NGS): Effect of quality and length trimming on transcriptome assembly

Sangeet Lamichhaney

Examensarbete / Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics,
395
Uppsala 2012

Master's Thesis, 30 HEC

Erasmus Mundus Programme – European Master in Animal Breeding and Genetics

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

De novo characterization of skeletal muscle transcriptome of Atlantic herring (Clupea harengus) using Next Generation sequencing (NGS): Effect of quality and length trimming on transcriptome assembly

Sangeet Lamichhaney

Supervisors:

Leif Andersson, SLU, Department of Animal Breeding and Genetics & UU, Department of Medical Biochemistry and Microbiology
Alvaro Martinez Barrio, UU, Department of Medical Biochemistry and Microbiology
Carl-Johan Rubin, UU, Department of Medical Biochemistry and Microbiology
Examiner:
Göran Andersson, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: EX0556
Programme: Erasmus Mundus Programme - European Master in Animal Breeding and Genetics
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2012
Name of series: Examensarbete / Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, 395
On-line publication: http://epsilon.slu.se

Key words: Atlantic herring, transcriptome, RNA sequencing, *De novo* assembly

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

De novo characterization of skeletal muscle transcriptome of Atlantic herring (Clupea harengus) using Next Generation Sequencing (NGS): Effect of quality and length trimming on transcriptome assembly

Sangeet Lamichhaney

Department of Animal Breeding and Genetics
Uppsala 2011

Masters Thesis, 30 HEC
Erasmus Mundus program
European Master in Animal Breding and Genetics

Table of contents

ABSTRACT

Atlantic herring *(Clupea harengus)*, one of the most abundant fish species on earth is an economically important marine species that is found in the Baltic Sea and on both sides of the Atlantic Ocean. Although it has been a popular species for marine fish population studies since long, yet the genomic information for Atlantic herring is scarce. Recent developments in ultra high throughput RNA sequencing methods has allowed rapid and cost effective generation of large sequence information, which can be used to characterize the transcriptome in any non-model species even when no reference sequence is available. Transcriptome sequencing from the skeletal muscle of a single specimen of Atlantic herring was performed using Illumina HiSeq 2000 platform that generated approximately 116 million reads (with 101 bp length). These short reads were trimmed for quality and were assembled into 115,046 contigs with an average length of 291 bp and N50 of 375 bp, thereby producing a draft transcriptome assembly with total size of 33.51 Mb. With the e-value threshold set to 10^{-4}, 46,979 contigs (40.84%) were identified to have matches against GenBank non-redundant (NR) proteins and Zebrafish unigenes database. Using the annotated transcriptome resource, 25,431 putative allelic variants (24,351 SNPs and 1080 indels) were identified. The present study provides a comprehensive muscle transcriptome resource which will be particularly useful for the validation of draft genome assembly of Atlantic herring that is currently being established within our group.

1. BACKGROUND

1.1. The Atlantic herring (*Clupea harengus)*

Atlantic herring is an economically important marine species found in the Baltic Sea and on both sides of the Atlantic Ocean. They are pelagic and shoaling fish and normally congregate in very dense populations (Andersson et al., 1981). It is one of the most abundant fish species on earth and it has a long history of heavy exploitation (Larsson et al., 2009). Atlantic herring has significant economic and cultural importance and possess exceptional heterogeneity in life history, morphology and behavior; hence it is a popular species for marine fish population studies (Hauser et al., 2001). One of the early studies identified 13 polymorphic protein loci in Atlantic herring using starch gel electrophoresis that were considered usable in routine population surveys (Andersson et al., 1981). Several population genetic studies have been done in Atlantic herring using microsatellite markers (Shaw et al., 1999, Ruzzante et al., 2006, Larsson et al., 2007, Gaggiotti et al., 2009, Larsson et al., 2010).

There are no available reports about the genome size of Atlantic herring. The closest relative of Atlantic herring, the Pacific herring *(Clupea pallasii)* has a diploid genome with 26 pairs of chromosomes (2n=52) with C-value (amount of DNA contained in the haploid nucleus) estimates ranging from 0.76 to 0.98 picograms (pg) (http://www.genomesize.com/). Genome size in the Pacific herring is thus estimated to be in the range of 0.7 to 0.9 GB (1 pg ~ 921 Mb) and we expect a similar size for Atlantic herring, which would make its genome half of the size of the zebrafish *(Danio rerio)* genome (http://www.ensembl.org, release 62). Atlantic herring (Clupeomorphs) is evolutionarily closer to zebrafish (Cypriniformes) as compared to other fish whose genome sequences are publicly available (medaka (*Oryzias latipes*), stickleback (*Gasterosteus aculeatus*), fugu (*Takifugu rubripes*) and tetraodon (*Tetraodon nigroviridis*), i.e. percomorphs) (Santini et al., 2009) (Figure 1).

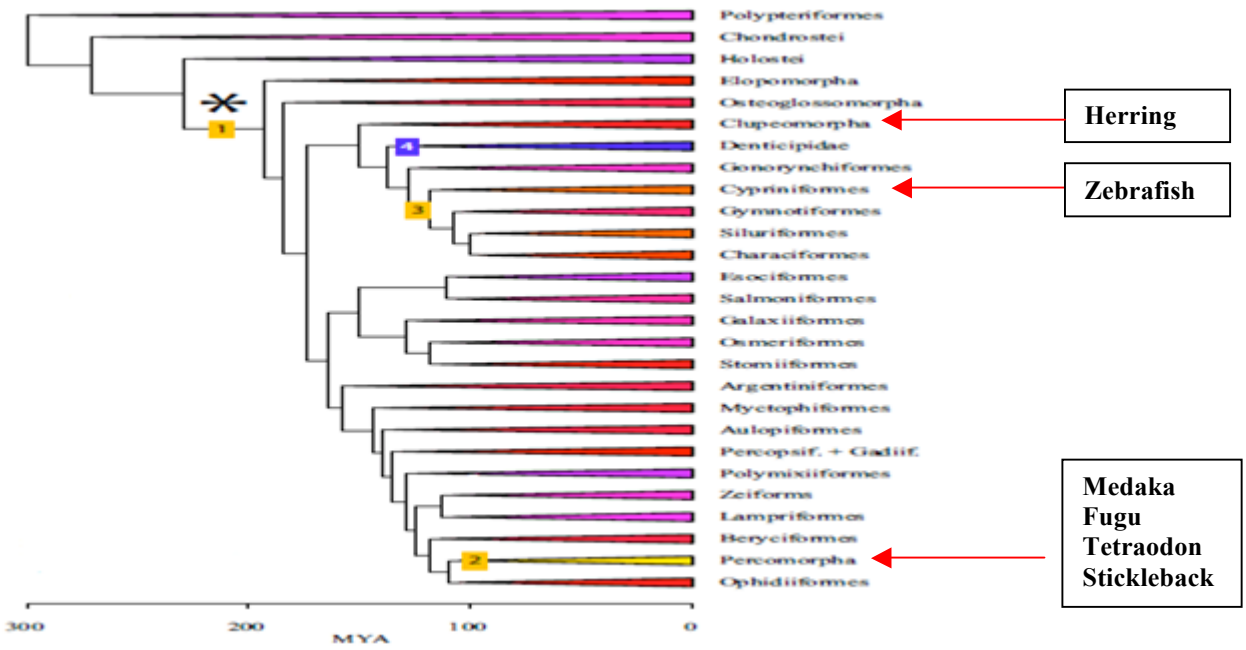


Figure 1: Evolutionary tree of ray-finned fish (* indicates fish specific whole genome duplication event, MYA: Time period in "Million years ago", Red arrow indicates evolutionary position of different fish whose reference genomes have been established fish: Cypriniformes (zebrafish), Percomorphs (medaka, fugu, tetraodon and stickleback) and fish under our study, herring (Clupeomorphs); Based on Santini et al., 2009)

Currently, about 418 Expressed Sequence Tags (ESTs) generated from different Atlantic herring tissues together with 248 nucleotide-sequence and 204 protein-sequence records are available from the NCBI database (http://www.ncbi.nlm.nih.gov/). These records include the mitochondrial genome of Atlantic herring (16.7 kb) with gene content (13 protein coding genes) and gene order similar to majority of the vertebrates (Lavoue et al., 2007). An effort to establish a draft genome assembly of the Atlantic herring by Illumina sequencing methods is currently undergoing by our group at IMBIM, Uppsala University, Sweden (PI Leif Andersson). This thesis is thus a parallel project to analyse the transcriptome of skeletal muscle from a single specimen of Atlantic herring, which will also serve to annotate the genome of the same individual.

1.2. The transcriptome

A transcriptome is the total set of transcripts present in a particular cell type at a specific time of a given organism. It is a representation of the total number of expressed transcripts in the

defined population of the cells, whose levels of expression is modulated by different developmental and differentiation stages, tissue types, or physiological conditions (Johansen et al., 2010). Thus, unlike the genome, which is a stable entity, the transcriptome is highly flexible with temporal and spatial variations and serves as a dynamic link between the genome and physical characteristics of an organism (Velculescu et al., 1997). Understanding the transcriptome provides information on various aspects of cell biology and biochemistry; e.g. the amount of gene activity during various stages of development or gene expression changes associated with a particular disease (Adams., 2008).

The key aim of a transcriptome study is to catalogue all the transcripts expressed in the tissue under study, determine the transcriptional structure of the genes (5' and 3' ends, start sites, splicing patterns and other post-transcriptional modifications) and to quantify expression level of each transcript under various physiological conditions (Wang et al., 2009). Understanding functional complexity of a transcriptome is a challenging task as most of the genes are transcribed in bidirectional manner and much more of the genome is transcribed than previously expected (Hegedus et al., 2009).

1.3. Methods used in transcriptome studies

1.3.1. Microarrays

The approach for large-scale studies of gene expression levels and transcriptome quantification has progressed from candidate gene-based detection of RNA by Northern blotting to various hybridization and sequence based approaches (Morozova et al., 2009). In recent past, the most widely used methodology for transcriptome analysis has been expression microarrays which is a high throughput approach involving extraction of mRNA from the tissue under study, labelling its cDNA with fluorescent dye and hybridization to high density DNA sequences immobilized in an ordered array on a solid surface (Schulze et al., 2001, Wang et al., 2009). Microarrays can be used to monitor the expression of thousands of genes

simultaneously, thereby enabling analysis of the temporal and spatial variations in the transcriptome (Mathavan et al., 2005). However, microarray technology is limited by sensitivity due to complex background hybridization and cross hybridization, poor dynamic range of detection compared to RNA-seq technology due to signal to noise ration at the low end and saturated signals at the high end of expression. In addition, normalization procedures to compare results across different experiments are complicated. Furthermore, microarrays are only able to quantify transcripts whose reference sequence is already known. These limitations make microarrays less useful for large-scale transcriptome analysis of non-model organisms whose reference genome are incomplete or yet to be established (Casneuf et al., 2007, Ledford, 2008, Pariset et al., 2009).

1.3.2. Sequence-based approaches

1.3.2.1. ESTs

Sequencing cDNA clones to generate Expressed Sequence Tags (EST) using traditional Sanger based technology was the early sequence based approach for transcriptome analysis (Adams et al., 1991), but this approach was limited by low throughput, high cost and quantification issues (Mortazavi et al., 2008, Wang et al., 2009). Tag based approaches like Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995) and SAGE-like methods (cap analysis of gene expression (CAGE) (Kodzius et al., 2006), massively parallel signature sequencing (MPSS) (Reinartz et al., 2002)) were developed later and these methods all use information from short sequence tags (14 -20 bp) information from 3' (or 5') end of transcripts to calculate "digital" transcript abundance. These tag-based methods overcame most of the limitations of hybridization-based systems but they were costly and could not detect splicing events (Sultan et al., 2008). In addition, as they were based upon traditional Sanger sequencing system, they still faced the limitations of laborious cloning procedures.

1.3.2.2. RNA sequencing

A potentially more comprehensive approach to high-throughput transcriptomics is the usage of a cost effective and ultra-high-throughput DNA sequencing technology; RNA-seq, which in recent years has revolutionized the analysis of eukaryotic transcriptomes (Mortazavi et al., 2008, Wang et al., 2009). RNA-seq has clear advantages over other approaches; it does not require bacterial cloning of cDNA and can generate short reads of cDNA at an extraordinary depth (Shi et al., 2011) and it is not limited to detecting transcripts mapped to existing genomic sequences, hence it is an attractive tool for transcriptome studies in non-model organism whose genomes are either not yet sequenced or poorly annotated (Vera et al., 2008). In addition, it allows detection and quantification of new splice isoforms and low abundant transcripts, determination of transcriptional boundaries of genes at single nucleotide resolution and identification of expressed Single Nucleotide Polymorphisms (SNPs) (Costa et al., 2010). Additional advantages of RNA-seq over microarrays include large dynamic range of expression levels over which transcripts can be detected with no upper limit for quantification and high levels of reproducibility of RNA-seq data for both technical and biological replicates (Wang et al., 2009).

There have been many papers published in recent years that have demonstrated the advantage of using RNA-seq for transcriptome analysis in different eukaryotic species (e.g. Mortazavi et al., 2008, Nagalakshmi et al., 2008, Lister et al., 2008, Blekhman et al., 2010, Shi et al., 2011). RNA-seq studies have also been done in fish and other aquaculture and marine species (Johansen et al., 2010, Liu et al., 2011, Franchini et al., 2011, Wei et al., 2011).

Roche 454, Illumina, and ABI SOLiD System sequencing platforms have become widely available over the past few years for generating ultra-high-throughput sequence data. Among these platforms, 454 sequencing technology has been most popular for transcriptome analysis

(Coppe et al., 2010, Fraser et al., 2011) but usage of Illumina sequencing technology is increasing gradually due to its continuous improvement in read length, quality and sequencing throughput (Surget-Groba et al., 2010, Cirulli et al., 2010). The Illumina systems have been widely used to sequence transcriptome of organisms with available reference genomes as well as non-model organisms. The latest Illumina sequencing machines generate between 75 to 100 bp reads with paired end sequence information, which has improved the construction of reliable *de novo* assemblies of transcripts for gene expression, novel gene discovery and comparative genomics studies (Franchini et al., 2011).

1.4. Aims of the study

Skeletal muscle tissues collected from a single specimen of the Atlantic Herring was used to generate paired short sequence reads (2 X 101 bp) using Illumina HiSeq 2000 technology to characterize the Atlantic herring skeletal muscle transcriptome. The major objectives of this study were;

- Construction of a *de novo* assembly of the transcripts expressed in the skeletal muscle and evaluate the quality of the assembly generated by different RNA-seq assembler that are currently available
- Comparative analysis of the assembly with the transcriptome of evolutionary close fish species whose draft genome assembly has been established (e.g. zebrafish)
- Estimation of the total number of transcripts expressed in skeletal muscle of Atlantic herring and evaluate the transcriptome redundancy
- Investigation of G protein coupled receptors (GPCRs) genes expressed in skeletal muscle in Atlantic herring
- Identification of putative allelic variants in the assembled transcripts

2. MATERIALS AND METHODS

2.1. Library construction and sequencing

Skeletal muscle tissue collected from a single specimen of the Baltic Herring (Atlantic herring found in Baltic Sea), caught in the archipelago of Stockholm in June, 2010 was used for RNA extraction and downstream transcriptome sequencing. mRNA selection, library preparation and sequencing were performed by the SciLife lab-Stockholm on an Illumina HiSeq 2000 sequencer. The pipeline of the protocol started with the purification of mRNA from the total RNA using polyA selection (Figure 2A) followed by fragmentation and conversion into single-stranded cDNA using priming of random hexamers priming. The second strand was generated to create double stranded cDNA (Figure 2B) for library preparation workflow. Library construction started by the generation of blunt-end DNA fragments using exonuclease activity (Figure 2C) followed by addition of "A" base (Figure 2D) to these blunt-ends to prepare for ligation of sequencing adapters (Figure 2E) (since each adapter has "T" base complementary overhang on 3' end). Final product was created after ligation, denaturation and amplification steps (Figure 2F). Average insert size of the cDNA library was 240 bp (Min: 180 bp, max: 400 bp).

This library was pooled in the flow cell of cBot cluster generation system (Figure 2G) that created clonal clusters from DNA template to prepare them for sequencing. DNA library samples were bound to complementary adapter oligos grafted in the flow cell and were copied by 3' extension using hybridized primers. These copies were further isothermally amplified to generate clonal clusters of approx. 1000 copies each, ready for sequencing. Sequencing was done on an Illumina Hiseq 2000 platform that generated paired end reads, each with 101 bp in length (Figure 2H). Base calling was done using Illumina BclConverter v1.7.1 and the final read dataset was deposited into the Uppnex system (at Uppsala University) for downstream bioinformatics analysis.

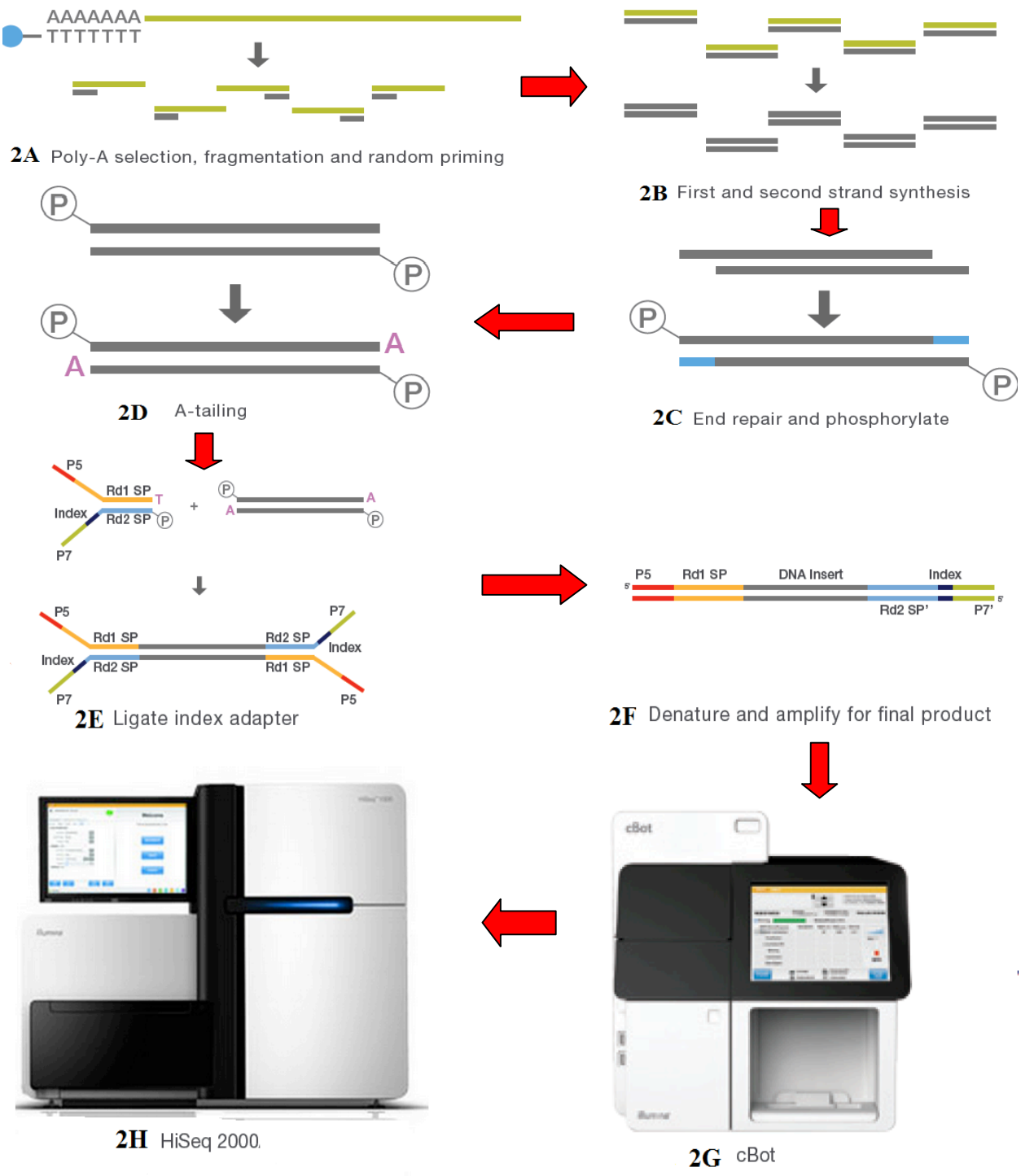


Figure 2: Workflow of library preparation and sequencing (mRNA is polyA selected and fragmented (2A), synthesis of first and second strand cDNA (2B), and creation of blunt end fragment (2C), Addition of "A" base to prepare for ligation to adapter sequence (2E). Final product is created (2F) which is amplified in cBot (2G) and sequenced in Illumina HiSeq platform (2H)

(Source: www.illumina.com/documents/%5Cproducts%5Cdatasheets%5Cdatasheet_truseq_sample_prep_kits.pdf)

2.2. Reads pre-processing

The reads datasets received from sequencing platform only contained sequences that had passed Illumina chastity filter. It is an Illumina sequencing pipeline quality filter, which uses the standard threshold of Chastity >= 0.6 (http://illumina.ucr.edu/ht/documentation/file-formats). Chastity for a given base call is defined as "the ratio of the highest of the four (base type) intensities to the sum of highest two", which is used to screen the clusters with low signal to noise ratio.

The quality of these raw sequence reads received was assessed using FASTX toolkit, version 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit) (Appendix 1). The Phred-like quality score (Q-score) which is the logarithmic calculation of base calling error probability P ($Q = -10\log_{10}P$), assigned to each base by Illumina pipeline software suite was plotted to estimate quality score distribution of the bases throughout the read. A cut-off threshold of Q-score 30 (i.e. a probability of an incorrect base call of 1 in 1000) was chosen and bases with Q-score less than 30 were trimmed using FASTX toolkit. A stacked-histogram graph for the nucleotide distribution was also plotted to estimate base diversity in the reads.

The trimmed reads were further screened to filter low quality and low complexity sequences as well as poly A/T tails using the latest version of the "Seqclean" tool (installed on March 4, 2011) (http://sourceforge.net/projects/seqclean/files) (Appendix 2). Simultaneously, the reads were also screened against the Univec database (containing sequences of vector origin and sequences of adapters, linkers and primers commonly used in the process of cloning cDNA or genomic DNA) (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) using "Seqclean" tool to avoid vector contamination and filter possible adapter sequences.

2.3. *De novo* assembly of sequencing reads

2.3.1. Inchworm RNA-seq assembly

In first phase of the study, *de novo* assembly of the reads was performed using the Inchworm RNA-seq assembler[1] (2011-01-20 release version) (http://inchworm.sourceforge.net/) (Appendix 3) which implements a *de Bruijn* graph algorithm (Pevzner et al, 2001) to construct transcripts from Illumina RNA-seq reads. Briefly, the cleaned reads were first split into Kmers (smaller pieces of reads of a user defined size) which were used as seeds to assemble contigs by extending them in both direction on the basis of sequence overlap between the Kmers (Grabherr et al, unpublished) (Figure 3).

Figure 3: Inchworm RNA-seq assembler algorithm described: A) reads are split into Kmers and stored in hash table; B) most frequently occurring Kmer is used as seed; C) construction of a *de bruijn* graph with each node representing series of overlapping Kmers; D) extension of Kmers at both ends in a coverage guided manner; E) extension continues till no Kmer exists to provide extension, generating a draft contig; F) entire cycle (B-E) repeats until all Kmers in hash table are exhausted generating assembly with set of linear contigs) (Based on: Grabherr et al., 2011)

The choice of the best Kmer size is particularly important for transcriptome assembly using *de Bruijn* graph algorithm; higher Kmer length results in a more contiguous assembly of highly expressed transcripts while poorly expressed transcripts are better assembled if lower Kmer lengths are used (Surget-Groba et al, 2010). To optimize our transcriptome assembly, an in-built script in Inchworm RNA-seq assembler was used to generate different

[1] Reads used for Inchworm assembly were not filtered by "Seqclean" tool

transcriptome assemblies using range of Kmer lengths (19 to 29), so we could evaluate the assembly statistics of each assembly. A Kmer size of 21 provided best result among all as it had the lowest number of contigs whereas N50 (i.e. the contig length such that using equal or longer contigs, we could gather half of the bases of the transcriptome assembled) and mean contig size were higher as compared to other assemblies (see "*de novo* assemblies of reads" in Results section 3.2.)

The assembly generated with 21-mer size was further reassembled using latest version of CAP3 assembler (installed on January 25th, 2011) (http://seq.cs.iastate.edu/cap3.html) (Appendix 4), that uses a traditional Overlap-Layout-Consensus algorithm (OLC) and merges the overlapping contigs in the assembly, in an attempt to join improperly assembled contigs from a single transcript into a single contig.

2.3.2. Trinity RNA-seq assembly

During the course of this study, a new RNA-seq assembly package "Trinity" was released which incorporated the previously released Inchworm RNA-seq assembler into one of its three software modules (Inchworm, Chrysalis and Butterfly) (http://trinityrnaseq.sourceforge.net/). The developers of "Trinity" claimed that the new incorporated assembler was more sophisticated and generated better assemblies compared to the previous Inchworm version, in particular, with regard to chimeric transcripts resolution and reports unique contigs for alternatively spliced isoforms and paralogous genes (Grabherr et al., 2011). Briefly, the algorithm of "Trinity" followed these subsequent steps: first, Inchworm generated transcripts for a dominant isoform, only reporting unique portions of alternatively spliced transcripts; second, Chrysalis clustered the Inchworm contigs that shared sequences supported by read pairs spanning over the potential junctions of the contigs into components, computed a *de bruijn* graph for each component and partitioned full read set

among these graphs; finally, Butterfly processed these individual graphs and reconstructed distinct transcript contigs for splice isoforms, and paralogous genes (Grabherr et al., 2011).

Trinity also estimated the coverage of each contig based on fragments per kilobase of exon per million fragments mapped (FPKM) score (measure of coverage that is normalized by both the transcript lengths, as well as by the total number of reads that could be mapped to any transcript) (Grabherr et al., 2011). The information about components, total contigs including splice isoforms and transcripts from paralogous genes that were clustered into a single component, their respective FPKM score and contig size could be extracted from the headers of each contigs assembled by Trinity RNA-seq assembler. The *de novo* assembly of the reads was re-performed using the "Trinity" RNA-seq assembler (2011-03-13 release version) (Appendix 5) with Kmer size 21 (this Kmer size was estimated to be best suited based on the assembly statistics such as total contig size, N50 and mean contig length) from the results of Inchworm RNA-seq assembly) (Table 4) and this assembly was used for further analyses.

2.4. Contigs validation

2.4.1. Alignment against phiX genome

Validation of the assembled contigs is an important step to distinguish between an actual transcript and spuriously assembled sequences. The first validation step was to align our contigs set with the "phiX" genome. The Illumina Hiseq 2000 sequencing protocol included the spiking of the samples in each lane with 1% "phiX" like control. PhiX (or phiX174) is a bacteriophage with very small and well-characterized genome (5386 bp), which is regularly used as a control for sequencing instruments. A single contig with a significant full length hit to "phiX" genome was expected as a result of good quality assembly. All contigs were aligned against phiX genome using the blastn program (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST) (installed on April 1st, 2011) (Appendix 6).

2.4.2. Alignment against Atlantic herring complete protein coding sequences from NCBI

Four full-length mRNA sequences (non-mitochondrial) of Atlantic herring reported to be expressed in muscle (Table 1) were downloaded together with 13-mitochondrial protein-coding genes (Table 2) from NCBI. We aligned these coding sequences against our contigs set using blastn tool (Appendix 6) to check if these previously reported Atlantic herring mRNAs were assembled in our transcriptome assembly.

Table 1: List of reported Atlantic herring full length mRNAs expressed in muscles*

Accession ID	Description	Length (bp)	Author
U20111	*Clupea harengus* clone TnI-alpha troponin-I mRNA	740	Hodgson,P.A, 1995
AF312364	*Clupea harengus* muscular cathepsin D mRNA	1191	Nielsen,L.B. et al, 2000
GQ455648	*Clupea harengus* alpha actin mRNA (fast)	1540	Mercer,R.C.C, 2009
EF495203	*Clupea harengus* alpha actin mRNA (slow)	1421	Mercer,R.C.C, 2009

***Source: National Centre for Biotechnology Information**

Table 2: List of Atlantic herring mitochondrial genes*

S.N.	Accession ID	Description	Length (bp)
1	YP_001293658.1	NADH dehydrogenase subunit 1	972
2	YP_001293659.1	NADH dehydrogenase subunit 2	1044
3	YP_001293660.1	Cytochrome c oxidase subunit I	1598
4	YP_001293661.1	Cytochrome c oxidase subunit II	690
5	YP_001293662.1	ATP synthase F0 subunit 8	165
6	YP_001293663.1	ATP synthase F0 subunit 6	684
7	YP_001293664.1	Cytochrome c oxidase subunit III	783
8	YP_001293665.1	NADH dehydrogenase subunit 3	348
9	YP_001293666.1	NADH dehydrogenase subunit 4L	294
10	YP_001293667.1	NADH dehydrogenase subunit 4	1380
11	YP_001293668.1	NADH dehydrogenase subunit 5	1833
12	YP_001293669.1	NADH dehydrogenase subunit 6	519
13	YP_001293670.1	Cytochrome b	1140

*** Source: Lavoué S, *et al,* National Centre for Biotechnology Information**

2.5. Annotation

Annotation of all assembled contigs (longer than 100 bp) was based on assignment of putative gene descriptions by comparing contigs with NCBI Non-redundant (NR) protein database (ftp://ftp.ncbi.nih.gov/blast/db, downloaded on 15th April, 2011) using blastx tool (Appendix 6). NR database consists of entries from all non-redundant GenBank coding sequences (CDS) translations along with entries from PDB (Protein data bank), SwissProt, PIR (Protein Information resource) and PRF (Protein research foundation) databases. Every alignment with an e-value < 10^{-4} was considered to be a significant hit.

A large proportion of assembled genes in different species have not yet been annotated completely (Shi et al, 2011), so we had every chance of missing the annotation of a number of contigs by comparing only to NR database. Hence, we further compared the set of contigs that had no hits to NR protein database to NCBI Unigene records for zebrafish (http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=7955, downloaded on 1st May, 2011) using tblastx tool (Appendix 6) with an e-value cut-off of 10^{-4}. Zebrafish Unigene records consist of 52,653 clustered sequence entries of all known genes and Expressed Sequence Tags (ESTs) from GenBank and dbEST respectively.

The features like all hits (list of hits to different subjects for each contig) and best hits (top hit to particular subject with maximum e-value), putative gene names, predicted proteins, similarity and E-value distribution as well as other statistics were parsed from BLAST output using a BLAST parser tool (http://www.atgc.org/BlastParser) along with custom Perl and UNIX scripts.

2.6. Transcriptome redundancy

Investigation of redundancy in the transcriptome assembly is a useful step to evaluate the quality of the assembly and as an initial step for the detection of alternative splice isoforms

and gene duplication events. We analysed the entries of NR proteins and zebrafish Unigene database that had BLAST hits to more than one contig to examine the level of transcriptome redundancy in the assembly.

2.7. Investigation of G-protein-coupled-receptor (GPCR) genes expressed in skeletal muscles of Atlantic herring

We investigated the contigs in the assembly that represented the GPCRs expressed in skeletal muscle of Atlantic herring. The GPCR protein sequences reported in various species were downloaded from GPCR database (GPCRDB) (http://www.gpcr.org/7tm/, version 2011.03.16). Contigs from our assembly that were homologous to the proteins in GPCRDB were identified using a comprehensive approach explained in Figure 4. First, all the contigs were aligned against the protein sequences from GPCRDB using blastx tool and e-value cut-off of 10^{-5}. Top hits (best hit to a particular protein with a low E-value) were extracted from the blast output and only those contigs that were aligned for at least 80% of their length were selected. This criterion of query coverage was set up to filter those hits showing false homology, an artifact likely due to the sharing of some short common sequence or motif region with the subject. In previous GPCRs identification studies in Xenopus (Ji et al, 2009) and fish (Niimura and Nei, 2005), protein sequences less than 250 amino acids (AA) (i.e. 750 bp coding sequences) were supposed to be too short for stretching seven transmembrane regions, the important feature of GPCRs. Hence, among the contigs with significant hits, contigs less than 750 bp (minimum length of nucleotides required to generate seven trans-membrane regions, which is an important feature of a GPCR protein) were also removed, finally generating putative GPCRs that should be expressed in skeletal muscle of Atlantic herring.

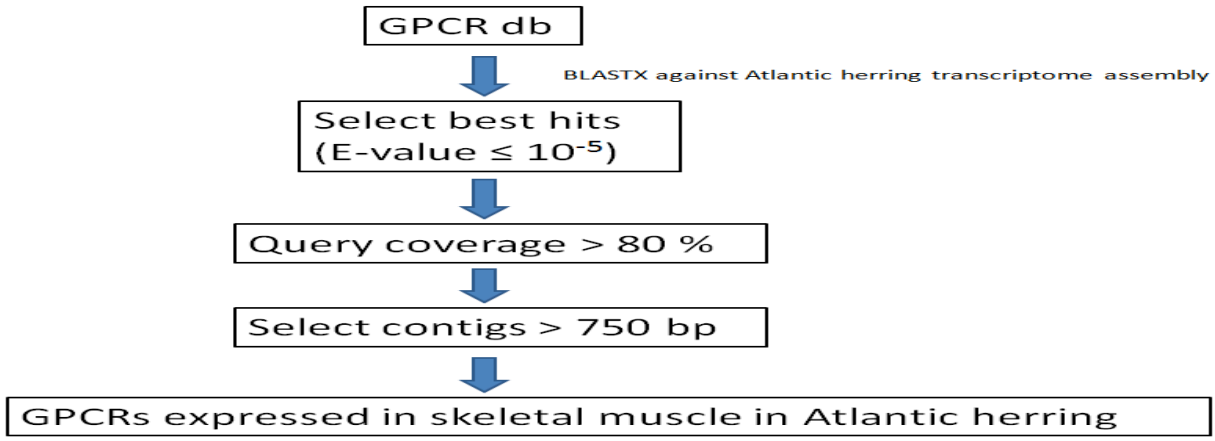


Figure 4: Strategy for identification of GPCRs expressed in skeletal muscle in Atlantic herring. The contigs were aligned against GPCRs from GPCRDB. The hits with e-value ≤ 10^{-5}, query coverage more than 80% and contig size at least 750 bp were selected as putative GPCRs.

2.8. Identification of putative allelic variants (SNPs/Indels) in the transcriptome assembly

Assembled contigs were scanned for variants using MosaikAligner (version 1.1.0017), samtools (version 0.1.16) and VarScan (version 2.2.5) packages (Appendix 7). First, MosaikAligner was used to perform a pairwise gapped alignment between all the filtered "clean" reads and reference sequences (in our case, assembled contigs). The alignment file was converted into "BAM" format using MosaikText tool (supplementary package of MosaikAligner). The "BAM" format is the standard pre-requisite input file format for samtools. Samtools generated a "pileup" file with the information about each aligned reference base, its corresponding read bases, read qualities and alignment mapping qualities. This "pileup" file was used as input file in the VarScan tool for calling SNPs and indels.

In order to identify quality SNPs and indels, specific screening criteria based on the read depth, read quality and p-value threshold were set up; only those variants with minimum read depth of 8, minimum base quality of 15 (Phred score at a variant position) and p-value threshold of 0.05 were declared as quality SNPs/indels. P-value is the estimation of significance of variant calling, which is calculated using a Fisher's Exact Test on the read counts supporting reference and variant alleles (http://varscan.sourceforge.net/support-

faq.html#output-p-value). Raw variants were further filtered based on the allele frequency; only those with a variant allele frequency between 20 to 60% were selected as putative significant variants to filter out the false positives due to sequencing errors.

3. RESULTS

3.1. Illumina sequencing and reads pre-processing

A total of 116.06 million reads that passed the Illumina Chastity filter were obtained in a single sequencing run, generating approximate 11.8 gigabase (Gb) of raw data (Table 3).

Table 3: The Atlantic Herring transcriptome dataset read statistics

	Total sequences (n)	Number of bases (bp)	Mean length (bp)	GC content (%)
Raw reads*	116,066,452	11,722,711,652	101	52.76
Trimmed reads**	116,066,452	7,428,252,928	64	52.22
Clean reads***	97,403,081	6,233,797,184	64	52.36

*Raw reads received from Illumina sequencing platform, **Reads after trimming first 12 and last 25 bases, ***Reads after filtering vector contamination, adapters, linkers and poly-A/T sequences from the Univec database

The quality score distribution of the raw reads showed that median quality score was higher than 30 (an error probability of 0.001 in log Phred scale) for almost two-third of the bases in the reads whereas the median quality degraded sharply towards the last one-third of the bases (Figure 5). Hence we decided to trim the last 25 bases of the reads considering them as of being “low quality” (less than 30). Similarly, nucleotide distribution plot of the raw reads showed a decrease in base diversity at the first 12 base positions, which were the sites that were primed by the “random hexamers” during mRNA-seq protocol (Figure 6). These first 12 bases were also trimmed; generating “Trimmed” reads with length of 64 bp (Table 3).

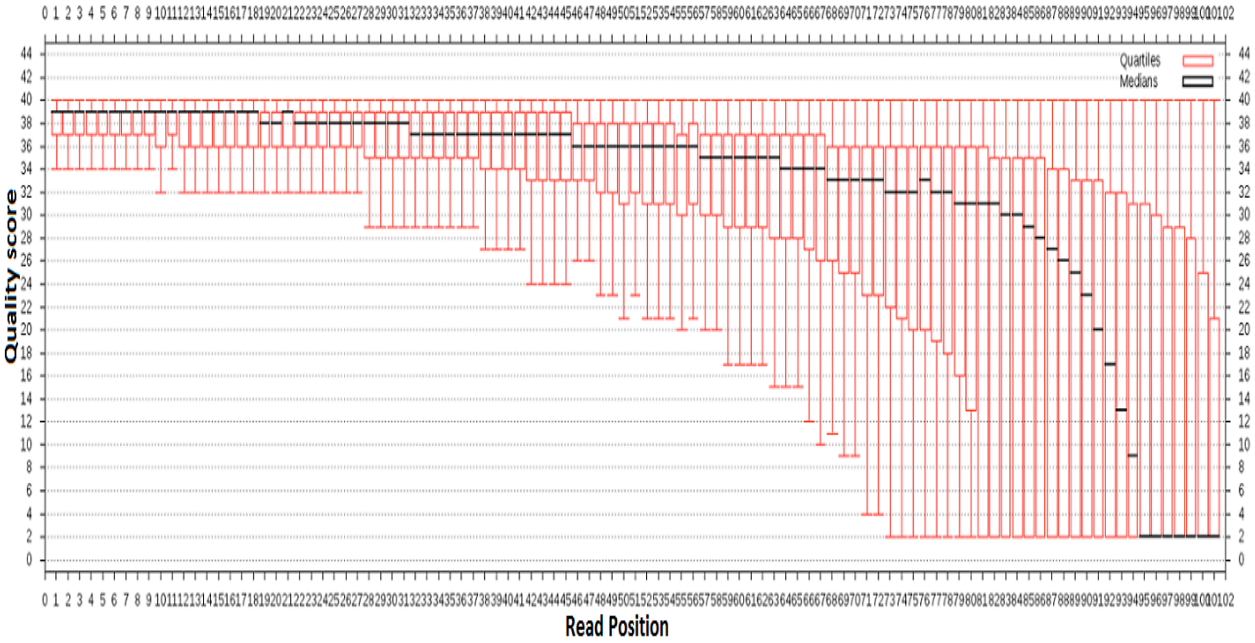


Figure 5: Quality distribution of raw reads

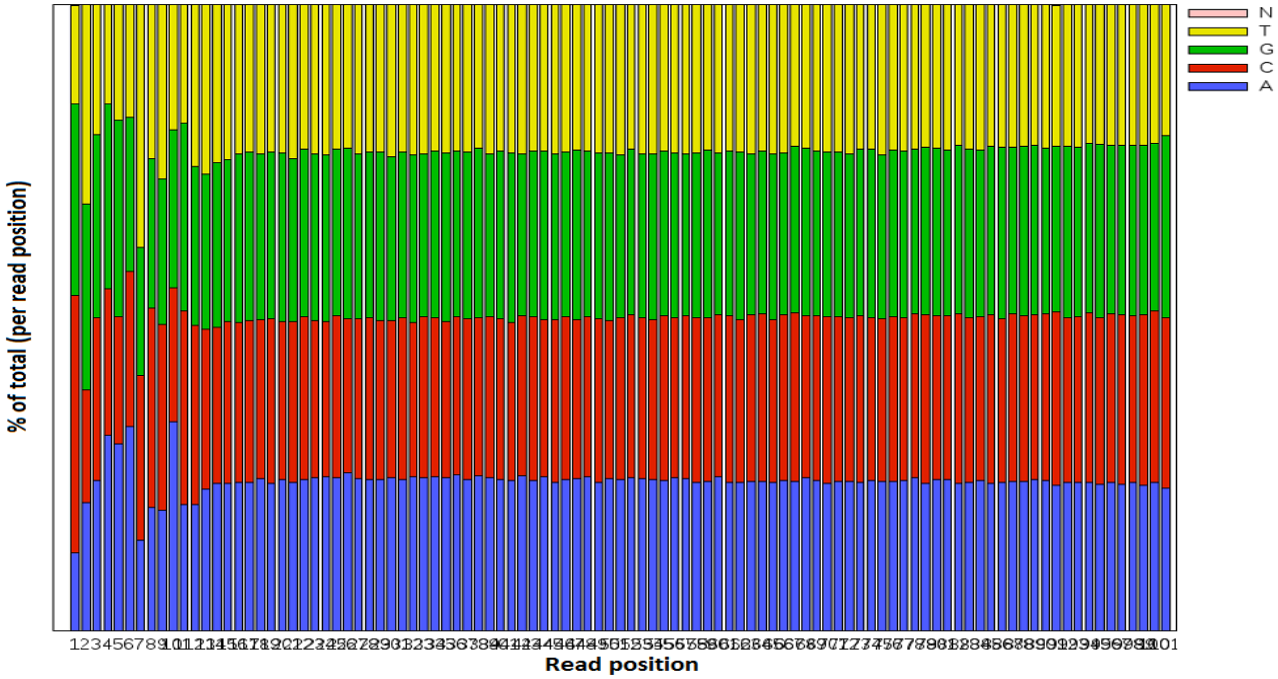


Figure 6: Nucleotide distribution of raw reads

The trimmed reads were further screened to remove low complexity sequences, poly A/T tails and vectors/adapters/linkers contained in the "Univec" database using the "Seqclean" tool. After this cleaning step, only 97.40 million "high quality" clean reads with 64 bp length (6.23 GB, 53.18% of raw data) remained (Table 3).

3.2. *De novo* assembly of reads

In the first phase of the study, all the high quality reads were assembled *de novo* using "Inchworm RNA-seq" assembler. We repeated the assembly process using a range of Kmer values (K=19 to K=29) and compared them based on assembly statistics to identify the optimum transcriptome assembly. The assembly statistics of choice were total number of contigs generated, N50 size, maximum contig length, total number of reads used in the assembly and their average coverage. Read mappings to the assembly were done using SOAPaligner, (version 2.21) (http://soap.genomics.org.cn/soapaligner.html). The assembly generated using K-mer value 21 had the lowest number of contigs with the largest N50 size and the longest mean contig length although the maximum length of the contigs was the lowest among the six Inchworm assemblies (Table 4). Based on these assembly statistics, Kmer 21 was selected as the optimal Kmer for our transcriptome assembly.

The assembly generated with Kmer 21 was further re-assembled with CAP3, in an attempt to merge the overlapping contigs. The CAP3 tool merged 52,088 contigs into 6,366 clusters and left remaining 195,688 singletons, thus generating a final assembly with 202,054 contigs. The N50 size (366 bp) and mean contig length (273 bp) was increased whereas the total sum of bases assembled was decreased by 11.23% when compared to the original Inchworm (using K=21) assembly.

Table 4: Summary statistics of *de novo* assemblies

Assembly type	No. of reads used for assembly	% reads assembled	Total contigs	Mean contig length	Max contig length	N50	Total sum of the assembly	Average coverage**
Inchworm* (using K=19)	116,066,452	77.57	255,553	242	37,683	280	62,022,150	92.91
Inchworm* (using K=21)	116,066,452	79.55	247,776	251	23,346	302	62,340,616	94.79

Inchworm* (using K=23)	116,066,452	80.79	253,089	246	29,414	292	62,500,492	96.03
Inchworm* (using K=25)	116,066,452	81.45	258,299	241	38,863	279	62,458,124	96.87
Inchworm* (using K=27)	116,066,452	82.02	266,923	234	42,550	263	62,664,533	97.22
Inchworm* (using K=29)	116,066,452	82.47	273,207	228	60,341	249	62,311,668	98.33
Trinity*** (Using K=21)	97,403,081	81.40	115,046	291	10,568	375	33,507,289	151.43

*Reads used for Inchworm assemblies were not filtered for low complexity sequences, poly A/T tails and vectors/adapters/linkers contaminations
**Total number of assembled read bases/Total number of bases in consensus sequence
***Trinity assembly was performed with reads filtered for low complexity sequences, poly A/T tails, vectors/adapters/linkers and vector contamination from the "Univec" database

Later, we re-performed the *de novo* assembly by using the "Trinity" RNA-seq assembler released during the course of our study. For this assembly, we used a Kmer value of 21 (the Kmer size estimated from previous Inchworm assemblies to yield better results). Over 79 million reads (81.40% of the total reads) were assembled into 115,046 contigs with length ranging from at least 100 bp to 10,568 bp with an average contig length of 291 bp and N50 size of 375 bp (Table 4). The total size of the assembly was 33.51 Mbp with an average sequencing depth of 151.43x.

The distribution of FPKM score (i.e. a measure of how transcripts are expressed relative to each other; see "Trinity RNA-seq assembly" in Materials and Methods section) showed that the majority of the contigs in the assembly had FPKM score less than 4 and only 10.70% of them had a score higher than 10 (with maximum score of 92,787) (Figure 7).

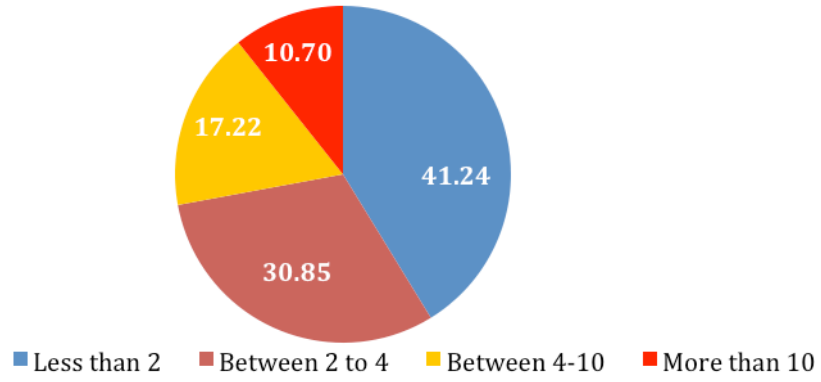


Figure 7: Distribution of FPKM scores between the contigs (White numbers within the chart shows the percentage of contigs with certain FPKM score)

The size distributions of these contigs showed that approx. 88% of total contigs were less than 500 bp (Figure 8). Although the majority of contigs were small (100 to 400 bp), there were 4,508 large contigs (longer than 1000 bp).

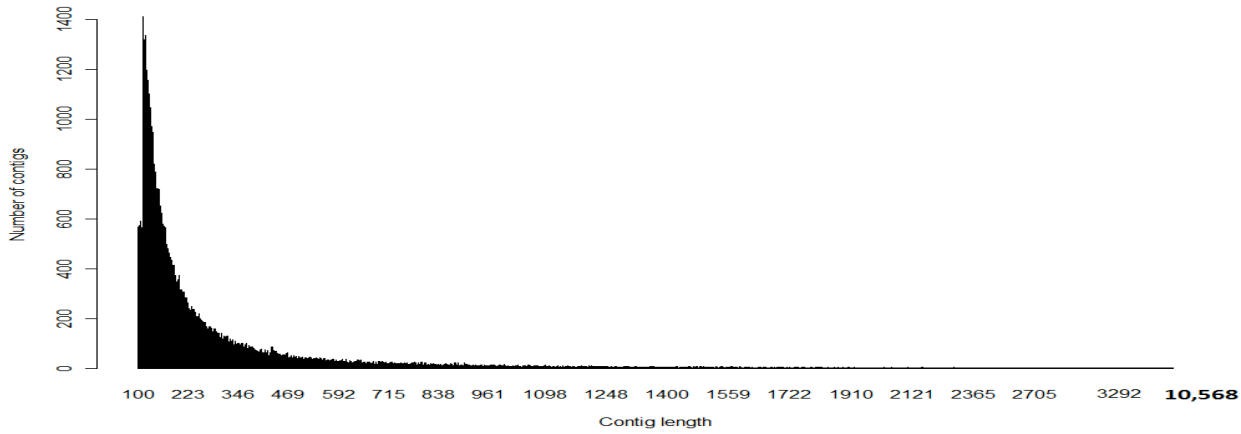


Figure 8: Length distribution of contigs from Trinity Assembly

Trinity grouped these 115,046 contigs into 101,554 different components (a component is a cluster of contigs that share common sequence supported by read pairs spanning over potential junctions; see "Trinity RNA-seq assembly" in Materials and Methods section). 90,488 components were singletons whereas the remaining 11,066 components had at least two to maximum eight contigs within each component.

A comparison between the assemblies produced using Inchworm merged with the CAP3 assembler and the Trinity assembly showed a clear improvement in favour of using Trinity. This decision was based on assembly statistics like total number of contigs, N50 size and the total size of the assembly (Figure 9 and 10). The total number of contigs in the Trinity assembly was reduced by approx. 40% with an increment in N50 size in comparison to Inchworm assembly. The total assembly size (33.51 Mbp) was about 3.5% of the expected genome size (~ 1 GB) for Atlantic herring. Due to the improvement reflected in the assembly statistics; we used the contigs produced with the "Trinity" assembly for our further analyses.

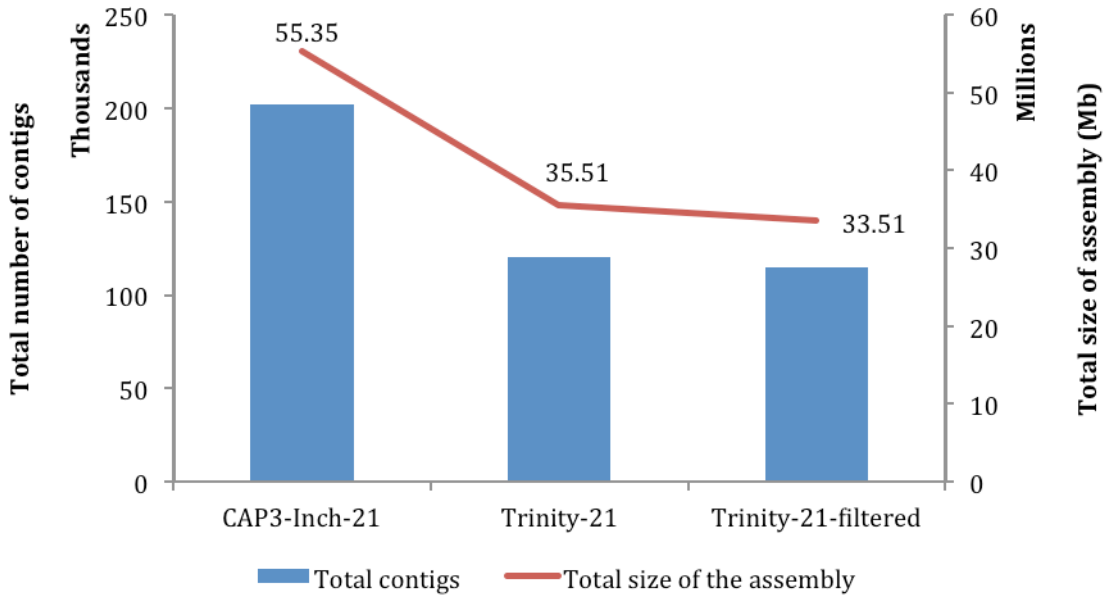


Figure 9: Comparison of total number of contigs and assembly size in various assemblies. CAP3-Inch-21: assembly generated by Inchworm with Kmer 21 using reads without filtering contamination from Univec database and merged by CAP3, Trinity-21: assembly generated by Trinity with Kmer 21 using reads without filtering contamination from Univec database, Trinity-21-filtered: assembly generated by Trinity with Kmer 21 using reads after filtering contamination from Univec database.

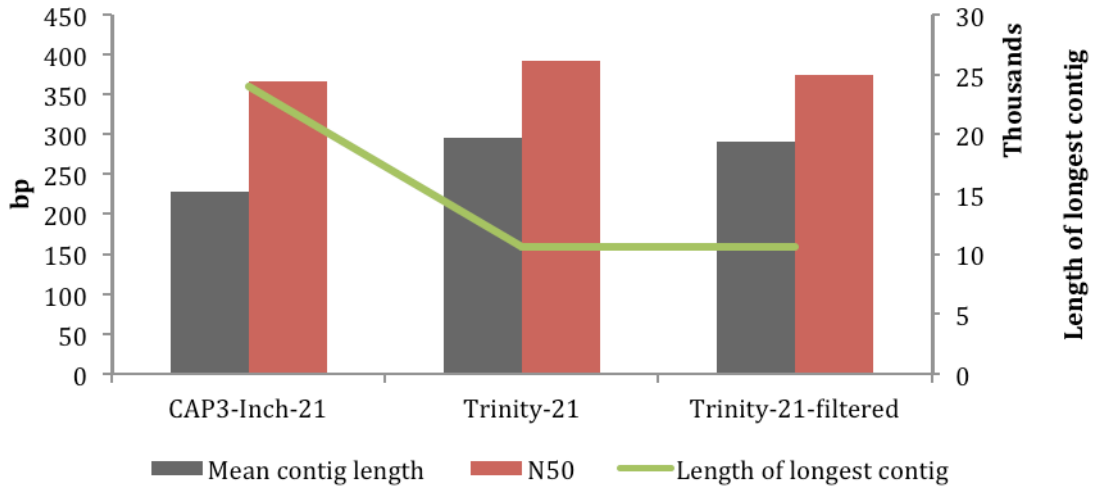
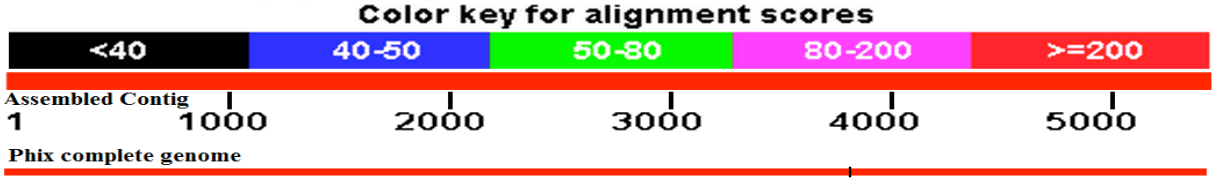


Figure 10: Comparison of mean contig length, N50 size and maximum contig size in various assemblies. CAP3-Inch-21: assembly generated by Inchworm with Kmer 21 using reads without filtering contamination from Univec database and merged by CAP3, Trinity-21: assembly generated by Trinity with Kmer 21 using reads without filtering contamination from Univec database, Trinity-21-filtered: assembly generated by Trinity with Kmer 21 using reads after filtering contamination from Univec database.

3.3. Contig validation

3.3.1. Alignment against phiX genome

The read dataset contained sequenced reads from phiX genome because the mRNA libraries sequenced were spiked with 1% phiX library as a control. By aligning raw reads to phiX genome, 0.7% of raw reads mapping to phiX genome could be extracted. Hence, one could expect that in case of a high quality assembly, all the reads from phiX would assemble into a single contig. Alignment of the contig set to phiX using blastn identified a single contig with full length alignment (100% of query coverage) to phiX with a significant E-score (0.0) (Figure 11), demonstrating that all phiX reads were correctly assembled into a single contig. This proved that the steps during the transcriptome assembly were correct and there was no mechanical failure during sequencing.



Accession	Description	Max score	Total score	Query coverage	E value	Max ident
28785	gi|262064129|gb|GQ153915.1| Enterobacteria phage phiX	6798	9866	100%	0.0	100%

Figure 11: Aligning contig set against the phiX complete genome identified a single contig with full-length hit to phiX

3.3.2. Alignment against available Atlantic herring complete protein coding sequences

All four full-length non-mitochondrial protein coding sequences reported to be expressed in muscles of Atlantic herring (source NCBI) had significant BLAST hits with the contigs of the transcriptome assembly (Table 5). Muscular cathepsin D mRNA (AF312364) produced a full-length alignment with a single contig, indicating that the transcript was perfectly assembled by Trinity. BLAST hits of TnI-alpha troponin-I mRNA (U20111) indicated that it had been assembled into two contigs. Fast and slow alpha actin mRNAs (GQ455648 and EF495203 respectively) had a common hit because they are the mRNAs from the alpha actin gene family and certain regions of these transcript sequences are shared between members of this gene family.

Table 5: Results of Blastn alignment of contigs set against Atlantic herring full-length mRNA expressed in skeletal muscle

Subject (Atlantic herring mRNA)	Length of subject (bp)	Query (Contig)	Length of query (bp)	Aligned co-ordinates of subject	% of subject aligned	ID %	E-score
AF312364	1191	comp2062_c0_seq1	1644	1 to 1191	100.00	99.58	0.0
U20111	740	comp11_c2_seq1	1360	7 to 536	71.63	86.23	1.7e-222
U20111	740	comp2563_c1_seq1	326	524 to 738	29.19	98.15	3.5e-108
GQ455648	1540	comp0_c0_seq1	1562	12 to 1526	98.38	99.60	0.0
EF495203	1421	comp0_c0_seq1	1562	70 to 1188	78.75	86.77	0.0

Alignment of the contig set against mitochondrial genes showed that all 13 mitochondrial transcripts had been assembled; six of them had full-length alignment whereas the remaining seven had more than 80% alignment (Table 6). The Trinity assembler could not assign unique contigs to each of these transcripts as these genes reside very close to each other in the mitochondrial DNA (mtDNA); hence they were assembled into four different contigs (Figure 12).

Table 6: Results of BLASTx alignment against Atlantic herring mitochondrial transcripts

Query (contig)	Query length (bp)	Subjects (Mitochondrial transcripts)	Subject length (bp)	% of subject aligned	ID %	E-score
comp196_c0_seq1	6620	YP_001293658.1	972	100.00	95.02	3.00E-127
		YP_001293659.1	1044	81.04	92.55	3.00E-20
comp29_c0_seq1	3363	YP_001293660.1	1598	100.00	96.94	3.00E-51
		YP_001293661.1	690	94.13	94.46	0-0
		YP_001293662.1	165	94.93	95.52	0.0
		YP_001293663.1	684	97.68	97.63	1.00E-71
comp26_c0_seq1	2985	YP_001293664.1	783	100.00	94.74	0.0
		YP_001293665.1	348	96.87	95.54	0.0
		YP_001293666.1	294	100.00	95.22	5.00E-113
		YP_001293667.1	1380	100.00	90.91	8.00E-24
comp26_c1_seq1	3743	YP_001293668.1	1833	98.69	96.44	4.00E-81
		YP_001293669.1	519	100.00	96.91	1.00E-137
		YP_001293670.1	1140	87.65	94.10	4.00E-121

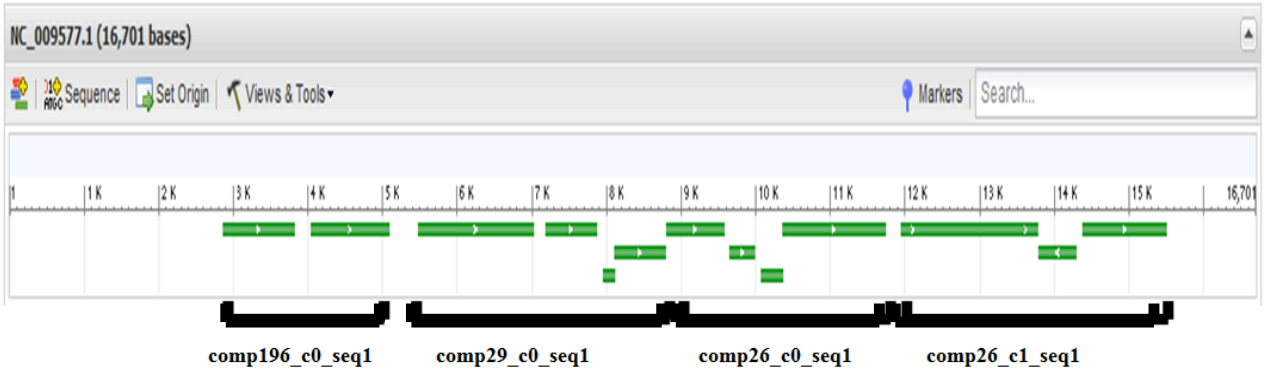


Figure 12: Distribution of genes in mtDNA in Atlantic herring and contigs from the assembly representing them (shown in black arrows). Note that multiple transcripts have been assembled into a single contig as the mitochondrial genes reside very close to each other)

3.4. Annotation

The assembled contigs were first aligned against NCBI non-redundant (NR) protein database using BLASTX with cut-off e-value of 10^{-4}. A total of 42,650 (37.08% of total) contigs had significant top BLAST hits with 23,696 unique records from the NR database. Among the contigs, those with longer assembled sequence had higher proportion of hits in the NR database (Figure 13). More than 90% of the contigs longer than 1500 bp and more than 85% of contigs ranging from 1000 to 1500 bp had matches to the NR database. Only 31.89% of total contigs less than 500 bp had significant NR hits. This result verifies the fact that the length of query sequence is crucial to determine the level of significance of a BLAST hit.

According to BLAST algorithm, shorter query sequences must possess higher similarity to subject in order to satisfy a certain E-value (Franchini et al, 2011).

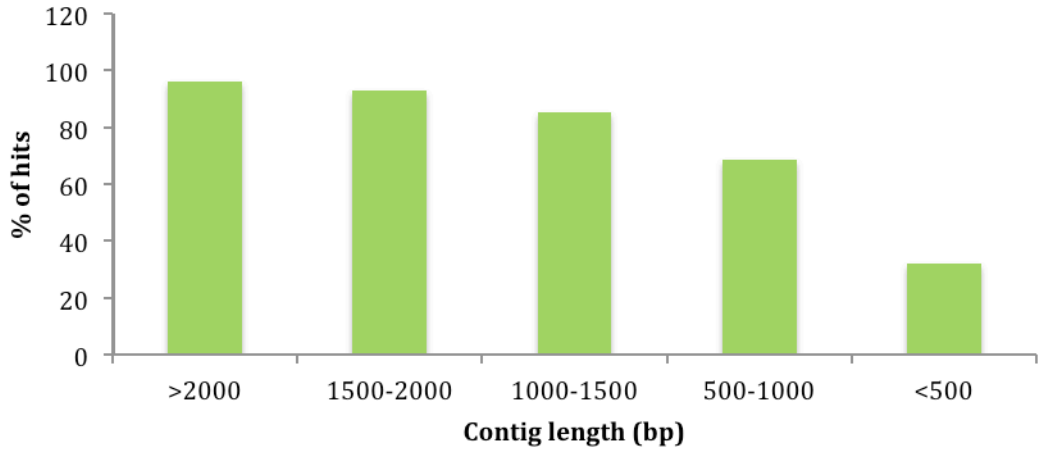


Figure 13: Effect of contig length on the proportion of hits to the NR database

The e-value distribution of the hits to the NR database showed that 13.21% of aligned sequences had a strong homology (e-value < 10^{-50}) whereas the remaining 86.79% had homology with an e-value in the range of 10^{-4} to 10^{-50} (Figure 14).

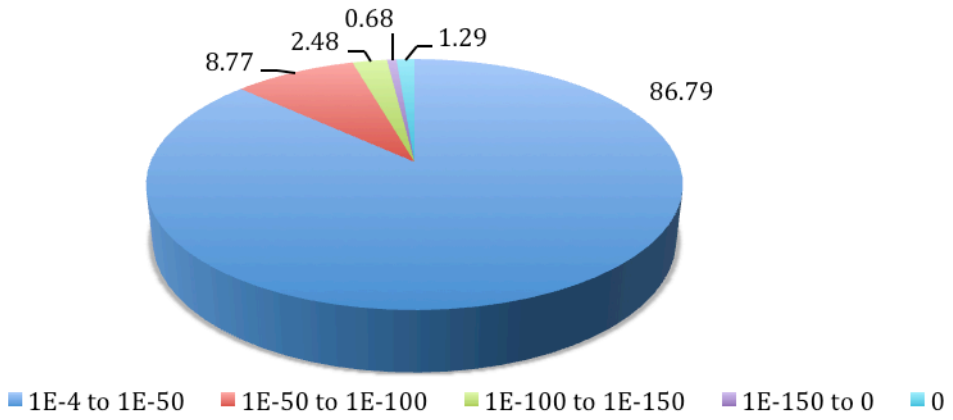


Figure 14: Distribution of E-value of BLAST hits

11.40% of the query contig sequences had a sequence identity of more than 80% to NR protein sequences whereas remaining 88.60% of the hits had sequence identity ranging from 13 to 80% (Figure 15).

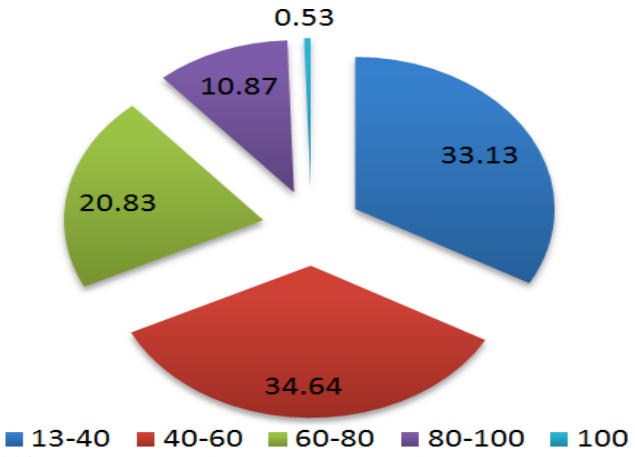


Figure 15: Distribution of % sequence identity between Atlantic herring contigs and NR proteins

The calculation of alignment coverage of BLAST hits showed that on an average, 75.85% of the length of the contig was aligned to NR protein sequences. More than 87% of these contigs had aligned more than 40% of their length and 56.78% had more than 80% (Figure 16).

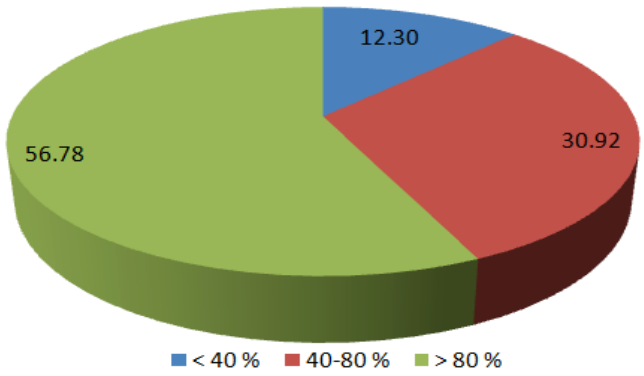


Figure 16: Distribution of the proportion of length of contigs aligned to NR proteins

Furthermore, we did an analysis of species distribution of BLAST hits that showed more than 75% of contigs had top matches (best hit chosen on the basis of lowest e-value) to annotated fish proteins in NR database (Figure 17) and the remaining hits to mammals, birds and other vertebrates. There were few hits to distantly related species and certain invertebrates that demonstrated the existence of few evolutionary conserved proteins within the assembled skeletal muscle transcriptome in Atlantic herring.

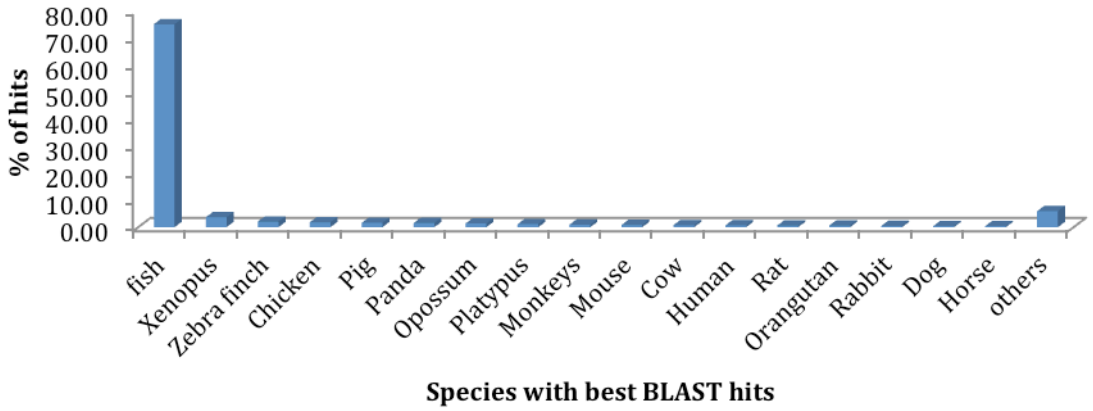


Figure 17: Species-wise distribution of NR protein hits

Among the hits to fish proteins, more than 50% of hits are against Zebrafish proteins, followed by several salmons species (18.44%) and tetraodon (16.85%) (Figure 18). The contigs also had hits to a few Atlantic herring proteins that were previously annotated in the NR database.

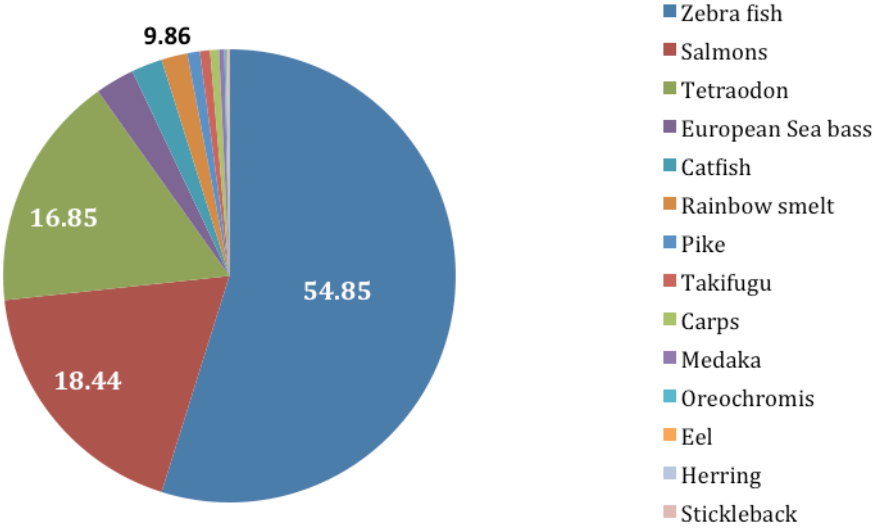
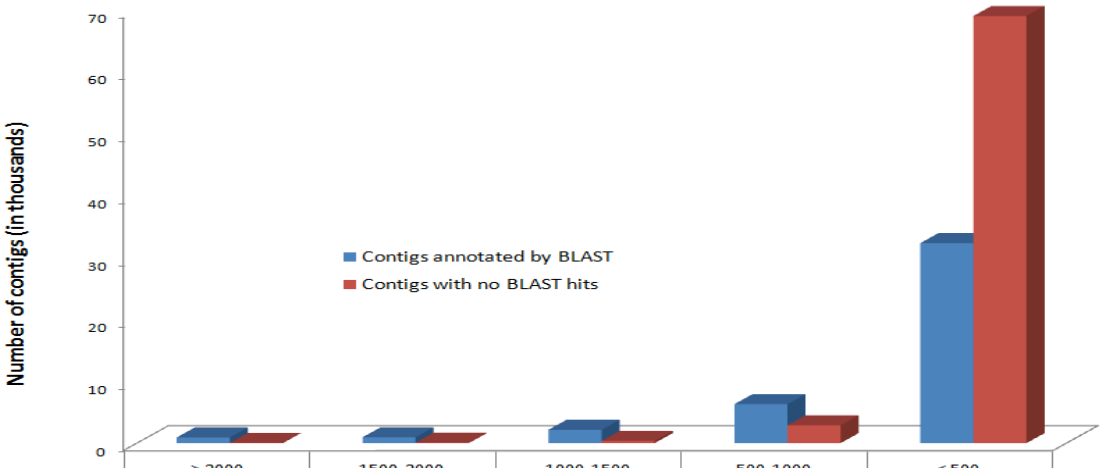


Figure 18: Species-wise distribution of fish protein hits

Comparison to zebrafish unigenes

All 72,396 (62.93% of total) contigs that did not have any hits to NR database were extracted and compared to Zebrafish Unigene records in an attempt to annotate contigs homologous to zebrafish transcripts, which are in the Unigene records but still not annotated in the NR database. Using TBLASTX, we found that 4,329 contigs (out of 72,396 contigs) had matches to 3,121 unique entries belonging to Zebrafish in Unigene database.

After merging results from the blast searches against NR and the zebrafish records in Unigene database, we could annotate a total of 46,979 contigs, which represented 40.84% of total transcriptome assembly. The remaining 59.16% of the assembly did not have any significant matches to annotated genes. A comparison of the size distribution of contigs with and without blast hits showed that the majority of contigs that do not have blast hits are small (less than 500 bp) (Figure 19).



	> 2000	1500-2000	1000-1500	500-1000	< 500
Contigs annotated by BLAST	916	952	2163	6297	32322
Contigs with no BLAST hits	36	72	370	2884	69034

Figure 19: Size distribution of contigs with and without BLAST hits

Further, we divided our original assembly into two sets; contigs with blast hits (annotated assembly) or without blast hits (un-annotated assembly). We obtained better assembly statistics for the annotated assembly compared to the un-annotated one (high N50 size and average contig length along with significantly high coverage) (Table 7). We also compared

these assemblies to total zebrafish transcripts reported in Unigene database (Table 7). The mean contig size and N50 of our annotated assembly was less in comparison to zebrafish unigenes, which demonstrated the presence of large number of small contigs in our assembly. As our contigs only covered skeletal muscle transcriptome of Atlantic herring, the total sum of our assembly was also less than zebrafish unigenes (which includes all the transcripts expressed in every tissues in zebrafish).

Table 7: Assembly statistics of original, annotated* and un-annotated assembly and comparison to zebrafish transcripts reported in Unigene database*****

Assembly	Total contigs	Max. Contig Size (bp)	Total sum of assembly (Mbp)	Average coverage	Mean contig size (bp)	N50 (bp)
Original assembly	115,046	10,568	33.56	151.43X	291	375
Annotated assembly	46,979	10,568	19.52	259.38X	669	669
Un-annotated assembly	68,067	3,270	13.99	23.62X	210	210
Zebrafish Unigene assembly	52,653	87,604	59.83	N/A	1,136	1,509

***Contigs with blast hit to NR protein and zebrafish unigenes database, ** Contigs with no blast hit to NR protein and zebrafish unigenes database, ***total annotated transcripts in zebrafish reported in Unigene records**

3.5. Transcriptome redundancy

The presence and degree of redundancy in the transcriptome assembly could help to explain the level of alternative splicing events (AS) (Coppe et al, 2010). We limited the investigation of transcriptome redundancy to only those contigs from our putative annotated assembly (Table 7). We considered that these contigs would be more reliable for any "biological" analysis compared to the original assembly. A total of 46,979 contigs were annotated with 26,817 entries (NR proteins and Zebrafish unigenes) and among these entries, 18,333 were represented by single contig each. If we consider that each of the entries corresponds to a different gene, these 18,333 (out of 46,979) contigs represented transcripts of different genes. The 8,484 remaining entries had hits to more than one contig (3.37 contigs per NR entry, on average), which was higher than the estimated alternative splicing events per gene in several teleost fish species (Table 8). The fraction of NR entries that were represented by at least two contigs was 31.64% (8,484/26,817), which was remarkably higher than the percentage of

unique alternative spliced genes in zebrafish but similar to other teleost fish (medaka, fugu and stickleback) (Table 8).

Table 8: Statistics of alternative splicing events in various Teleost fish and comparison in Atlantic herring

	Atlantic Herring*	Zebrafish**	Medaka**	Fugu**	Stickleback**
Unique AS genes (% of total genes)	31.64	17.0	31.2	43.2	32.4
AS events/gene	3.37	1.74	1.67	2.21	1.65

*Calculation of putative alternative splicing events in Atlantic herring was based on Transcriptome redundancy analysis, which may only partially explain true alternative splicing events
**Source: Liu et al., 2010

3.6. G-protein-coupled-receptor (GPCR) genes expressed in skeletal muscles of Atlantic herring

Using the strategy shown in Figure 4, 127 contigs from the annotated assembly were identified that had homology with 60 unique entries in GPCR database. Among the four families of GPCR super-family, 28 proteins from class A rhodopsin-like family, 27 from class B secretin-like family and 5 from class C Metabotropic glutamate/pheromone family were identified (Table 9). There were no hits to annotated vemeronasal receptors proteins listed in the GPCR database.

Table 9: GPCR homologs identified in skeletal muscle of Atlantic herring

GPCR family	Number of matches from GPCRDB	Number of homologous contigs in Atlantic herring
Class A Rhodopsin like	28	56
Class B Secretin like family	27	47
Class C Metabotropic glutamate/pheromone family	5	24

Among these 60 hits, only 18 were against GPCRs from fish (eight in Zebrafish and ten in Tetraodon) (Table 10 & 11). However, only eight GPCRs entries among these hits had functional annotation in the GPCR database. We manually checked these eight GPCR proteins but surprisingly found that three of these proteins were actually non-GPCR proteins but were still included in GPCR database.

Table 10: Hits to fish rhodopsin-like GPCRs

Contigs	Contig size	Homologous GPCR protein hits	Species
comp30718_c0_seq1	822	Slit2	Zebrafish
comp24036_c1_seq1	910	Sphingosine 1-phosphate receptor 2	Zebrafish
comp26407_c1_seq1	1146	Cxcr7b protein	Zebrafish
comp13621_c0_seq1	1397	PREDICTED: Leucine-rich repeat-containing G protein-coupledreceptor 4 precursor-like	Zebrafish
comp31703_c0_seq1	768	Slit1b	Zebrafish
comp27908_c1_seq1	777	Slit (Drosophila) homolog 3	Zebrafish
comp20545_c0_seq1	761	Uncharacterized	Tetraodon
comp21148_c0_seq1	764	Uncharacterized	Tetraodon
comp12410_c0_seq1	827	Uncharacterized	Tetraodon
comp14272_c1_seq1	849	Uncharacterized	Tetraodon
comp20630_c0_seq1	894	Uncharacterized	Tetraodon
comp6782_c0_seq1	903	Uncharacterized	Tetraodon
comp9733_c2_seq1	1108	Uncharacterized	Tetraodon
comp1210_c0_seq1	1184	Uncharacterized	Tetraodon
comp4596_c0_seq1	1618	Uncharacterized	Tetraodon

Table 11: Hits to fish secretin-like GPCRs

Contigs	Contig size (bp)	Homologous GPCR protein hits	Species
comp10440_c0_seq1	1121	PREDICTED: CD97 antigen-like	Danio rerio
comp11045_c0_seq1	872	PREDICTED: cadherin EGF LAG seven-pass G-type receptor 1a	Daniorerio
comp18836_c0_seq1	778	Uncharacterized	tetraodon

3.7. Identification of putative allelic variants (SNPs/Indels)

As in "Transcriptome redundancy" and "GPCRs identification analysis", we only considered annotated assembly (Table 7) for performing variant calling. We considered the following specific parameters to be observed in any position where a variant would be called in order to assure a quality call: minimum read depth of 8, minimum base quality at a position to count a read of 15 and a p-value threshold of 0.05. A total of 149,396 raw variants were called; among these raw variants, only those with variant frequency in the range of 20 to 60% were selected and kept as putative variants. As a result after filtering, a total of 25,431 putative polymorphic positions (24,351 SNPs and 1080 indels) were identified with the average SNP density of 1.247 SNPs per Kb. Almost two-thirds of these putative SNPs were transitions (Figure 20).

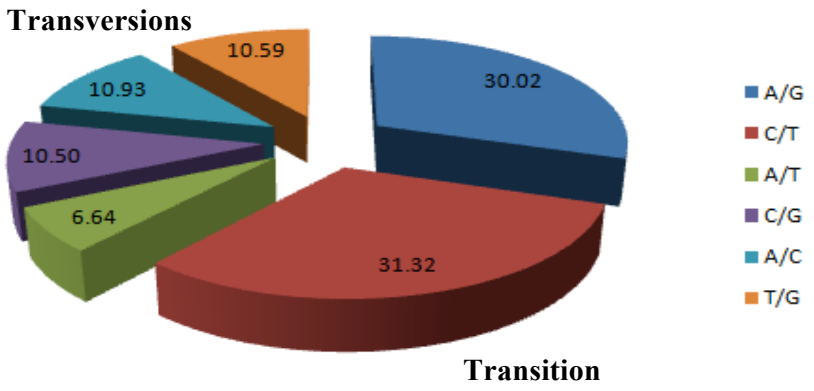


Figure 20: Proportion of transitions (A ←→G or C ←→ T changes) and transversions (A ←→ T, C ←→ G, A ←→ C and T ←→ G changes) identified in the putative SNP set

The distribution of variant frequencies showed that the majority of putative SNPs had variant frequency in the range of 30 to 50% with the average variant frequency of 0.397 (Figure 21).

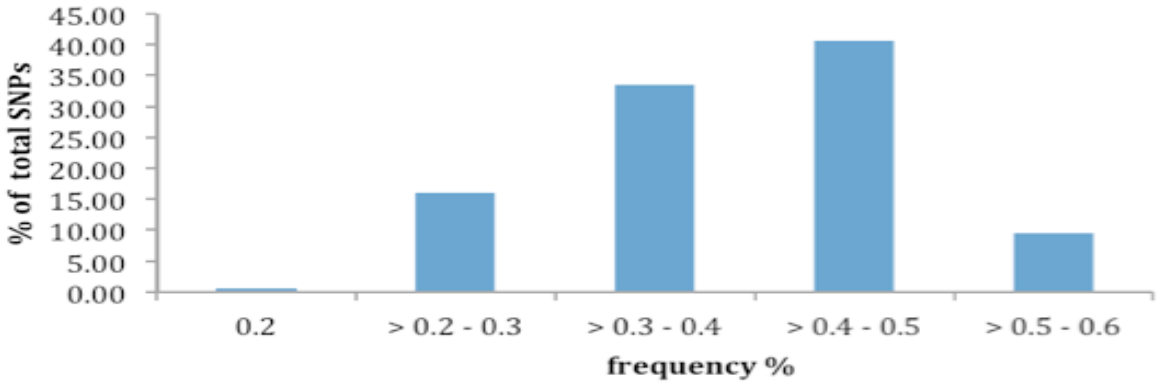


Figure 21: Distribution of variant frequency of putative SNPs in the Atlantic herring transcriptome from a single fish

A total of 8,719 annotated contigs were found to contain putative SNPs. Of these contigs, 47.95% had only one SNP and 88.70% had five or fewer SNPs per contig (Figure 22).

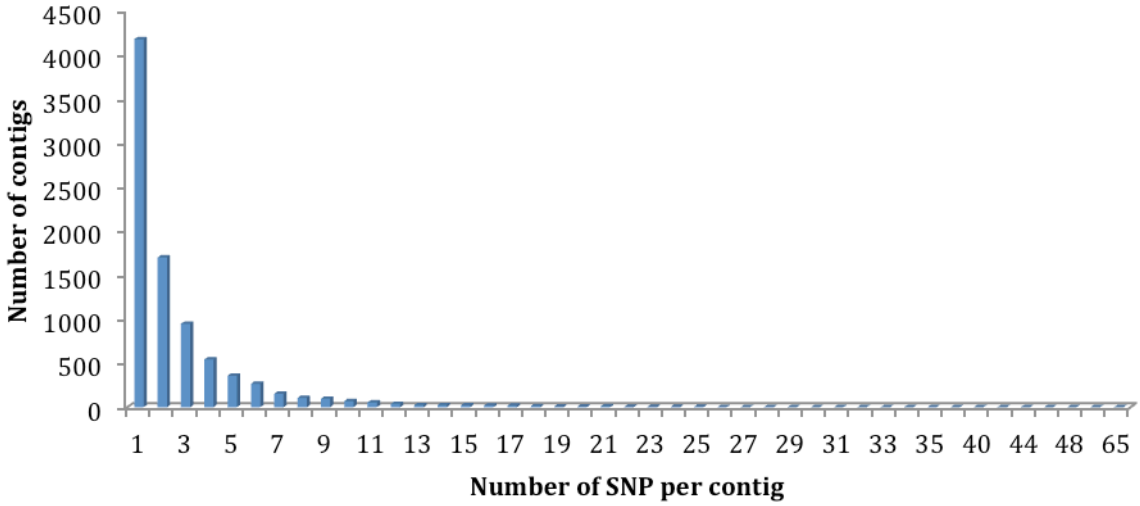


Figure 22: Distribution of putative SNPs per contig (The histogram shows the number of contigs by the total number of SNPs identified in each of these contigs)

In order to check the quality of our putative SNPs, we looked into the SNPs in a contig that was homologous to zebrafish Glucose-phosphate isomerase (GPI) gene. Andersson et al., (1981) had identified five alleles within this locus using starch gel electrophoresis in Atlantic herring. We could detect that two SNPs placed within the open reading frame (ORF) of the GPI-homologue contig were called with a significant p-value with high coverage of reads (Table 12). The sequenced specimen seemed to be heterozygous for both the alleles (variant allele frequency ~ 50%).

Table 12: Statistics of polymorphic positions detected in ORF of contig homologous to zebrafish GPI gene*

Position	Reference base	Variant base	Varfreq%	Reads1	Reads2	p-value	Type of mutation
1416	A	G	47.98	3747	3436	0	Synonymous (F→ F)
1751	C	T	46.88	3528	3115	0	Non-synonymous (N → D)

*Position: polymorphic position in the contig, Varfreq%: frequency of variant allele, Reads1: coverage of reads for reference base, Reads2: coverage of reads for variant base, p score: significance of variant calling.

4. DISCUSSION

The characterization of the transcriptome in a non-model organism that lacks genomic sequence information using ultra-high throughput mRNA sequencing technology is a fast, efficient and a cost effective approach. With such next generation sequencing technologies, *de novo* transcriptome assemblies of various aquatic species have been established in the recent years: pout fish (*Zoarces viviparous*) (Kristiansson et al., 2009), cichlid fish *(Amphilophus zaliosus*) (Elmer et al., 2010), african abalone (*Haliotis midae*) (Franchini et al., 2010), european eel (*Anguilla anguilla*) (Coppe et al., 2010), catfish (*Ictalurus sps*) (Liu et al., 2011), guppy (*Poecilia reticulata*) (Fraser et al., 2011) e.t.c.

In this study, we were able to establish a draft muscle transcriptome of an economically important marine fish, Atlantic herring; using the Illumina HiSeq-2000 platform. A cDNA library from skeletal muscle tissue from a single specimen of Atlantic herring was sequenced to generate pair end reads (two reads with 101 bp each). We obtained 116.06 million raw reads that were filtered for quality and contamination, thus generating 97.40 million "high quality" reads.

There is a need to adopt different strategies to improve the accuracy of a *de novo* assembly. The transcriptome assembly of non-model organisms is still a challenging task despite the rapid development of assemblers in recent time (Haas and Zody, 2010). The quality of a transcriptome assembly is highly dependent upon Kmer size (the user defined length of sequence which overlaps between reads and it is used to assemble and extend contiguous nodes of *de bruijn* graph). The choice of longer Kmer results in better assembly of highly expressed transcripts and poorly expressed transcripts are better assembled with low Kmer (Surget-Groba et al, 2010). In most of the studies, an optimal Kmer size that is a compromise between these two extremes is chosen. We tested the performance of various Kmer lengths

taking the advantage of Inchworm RNA-Seq assembler that could generate different assemblies through a range of Kmer sizes (19 to 29). The assembly with 21-mer (Inch-21 assembly) had the best assembly statistics (lowest number of contigs (247,776), highest N50 (302 bp) and mean contig size (251 bp)); hence a 21-mer was estimated to be an optimal Kmer for the assembly of Atlantic herring transcriptome.

The Inch-21 assembly still had a larger number of transcripts than expected in Atlantic herring, which may be partially explained by inappropriate merging of read sequences for the same transcript into a unique contig. Hence, the Inch-21 assembly was further reassembled using CAP3 assembler merging 52,088 contigs into 6,366 clusters and generating a new assembly (Inch-21-CAP3) of 202,054 contigs. There was an improvement in N50 (366 bp) and mean contig size (273 bp) but still the total number of contigs was higher than expected.

Other important factors that affect the quality of a transcriptome assembly may include non-uniform transcriptome coverage due to variation in gene expression level; repeats, allelic variants and alternatively spliced transcripts that impedes contig elongation and erroneous fusion of reads to assemble chimeric transcripts (Grabherr et al., 2011). We identified a number of contigs from the Inch-21 assembly (results not shown) that had hits to different RefSeq genes at various positions in a single contig, possibly due to assembly of chimeric contigs. The generation of high number of contigs and evidence of chimeras hinted us that Inchworm RNA-Seq assembler was not an efficient assembler for a high coverage RNA-Seq data.

During the course of this study, a new RNA-seq assembler package called “Trinity” was released which added new software modules to the existing Inchworm assembler and expected to resolve the issues dealing with chimeric transcripts resolution and report unique

contigs for alternatively spliced isoforms and paralogous genes (Grabherr et al., 2011). The results of the Trinity assembly in whitefly (insect lacking a reference genome) transcriptome showed that full-length reconstruction of a large fraction of transcripts across a broad range of expression level and recognition of alternative splice isoforms and transcripts from recently duplicated genes was possible (Grabherr et al., 2011). The performance of "Trinity" thus motivated us to use this tool for the assembly of the Atlantic herring transcriptome.

We generated a total of 115,046 contigs (≥ 100 bp) from the "Trinity" assembly, which showed clear improvements in assembly statistics (total contigs assembled, total assembly size, N50 and mean contig size, average coverage) in comparison to the Inchworm assembly (see Table 4). The N50 (375 bp) and mean contig size (291 bp) of the assembly in our study was higher in comparison to other studies using similar Illumina RNA-Seq technology and assembly parameters (South-African abalone transcriptome: N50-356 bp, mean contig size-260 bp, assembled with "CLC Genomics Workbench v4.0 tool" (Franchini et al., 2011); Tea transcriptome: N50-225 bp, mean contig size-208 bp, assembled with SOAP de novo tool (version 1.03) (Shi et al., 2011)). The average read depth of the assembly was high (151.43x) and the distribution per contig (measured as FPKM score) had a broad range across the assembly (see figure 7). We expected this behaviour because variation in expression levels of different transcripts in a particular tissue causes wide variation in the coverage of contigs assembled.

After assembly, we utilized bioinformatics methods to validate the transcriptome assembly. First, we aligned our contig sets against the phiX genome, which was spiked in the samples as a control for the sequencing instrument. The blast result showed that all the reads from phiX were assembled into a single contig that had 100% similarity and full-length alignment to phiX genome (see figure 11). We also matched our contigs against mitochondrial transcripts

and four full-length mRNAs expressed in the skeletal muscle of Atlantic herring that had been previously reported in the NCBI database. The results showed that all of these transcripts were contained in the assembly, the majority of them being reconstructed in full-length. This validation analyses demonstrated that the assembly steps worked correctly.

The contigs from the assembly were further compared to the NR protein database for annotation. This comparison returned 42,650 contigs (37.08% of total assembled contigs) with significant hits to 23,696 unique NR proteins. The majority of the long contigs (>1500 bp) had hits whereas most of the small contigs (<500 bp) had no hits showing high reliability of the long assembled contigs in our assembly. The homologous sequence in the contigs covered 75.85% of the total contig size on average; this showed that most parts of the transcripts had been assembled. More than 75% of the top-blast matches were against fish proteins (50% among those matches specifically against zebrafish), which demonstrated the fact that Atlantic herring is evolutionary closer to zebrafish than to other fish whose protein annotations are available.

The NR database only contains annotated protein sequences; therefore aligning contigs only against NR proteins might miss some of the homologous sequence, which is yet to be annotated. Hence, we compared the remaining contigs with no hits to NR database to zebrafish unigenes records (contains all mRNA and ESTs reported to date in zebrafish), which identified additional 4,329 contigs with significant hits to 3,121 Zebrafish unigenes. Thus, by comparing contigs to annotated sequences from various species, 46,979 contigs (41% of the entire transcriptome) were annotated. The annotated contigs had better assembly statistics compared to the original assembly. They presented a higher N50 size (669 bp) and mean contig size (415 bp) along with significantly high read depth (259.38 x). The number of annotated contigs was close to the total number of transcripts reported in zebrafish (51,569,

Ensembl, release 62). The remaining 68,067 contigs (59.16% of total assembly) did not have significant homology to any existing genes.

The percentage of contigs matching to known sequences was low (41%) but was comparable to results from recent studies where Illumina sequencing technology was used for *de novo* transcriptome assembly of non model species: 16.2% in whitefly transcriptome; Wang et al, 2010), 32.6% in tea transcriptome; Shi et al., 2011, 30.9% in catfish transcriptome; Liu et al., 2011 and 16.8% in South African Abalone transcriptome; Franchini et al., 2011. The common explanation for the low percentage of hits to sequence database in each of these studies was the lack of genomic information in these non-model species. But, we believe that lack of genomic information only partially explains the low percentage of hits to sequence database. For example, a large proportion (~88%) of the contigs in the assembly were small (less than 500 bp), some of them may be too short to produce statistically significant blast alignments. Although a strict laboratory protocol was followed for polyA selection to sequence only mRNAs from tissues, high abundance of other forms of RNAs (e.g. rRNA, tRNA and microRNAs) in the cell may allow sequencing of considerable amount of these non coding RNAs, resulting in the assembly of a significant number of small sized contigs that would obviously have no hits to homologous protein coding transcripts. The average read depth of these un-annotated contigs was lower (23.62x) compared to the annotated assembly (259.38x), which also hinted the possibility of spurious assembly of some of these sequences that do not have any biological relevance.

We also analysed the nature of these hits with entries from NR database and zebrafish unigenes in an attempt to evaluate the level of transcriptome redundancy. Transcriptome redundancy (existence of contigs sharing similar sequences) is an expected event in a transcriptome assembly, which is either due to the assembler or because of the existence of

certain biological phenomenon (e.g. alternative splicing and gene duplication). RNA-seq assemblers may not be able to assemble all sequences belonging to same transcript into a single contig because of the assembly artifacts and sequencing errors, resulting in redundant contigs. Allelic and splice isoforms as well as transcripts from gene duplicates also share certain specific sequence that may result in redundancy (Coppe et al, 2010). 31.64% of the total significant entries were aligned to more than one contigs (3.37 contigs hits per protein), which was comparable to results of European eel transcriptome assembly (29% of entries represented by at least 2 contigs, 3.3 per entry on average, Coppe et al., 2010). But this level of redundancy was higher than the number of genes with alternative splicing events in zebrafish (Liu et al., 2010) (see table 8). This result demonstrated that the amount of transcriptome redundancy in the assembly was higher than the expected redundancy due to biological events (e.g. splice isoforms and gene duplicates) and this level of redundancy in the present annotated assembly can only partially explain such events.

GPCRs are one of the largest super-families of membrane protein, which make up almost 1-2% of vertebrate genome and are characterized by highly conserved seven trans-membrane (TM) hydrophobic regions (Metpally and Sowdhamini, 2005). Fish also have this super-family, but the number of GPCR genes reported in fish is substantially lower compared to mammals (Niimura and Nei, 2005), most probably due to the lack of olfactory receptors in fish. Investigation of putative GPCRs homologous contigs in the transcriptome assembly identified 127 unique contigs that had hits to 60 unique GPCRs from different species listed in the GPCR database. A study by Metpally and Sowdhamini, 2005 had identified 466 GPCRs in *Tetraodon nigroviridis* genome. However comparison to this number will be biased, as we do not expect all of the GPCRs to be expressed in the skeletal muscle. The number of unique GPCRs (60) identified in our study is comparable to the study by Jean-Baptiste et al., (2005) who reported 42 unique GPCRs expressed in skeletal muscles of

different mammals (human, dog, monkey, mouse, rat and rabbit). However, we discovered that the GPCR database had included certain non-GPCR sequences. With closer inspection of our results (data not shown) several of the protein sequence from GPCR database that had hits to our contigs did not correspond to GPCRs. Hence, the results of the estimation of GPCRs based on the information of GPCR database in our study was not reliable. More detailed studies with an alternative reliable approach will be needed to estimate the true number of GPCRs expressed in muscle tissue from Atlantic herring.

We also scanned the frequency of allelic variations in the assembled transcripts. Since no reference genome was available for this species, we mapped the reads to the contigs sequences as a reference for variant calling. We discarded the contigs that did not receive any BLAST annotation for this analysis, as majority of them were not reliable as true protein coding transcripts. A total of 25,431 putative allelic variants (24,351 SNPs and 1080 indels) were identified, with an average SNP density of 1.247 SNPs per Kb. Two SNPs were identified in a contig that was homologous to the Glucose-Phosphate Isomerase (GPI) gene. In a previous study, five alleles were identified in the Atlantic herring GPI gene (Andersson et al, 1981). One of the non-synonymous SNP identified in our study that changes the charge of the protein might explain part of the variation previously detected at the GPI locus in Atlantic herring using starch gel electrophoresis. The variant frequency at both polymorphic sites was ~50%, hence the Atlantic herring specimen sequenced seemed to be heterozygous for these two sites.

5. CONCLUSION

The present study described the characterization of skeletal muscle transcriptome of Atlantic herring *(Clupea harengus)*, one of the most abundant and economically important marine fish species. Using paired end reads generated by Illumina sequencing technology from skeletal

muscle of a single specimen, we assembled 115,046 contigs (> 100bp) using Trinity RNA-seq assembler with 46,979 contigs (~42% of total assembly) annotated to 26,817 known genes. The results demonstrated that ultra-high throughput RNA-Seq technology in combination with downstream bioinformatics strategies could be applied as a fast and cost efficient approach for *de novo* transcriptome assembly of a non-model organism without prior genomic information. The annotated contigs were used to identify GPCRs expressed in the skeletal muscles of Atlantic herring. These contigs also allowed us to detect polymorphic positions in the coding sequences of reference individual that may be useful for the further genetic studies. As this study is a part of Atlantic herring genome sequencing project, the draft transcriptome assembly generated in this study would certainly be a useful resource for the validation of genome assembly of Atlantic herring that is currently being conducted within the group.

ACKNOWLEDGEMENTS

I would like to express sincere gratitude to my supervisors professor leif andersson, dr. Alvaro martinez barrio and dr. Carl-johan rubin for providing me the unique opportunity to work in this interesting project within their group. It is all due to their encouragement, guidance and support that helped this thesis become a reality. A word of special appreciation also goes to dr. Görel sundstrom who was always ready to provide expert opinions about the "biological" interpretation on the "bioinformatics" analysis of our datasets. I am also thankful to all the personnels at scilifelab, stockholm who generated the sequence reads and made it available to us for downstream analysis. I want to thank dr. Manfred grabherr, the developer of "trinity" rna-seq assembler, who provided valuable advise on the various issues of rna-seq assembly using inchworm and trinity assembler. I would also like to thank nima rafati; my colleague who was working in the same project, whose collaboration made the job much easier. I also wish to express my sincere thanks to all other members of genome-seq group who

participated in the project meetings and provided important inputs for the study. Last but not least, i would like to thank my wife for her patience, support and love throughout my life.

APPENDIX

Appendix 1: Running FASTX tools

The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files pre-processing.

A. Generating FASTQ reads quality statistics
$ fastx_quality_stats –i raw_reads.fasta –o quality_stats.txt

B. Generating FASTQ Quality chart
$ fastq_quality_boxplot_graph.sh -i quality_stats.txt -t quality score distribution

C. Generating Nucleotide distribution chart
$ fastx_nucleotide_distribution_graph.sh -i quality_stats.txt -t Nucleotide distribution

Appendix 2: Running Seqclean tool

A script for automated trimming and validation of ESTs or other DNA sequences by screening for various contaminants, low quality and low-complexity sequences

$ seqclean seqfile raw_reads.fasta -v Univec.fasta -r cleaning-report.cln -o cleaned-reads.fasta -l 64

Appendix 3: Running Inchworm RNA-seq assembler for de-novo assembly

A. Extracting FASTA sequences from FASTQ files

$ Inchworm-03132011/**util/**fastQ_to_fastA.pl -I left.fq > left-reads.fa

$ Inchworm-03132011/**util/**fastQ_to_fastA.pl -I right.fq > right-reads.fa

$ Cat left-reads.fa right-reads.fa > both-reads.fa

B. Running Inchworm

$ Inchworm --reads both-reads.fa --run_inchworm --DS -K 21 -L 100

Appendix 4: Running CAP3 for merging assembly

$cap3 Inch-21.fasta

Appendix 5: Running Trinity RNA-seq assembler

$ Trinity.pl --seqType fa --left reads-forward.fasta –right reads-reverse.fasta --output Trinity-output --run_butterfly -- min_contig_length 100 --paired_fragment_length 240

Appendix 6: Running blastall package

blastn search nucleotide databases using a nucleotide query
$ blastall –p blastn –d formatted-database.fasta -i query.fasta -o blastn-output -m 8 -e 1e-5 -a 8 -W 11

blastx Search **protein** database using a **translated nucleotide** query
$ blastall –p blastx –d formatted-database.fasta -i query.fasta -o blastx-output -m 8 -e 1e-5 -a 8 -W 3

tblastx Search **translated nucleotide** database using a **translated nucleotide** query
$ blastall –p tblastx –d formatted-database.fasta -i query.fasta -o tblastx-output -m 8 -e 1e-5 -a 8 -W 3

Appendix 7: Mapping reads to contigs for variant calling

A. MosaikAligner
$ MosaikBuid –fr reads-forward.fasta –fr2 reads-reverse.fasta –out formatted-reads.dat
$ MosaikBuild –fr Trinity-assembly.fasta –oa Trinity-assembly.dat
$ MosaikAligner –in formatted-reads.dat –out aligned-reads.dat –ia Trinity-assembly.dat –mm 4 –p 8
$ MosaikText –in aligned-reads.dat –bam aligned-reads.bam

B. Samtools
$ samtools sort –on aligned-reads.bam aligned-reads-sort
$ samtools pileup –f Trinity-assembly.fasta aligned-reads-sort.bam > aligned-reads-sort.pileup
C. varscan
$ VarScan pileup2cns aligned-reads-sort.pileup --min-coverage 8 --min-reads 2 --min-avg-qual 15 --min-var-freq 0.2 --p-value 0.05

REFERENCES

Adams, M.D., Kelley, J.M., Gocayne, J.D., *et al.* (1991) Complementary DNA Sequencing: Expressed Sequence Tags and human. *Genome Project. Science*, 252: 5013, 1651-1656.

Andersson, A., Ryman, N., Rosenburg, R., Stahl, G. (1981) Genetic variability in Atlantic herring (*Clupea harengus harengus*): description of protein loci and population data. *Hereditas*, Volume 95, Issue 1, pages 69–78.

Blekhman, R., Marioni, J.C., Zumbo, P., Stephens, M., and Gilad, Y. (2010) Sex-specific and lineage-specific alternative splicing in primates," *Genome Research*, 20: 2, 180–189.

Casneuf, T., Van de Peer, Y., Huber, W. (2007) In situ analysis of cross-hybridization on microarrays and the inference of expression correlation. *BMC Bioinformatics*, 8:461.

Cirulli, E.T., Singh, A., Shianna, K.V., et al. (2010) Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biology*, 11(5).

Coppe, A., Pujolar, J.M., Maes, G.E. *et al*. (2010) Sequencing, de novo annotation and analysis of the first Anguilla anguilla transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. *BMC Genomics*, 11:635

Costa, V., Angelini, C., De Feis, I., Ciccodicola, A. (2010) Uncovering the Complexity of Transcriptomes with RNA-Seq. *Journal of Biomedicine and Biotechnology*, doi:10.1155/2010/853916.

Elmer, K.R., Fan, S., Gutner, H.M., Jones, J.C., Boekhoff, S., Kuraku, S., Meyer, A. (2010) Rapid evolution and selection inferred from transcriptomes of sympatric crater lake cichlid fishes. *Mol Ecol*, 19 (Suppl. 1): 197-211.

Franchini, P., Van der Merwe, M., Roodt-Wilding, R. (2011) Transcriptome characterization of the South African abalone Haliotis midae using sequencing-by-synthesis. *BMC Research Notes*, 4:59.

Fraser, B.A, Weadick, C.J. Janowitz, I. *et al*. (2011) Sequencing and characterization of the guppy (Poecilia reticulata) transcriptome, *BMC Genomics*, 12:202.

GAGGIOTTI, O.E., BEKKEVOLD, D., JØRGENSEN, H.B., FOLL, M., CARVALHO, G.R., ANDRE, C., RUZZANTE D.E. (2009) DISENTANGLING THE EFFECTS OF EVOLUTIONARY, DEMOGRAPHIC, AND ENVIRONMENTAL FACTORS INFLUENCING GENETIC STRUCTURE OF NATURAL POPULATIONS: ATLANTIC HERRING AS A CASE STUDY. *EVOLUTION*, 63(11):2939-2951.

Hauser, L., Turan, C., Carvalho, G.R. (2001) Haplotype frequency distribution and discriminatory power of two mtDNA fragments in a marine pelagic teleost (Atlantic herring, *Clupea harengus*). *Heredity*, 87, 621–630.

Haas, B.J. & Zody, M.C. (2010) Advancing RNA-Seq analysis. *Nature biotechnology*, 28:5

Hegedus, Z., Zakrzewska, A., Ágoston, V.C., *et al.* (2009) Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. *Molecular Immunology*, 46 : 2918–2930.

Ga¨el Jean-Baptiste, G., Zhao Yang, Z., Khoury, C. et al. (2005) Peptide and non-peptide G-protein coupled receptors (GPCRs) in skeletal muscle. *Peptides,* 26: 1528–1536

Ji, Y., Zhang, Z. and Hu, Y. (2009) The repertoire of G-protein-coupled receptors in *Xenopus tropicalis*. *BMC Genomics*, 10:263.

Johansen, S.D., Karlsen, B.O., Furmanek, T., *et al*. (2010) RNA deep sequencing of the Atlantic cod transcriptome. *Comparative Biochemistry and Physiology*, doi:10.1016/j.cbd.2010.04.005.

Kodzius, R. *et al.* (2006) CAGE: cap analysis of gene expression. *Nature Methods* 3, 211–222.

Kristiansson, E., Asker, N., Forlin, L., Larsson, D.G.J. (2009) Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. *BMC Genomics*, 10:345.

Larsson, L.C., Laikre, L., C André, C., Dahlgren, T.G., Ryman, N. (2010) Temporally stable genetic structure of heavily exploited Atlantic herring (Clupea harengus) in Swedish waters. *Heredity*, 104, 40–51.

Larsson, L.C., Laikre, L., Palm, S., André, C., Carvalho, G.R., Ryman, N. (2007). Concordance of allozyme and microsatellite differentiation in a marine fish, but evidence of selection at a microsatellite locus. *Mol Ecol*, 16(6):1135-47.

Lavoue, S., Miya, M., Saitoh, K., Ishiguro, N.B., Nishida, M. (2007) Phylogenetic relationships among anchovies, sardines, herrings and their relatives (Clupeiformes), inferred from whole mitogenome sequences, *Mol. Phylogenet. Evol*, 43 (3), 1096-1105.

Ledford, H. (2008) The death of microarrays? *Nature* 455, 847.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., Ecker, J.A. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523-536.

Liu, S., Zhou, Z., Lu, J., *et al.* (2011) Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. *BMC Genomics*, 12:53.

Jianguo Lu, J., Eric Peatman, E., Wenqi Wang, W., Qing Yang, Q., *et al.* (2010) Alternative splicing in teleost fish genomes: same-species and cross-species analysis and comparisons. *Mol Genet Genomics* 283:531–539.

Mathavan, S., Lee, S.G.P, Mak, A., *et al*. (2005) Transcriptome Analysis of Zebrafish Embryogenesis Using Microarrays. *PLoS Genetics*, 1: 2.

Metpally, R.P.R. and Sowdhamini, R. (2005) Genome wide survey of G protein coupled receptors in *Tetraodon nigroviridis. BMC Evolutionary Biology* 2005, 5:41

Morozova, O., Hirst, M., Marra, M.A. (2009) Applications of New Sequencing Technologies for Transcriptome Analysis. *Annu. Rev. Genomics Hum. Genet*. 10:135–51.

Mortazavi, A., Williams. B.A., McCue. K., Schaeffer. L., Wold. B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, Vol 5: 7.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320:1344-1349.

Niimura, Y. and Nei, M. (2005) Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *PNAS*, 102:17, 6039–6044.

Pariset, L., Chillemi, G., Bongiorni, S., Romano Spica, V., Valetini, A., (2009) Microarrays and high-throughput transcriptomics analysis in species with incomplete availability of genomic sequences. *New Biotechnol*. 25, 272–279.

Pevzner, P.A., Tang, H., Waterman, M.S. (2001) An Eulerian path approach to DNA fragment assembly. *PNAS,* 98: 17, 9748-9753.

Reinartz, J. *et al.* (2002) Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief. Funct. Genomic Proteomic* 1, 95–104.

Ruzzante, D.E., Mariani, S., Bekkevold, S. et al. (2006). Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *Proc Biol Sci,* 273(1593):1459-64.

Santini, F., Harmon, L.J., Carnevale, G., Alfaro, M.E. (2009) Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evolutionary Biology,* 9:194

Shaw, P.W., Turan, C., Wright, J.M., O'connel, M., Carvalho, G.R. (1999) Microsatellite DNA analysis of population structure in Atlantic herring (Clupea

harengus), with direct comparison to allozyme and mtDNA RFLP analyses, *Heredity*, 83, 4:490-9.

Schulze, A., Downward, J. (2001) Navigating gene expression using microarrays-a technology review. *Nature Cell Biology*, 3.

Shi, C.Y., Hua, Y., Wei, C.L., *et al.* (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics*, 12:131

Sultan, M., *et al*. (2008) A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321, 956.

Surget-Groba, Y., Montoya-Burgos, J.I. (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res*, 20: 1432-1440.

Surget-Groba Y, Montoya-Burgos JI: Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Res* 20(10):1432-1440.

Velculescu, V.E., Zhang, L., Zhou, W., *et al*., 1997 Characterization of the Yeast Transcriptome. *Cell*, 88: 243–251.

Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* 270, 484–487 (1995).

Vera, J.C., Wheat, C.W., Fescemyer, H.W., Frilander, M.J., Crawford, D.L., Hanski, I., Marden, J.H. (2008) Rapid transcriptome characterization for a non model organism using 454 pyrosequencing. *Molecular Ecology*, 17:1636-1647.

Wei, Z., Liu, X., Feng, T., Chang, Y. (2011) Novel and Conserved MicroRNAs in Dalian Purple Urchin (*Strongylocentrotus nudus*) identified by Next Generation Sequencing. *Int. J. Biol. Sci.* 7(2): 180-192.

Wang, X.W., Luan, J.B., Li, J.M., Bao, Y.Y., Zhang, C.X. & Liu, S.S. (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics* 2010, 11:400

Wang, Z., Gerstein, M., Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 57-63, doi: 10.1038/nrg2484.