



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

# Transcriptome Analysis of Atlantic Herring (*Clupea harengus*) using Next Generation Sequencing (NGS))

*Nima Rafati*



---

Examensarbete / Swedish University of Agricultural Sciences,  
Department of Animal Breeding and Genetics,  
405  
Uppsala 2013

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

## **Transcriptome Analysis of Atlantic Herring (*Clupea harengus*) using Next Generation Sequencing (NGS))**

*Nima Rafati*

### **Supervisors:**

Leif Andersson, SLU, Department of Animal Breeding and Genetics  
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:** 2013

**Name of series:** Examensarbete / Swedish University of Agricultural Sciences,  
Department of Animal Breeding and Genetics, 405

**On-line publication:** <http://epsilon.slu.se>

**Key words:** Atlantic Herring, Transcriptome, RNA-Seq, Next Generation Sequencing





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

## Transcriptome Analysis of Atlantic Herring (*Clupea harengus*) using Next Generation Sequencing (NGS)

*Nima Rafati*



Education and Culture

**Erasmus Mundus**



---

Department of Animal Breeding and Genetics  
Examensarbete  
Uppsala 2011

Master's Thesis, 30 HEC  
Erasmus Mundus programme – European  
Master in Animal Breeding and Genetics

---



<b>1</b>	<b>Abstract</b>	<b>1</b>
<b>1.</b>	<b>Introduction</b>	<b>2</b>
1.1	What is Transcriptome?	4
1.1.1	One gene, many proteins	6
1.2	Transcriptome analysis	6
1.3	RNA-Seq methodology	8
<b>2</b>	<b>Materials and Methods</b>	<b>10</b>
2.1	Data	11
2.2	Contamination checking	11
2.3	Transcriptome assembly tools	11
2.4	Assembly validation	12
2.5	Gene annotation	12
2.6	Finding full length sequences	13
2.7	Gene expression	13
2.8	Multiple sequence alignment (MSA)	13
<b>3</b>	<b>Results</b>	<b>14</b>
3.1	Data trimming	14
3.2	Transcriptome assembly	14
3.3	Gene annotation	17
3.3.1	BLAST on available herring genes	17
3.3.2	BLAST on non redundant (nr) protein database	18
3.3.3	BLASTX against conserved domain database (CDD)	19
3.3.4	BLASTN on Oryzias latipes (medaka) UniGene and Danio rerio (zebrafish) muscle ESTs	20
3.3.5	Finding full length sequences	20
3.3.6	Alignment on eukaryotic tRNAs	20
3.4	Gene expression	20
3.5	Glucose 6-phosphate isomerase (GPI)	22
3.6	Alpha actin	23
<b>4</b>	<b>Discussion</b>	<b>23</b>
<b>5</b>	<b>Conclusion</b>	<b>28</b>
<b>6</b>	<b>Acknowledgments</b>	<b>28</b>
<b>7</b>	<b>References</b>	<b>29</b>
<b>8</b>	<b>Appendix</b>	<b>33</b>
8.1	Genome/Transcriptome assembly	33
8.2	Quality of sequenced reads	35
8.3	Determining optimal k-mer	36
8.4	N50 and N90 calculation	36
8.5	Scripts	37
8.5.1	N50 and N90	37
8.6	Software commands	39

8.6.1	SeqClean .....	39
8.6.2	Inchworm.....	39
8.6.3	SSPACE.....	39
8.6.4	CAP3.....	39
8.6.5	Trinity.....	40
8.6.6	SOAPdenovo and GapCloser .....	40
8.6.7	Jellyfish .....	41
8.6.8	SOAPaligner .....	41
8.6.9	CD-HIT.....	41
8.6.10	UCLUSTAL.....	41
8.6.11	BLAST .....	42
8.6.12	Mosaik .....	42
8.6.13	Samtools .....	42
8.6.14	VarScan .....	42
8.6.15	Supplementary figures.....	44



## 1 ABSTRACT

The Atlantic herring is one of the most abundant fish species in the Northern hemisphere especially in the Northeast Atlantic. There are various stocks of this fish due to their spawning time and their vast distribution. Many studies have tried to characterize herring populations and these efforts can be combined with massively parallel sequencing technologies to develop genetic resources. The transcriptome is a repertoire of RNAs in cells produced by transcription and messenger RNAs (mRNAs) are responsible for gene expression. This repertoire may change during different life stages and environmental conditions. Transcriptome studies have been mainly dependent on a reference genome and developed tools are not applicable for non-model organisms for which the reference genome is missing or only partially available. In this study we present a *de novo* transcriptome assembly by using different strategies and one specific transcriptome assembler, Trinity. The RNA was collected from muscle of a male spring spawning herring and then sequenced with an Illumina Hiseq 2000 machine. After trimming for low quality nucleotides and possible contaminations, the assembly resulted in 131,788 contigs with a total size of 40 million nucleotides (Mnts). The transcriptome generated by Trinity was compared with other assemblies, assembled by a genome assembler (SOAPdenovo), Inchworm, SSPACE, and CAP3. The results showed that Trinity developed a more reliable assembly. We validated almost 47% of all contigs by comparison to available databases such as nr and other vertebrate transcripts. Also, we quantified the relative expression of transcripts by counting the number of aligned reads per kilobase per million mapped reads (RPKM). Furthermore, we performed in-depth studies of two genes. We identified two copies of Glucose 6-phosphate isomerase (GPI) on draft genome assembly. Our results showed that the sampled fish was heterozygous at the GPIb locus. Among the alpha actin isoforms, we identified the alpha actin b transcript (ACTA1b) in fast muscle corresponding to spring spawning herring fish characteristics.

# 1. INTRODUCTION

The Atlantic herring (“sill” in Swedish) is found among the 140 native fish species in Sweden and close countries in northeast Atlantic Ocean and together with sardines, capelin, anchovies, menhaden, and small mackerels, it makes up more than one quarter of all saltwater landings (Encyclopædia Britannica). Herring is globally among the most abundantly caught fish and it is important for Swedish economy. The global herring capture production reached to 4 million tons in 1960s and dropped to 887,553 tons during 1980s when Sweden had the highest capture in global production (~13%) (Figures 1 and 2). Also, it is an important food source being 15% of all herring captures allocated for human consumption and it has played a main role in Swedish cuisine since the Middle age. Herring is served pickled at Midsummer and Christmas, and also the fermented herring is popular in northern part of Sweden (Swedish fish database).

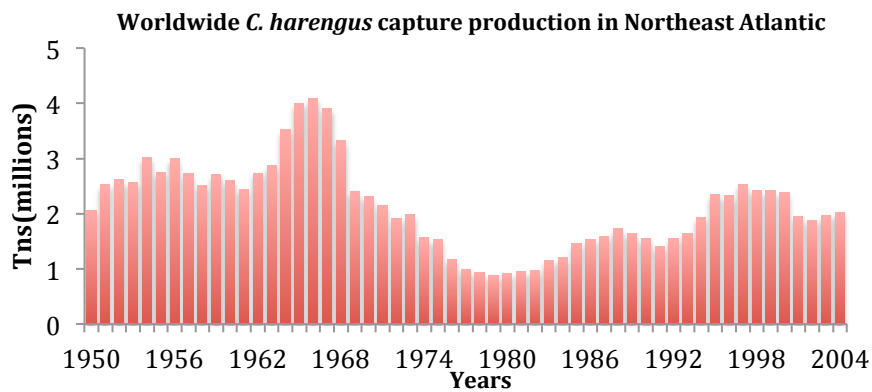


Figure 1- Worldwide Herring capture production between 1950 and 2004 (FAO fishery statistics).

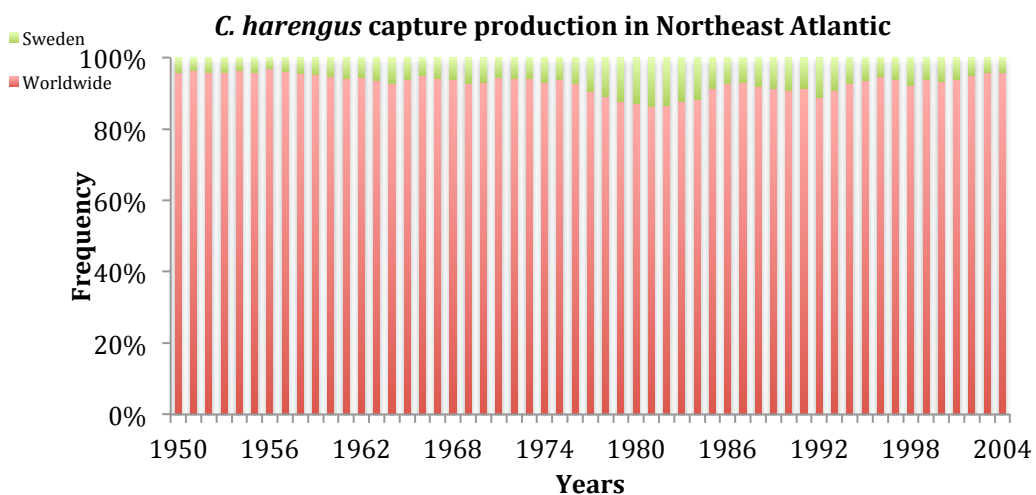


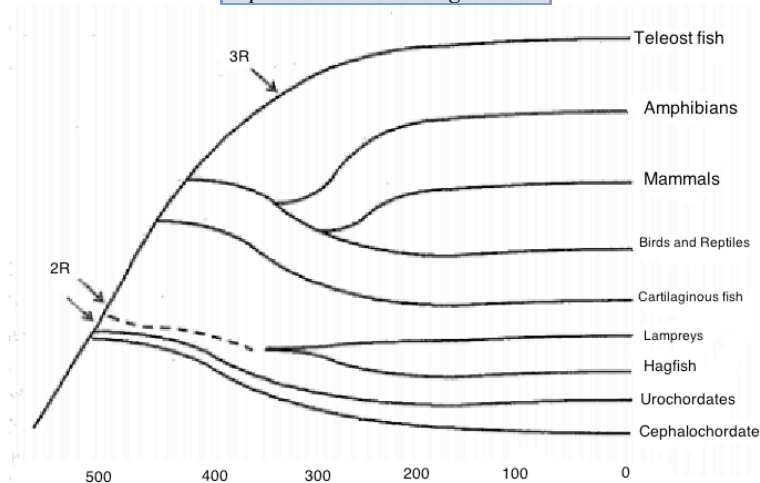
Figure 2- Sweden and worldwide herring capture production between 1950 and 2004 (FAO fishery statistics)

From an evolutionary point of view, herring belongs to Teleost infraclass, which is one of the most important diversifications (see table 1 for taxonomy information). Most of the living fish are part of this infraclass, which includes near 29,000 species with around 200 species in the *Clupidae* family and the Atlantic herring (*Clupea harengus*) is one of them. Hox gene family studies, molecular, and paleontological data suggested that the herring diverged from other teleost

fish like zebrafish, medaka, stickleback, tetradon, and fugu between 147 and 314 million years ago (mya) (Moyle et al., 1992 and Peng et al., 2009). Teleosts fish have gone through a whole genome duplication approximately 420 mya, based on gene family studies and large scale genomic analyses (Figure 3; reviewed in Sundström 2010).

**Table 1- Taxonomy information of Atlantic Herring (NCBI).**

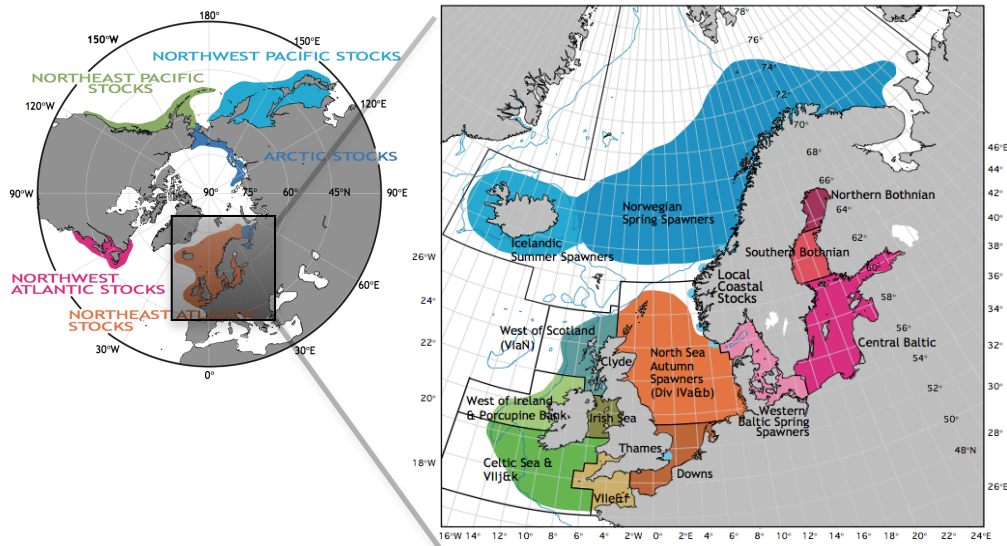
Kingdom:	Animalia
Phylum:	Chordata
Subphylum:	Vertebrata
Superclass:	Osteichthyes
Class:	Actinopterygii
Subclass:	Neopterygii
Infraclass:	Teleostei
Order:	Clupeiformes
Family:	Clupeidae
Subfamily:	Clupeinae
Genus:	Clupea
Species:	<i>harengus</i>



**Figure 3- Chordate evolution and time points for whole genome duplication. Time scale is in million of years (Sundström, 2010).**

Herring is a bony fish, schooling in coastal waters where the light is intense and they use it for controlling their vertical distribution. Schooling, fast movement, and their silvery body, reflecting light, act as anti-predator device (FishBase). The prey of herring is dependent on location, season, and age but mainly uses small planktons and copepods when it is adult. They can grow until 45 cm and their usual length is 20-30 cm with a maximum weight of 1,050 g (FishBase). The lifespan is quite variable, ranging between 4 to 20 years and they enter sexual maturity in 3-9 years of age (FishBase and HerringNetwork). Herring is a diploid organism with 52 chromosomes and males are heterogametic (Doucette and Fitzsimons 1982). However, unlike mammals and birds, teleost fish have diversity in sex-determination systems influenced by multiple autosomal genes and some environmental factors e.g. temperature (Volf, 2002).

Herring has a wide distribution at both sides of the Atlantic Ocean and within each region they are grouped by spawning time. Figure 4 shows various populations with different spawning time in the northeast Atlantic. There is at least one herring stock spawning any month of the year. The female lays 30,000 eggs on average, ranging from 20,000 to 40,000, on the seabed, on beds of algae, or stones (HerringNetwork and FishBase).



**Figure 4- Left) Herring distribution in the worldwide classified in 5 stocks. Right) Herring distribution in northeast Atlantic according to their spawning time (HerringNetwork).**

Herrings have been classified as species, sub-species, and stocks ecologically and morphologically apart from their genetic differences and reproductive isolations (Ryman et al., 1984). Although many studies have suggested that these differences among herring stocks are due to environmental effect, some researchers have tried to monitor the variation by comparing markers in different population. For instance, Ryman et al. (1984) reported 54 alleles for 17 loci indicating a large effective population size, gene flow between different populations, or a recent colonization that have not yet gone into genetic drift. However, the overall variation within and between populations could not be addressed thoroughly by this limited set of genes since the analysed loci may not have undergone selection and consequently did not represent the correct picture of polymorphisms present in each population. The massively parallel sequencing technologies have revolutionized biology and computer sciences by generating high amount of information in short time and lower costs. This information enables us to identify variation at a single-base resolution either in DNA (genome) or RNA (transcriptome). The aim of this study was to generate a muscle transcriptome of the Atlantic herring and it is a part of major effort to characterize genetic differentiation of whole genome.

### 1.1 What is Transcriptome?

A transcriptome is a set of synthesized RNA molecules from both coding and non-coding regions of DNA in one cell (Brown, 2002). There are different RNA molecules that have specific roles in biological process. Transcription is the process generating the transcriptome and proteins are synthesized by transcriptome translation into amino acid polymers (Figure 5) (Brown, 2002). RNA synthesis is subject to strong spatial and temporal regulation, with some RNAs being exclusive to some cell types or time points, while other RNAs are continuously expressed (Brown, 2002 and Alberts et al., 2007).

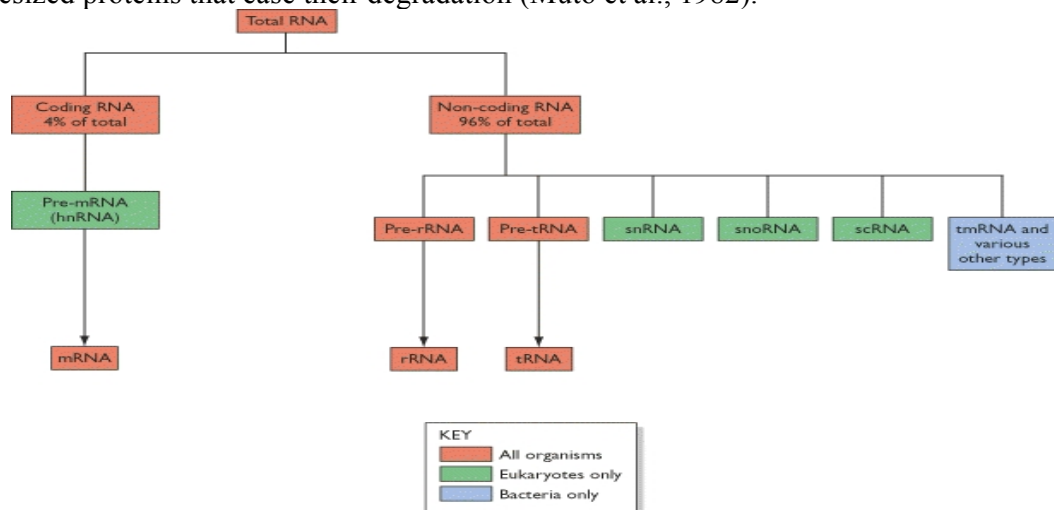


**Figure 5-Genome, Transcriptome, and Proteome.**

RNAs can be classified into two main categories (Figure 6): coding and non-coding. It has been shown in eukaryotes, bacteria, and archaea that 4% of the cellular RNAs are coding RNA or messenger RNA (mRNA). They are transcribed as complementary sequence of DNA that will be translated into protein by help of transfer RNA (tRNA) leading amino acids to their location in peptides corresponding to their codon on mRNAs (Brown, 2002).

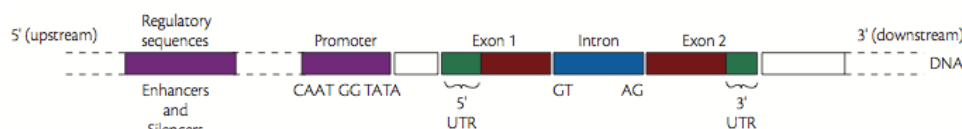
Ribosomal RNAs (rRNAs) are the most abundantly synthesized non-coding RNAs (80 to 90%) and vital component of ribosomes, the engines of RNA translation into amino acid chain. Other kinds of RNAs are present in limited numbers and participate in regulation of RNA expression (Wilhelm and Landry 2009, Brown, 2002).

Eukaryotes, have other kinds of non-coding RNAs such as small nuclear RNA (snRNA) involved in mRNA processing, small nucleolar RNA (snoRNA) which has a key role in rRNA processing, and small cytoplasmic RNA (scRNA) which functions are unclear (Brown, 2002). Micro RNAs (miRNA) are another class of non-coding RNAs that plays an important role in the regulation of gene expression especially in organogenesis and development (He and Hannon, 2004). In bacteria and archaea, there are some non-coding RNAs other than the mentioned classes such as transfer-messenger RNA (tmRNA). tmRNA acts as tRNA and attaches short peptide fragments onto synthesized proteins that ease their degradation (Muto et al., 1982).



**Figure 6- The RNA classification. mRNA: messenger RNA, rRNA: ribosomal RNA, tRNA: transfer RNA, snRNA: small nuclear RNA, snoRNA: small nucleolar RNA, scRNA: small cytoplasmic RNA, tmRNA: transfer-messenger RNA (Genomes 2, 2002).**

Eukaryotic protein-coding genes include untranslated regions at ends of gene (5'-UTR and 3' UTR), coding (exons), and non-coding (intron) sequences (Figure 7). Also, in the upstream regions there are regulatory elements controlling the level of expression and transcription initiation, so-called promoters, which facilitate the transcription.



**Figure 7- Eukaryotic protein-coding gene structure.**

In the transcription reaction the gene is completely transcribed including UTR, exons, and introns but only exons are finally translated into proteins. Therefore, introns should be removed through a process so-called intron splicing (Figure 8). snRNAs and a protein called spliceosome play key roles in this biochemical reaction (Brown, 2002).

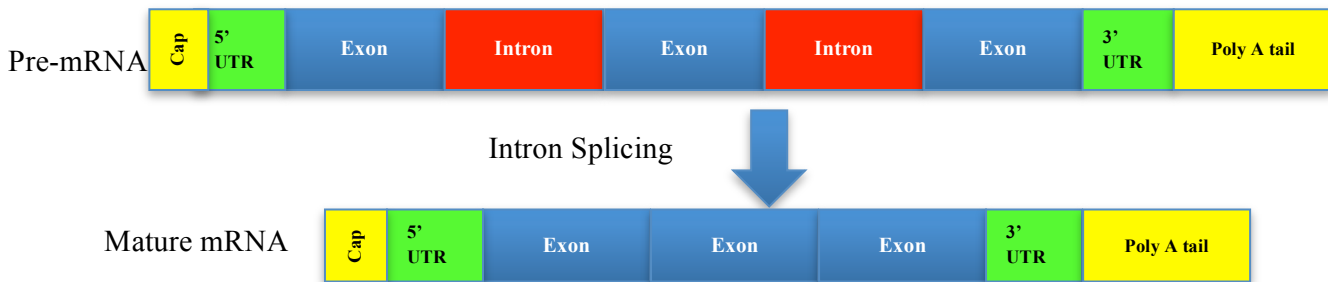


Figure 8- Intron splicing.

### 1.1.1 One gene, many proteins

“One gene one protein” theory was valid up until the discovery of the splicing pathway in the 1980s. This discovery showed that each gene generates various alternate transcripts (isoforms) by using alternative initiation, splicing, or poly-adenylation sites. Consequently, one gene can produce many proteins; in other word, each gene can be expressed differently, producing the so-called alternative expression (Black, 2000 and Brown, 2002). Figure 9 shows different patterns of pre-mRNA processing and final transcripts.

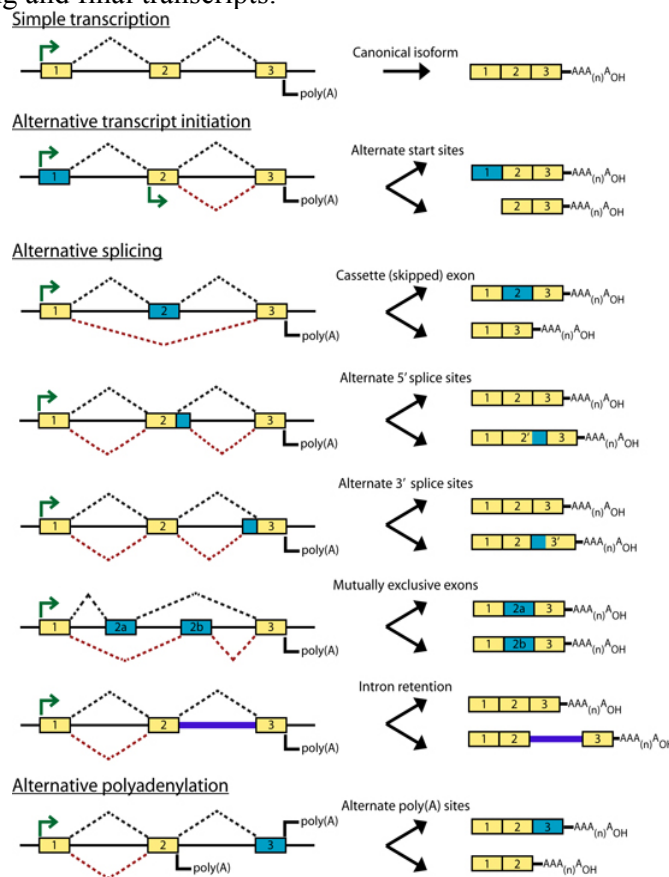


Figure 9- Alternative-splicing classification.

## 1.2 Transcriptome analysis

High throughput transcriptome analysis started with microarray technique by monitoring expression level of annotated genes. Later on, tiling arrays, the following generation of microarray, was used as a new tool to detect un-annotated genes. In this method, designed probes interrogating a genome, regardless of any gene annotation, were used. Although experiments using tiling arrays revealed that the transcriptome complexity had been underestimated, the resolution was low due to a limitation of designed probes to cover the genome properly. Also,

this technique is highly affected by the amount of input RNA influencing the sensitivity, specificity and splice event detection (Mortazavi et al., 2008, Lister et al., 2009, Marguerat et al., 2009, and Marguerat and Bähler 2010).

Another approach taken was expressed sequence tag (EST) sequencing of complementary DNA (cDNAs). This approach has been used as a main method for transcriptome discovery in model organisms. However, there are some constraints in this method. For instance, a part of transcriptome is usually covered with low capability in quantifying the expression (Mortazavi et al., 2008).

There are other known methods for transcriptome analysis such as serial analysis of gene expression (SAGE) and massively parallel signature sequencing (MPSS) which are based on using cDNA libraries and sequencing (Marguerat and Bähler 2010). These methods were dependent on chain termination sequencing that Fredrick Sanger invented in 1977.

In the beginning the aim of Sanger technique was genomic sequencing but in the 1990s scientists used it to sequence cDNA (Scheibye-Asinga et al., 2009). After the sequencing technology development and industrialization of massively parallel sequencing technologies (so-called Next Generation Sequencing or Now Generation Sequencing- NGS) the cost of sequencing per base has remarkably decreased (Figure 10) (Lister et al., 2009).

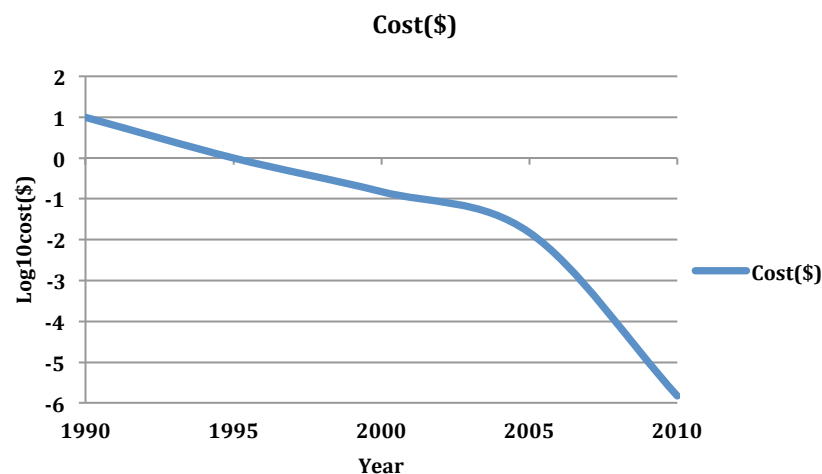


Figure 10- DNA sequencing cost from 1990 to 2010 (<http://www.synthesis.cc/2010/08/recent-dna-cost-and-productivity-figures-from-the-economist.html>).

The main difference between NGS and Sanger sequencing output is the shorter lengths of the sequence reads obtained by three main NGS platforms: 454 Life Sciences (a Roche company), the Illumina Genome Analyser (formerly called Solexa), and AB SOLiD system (owned currently by Life technologies) (Table 2) (Marguerat et al., 2009, Scheibye-Asinga et al., 2009, Marguerat and Bähler 2010). The main challenge in using the NGS products is their assembly which needs complicated algorithms that are computationally intensive (Pop 2009, Marguerat et al., 2009, Scheibye-Asinga et al., 2009).

Table 2- NGS generated reads' length (Marguerat and Bähler 2010).

Technology	Read length (nts)	Year of invention
454-pyrosequencing <sup>#</sup>	200-600	2005
Illumine/Solexa <sup>¶</sup>	35-100	2006
SOLiD	25-50	2007

<sup>#</sup> With the GS FLX Titanium series in 2008, the sequencing length increased to 400-600 nts.

<sup>¶</sup> Hiseq 2000 system can produce reads with the length of 100 nts.



By evolving broad applications of NGS, the RNA-Seq approach was born; RNA-Seq has been used to detect alternative splicing, novel genes and transcripts, and genomic structural variations. RNA-Seq can measure and quantify the expression level with a higher sensitivity than other aforesaid methods by counting the number of reads hitting a transcript or the average number of hits per base (Marguerat et al., 2009). There are two parameters for expression measurement: reads per kilo base per million mapped reads (RPKM) (Mortazavi et al., 2008) and fragments per kilobase per million mapped reads (FPKM) (Trapnell et al., 2010). For calculation of these parameters, first, reads are realigned to transcripts and then the coverage is normalized by number of aligned reads -relative to all mappable reads- and length of transcripts. In FPKM, the number of paired reads is calculated while in RPKM single reads are considered. These estimations of transcript expression are in good agreement with acquired results from microarrays provided adequate depth of sequencing (Mortazavi et al., 2008).

In addition, RNA-Seq overcomes some shortcomings of microarray transcription detection such as probe cross-hybridization, low sensitivity and specificity (especially for low expressed genes), dye-based detection issues, and probe designing constraints that limit the RNA splice detections and unmapped genes. Another benefit of RNA-Seq is its ability to recognize paralogous genes and quantifying their level of expression by read count. In addition to the advantages mentioned above, RNA-Seq also makes it possible to identify polymorphisms, alternative expression, and it is independent of a reference genome (Lister et al., 2009).

As mentioned before one of the drawbacks of NGS is the short read length. Therefore, the main challenge in this method is its application to assembly (Mortazavi et al., 2008 and Marguerat et al., 2009). Hence, a read length increase would likely enable us to detect splice patterns more efficiently and develop *de novo* assembly of transcriptomes in cases that the reference genome is not available (Lister et al., 2009).

### **1.3 RNA-Seq methodology**

RNA-Seq basically follows the DNA sequencing approach and it is applicable for all organisms provided that enough RNA is available. Also, it is dependent on sequencing technology. In the first step, it is crucial to extract mRNAs and remove the rest of RNAs especially rRNA, since more than 90% of RNAs in cell are in ribosomal form. There are three different methods applied for rRNA removal (Wilhelm and Landry 2009):

- 1- The rRNA depletion kits such as RiboMinus (Invitrogen). In this method antisense sequence of ribosomal transcripts conjugating to either biotinylated or magnetic beads are used.
- 2- In the poly-A enrichment, the same kind of beads as previous method is used with this difference that they carry oligo dT and the RNA molecules with poly-A tails will attach to them.
- 3- Using enzymes degrading uncapped RNAs

After mRNA purification, they are fragmented by either chemical or physical approaches to generate fragments of 200-300 nts. These fragments go through amplification by reverse transcriptase using random primer to generate cDNAs (Mortazavi et al., 2008, Marguerat and Bähler 2010). Next, adapters are attached to cDNAs and cDNA libraries are sequenced by NGS technologies.

The library preparation is highly important for RNA-seq as this step dictates how well the cDNA sequence data will mirror the complexity of the original sampled RNA. In the original protocol, which is specific for genomic DNA, adapters are added to double stranded DNA (dsDNA) fragments while the transcript is a single stranded molecule and the information on transcriptional



direction will be lost. Consequently, two different protocols were introduced: strand-specific and double-stranded.

With the strand-specific protocol, the purified mRNAs are marked by different methods such as pre-treating with sodium bisulphate to detect the original transcript. With the double stranded, the original transcript is not known (Lister et al., 2009, Marguerat and Bähler 2010).

After sequencing, different approaches are applied to use this information depending on genome availability of the organism. For model organisms, reads are usually mapped to the genome by software such as Bowtie and then the aligned reads are quantified by Cufflinks or ERANGE (Enhanced Read Analysis of Gene Expression) (Mortazavi et al., 2008). However, in non-model organisms (species without assembled genome), the transcriptome must be assembled *de novo* (see Appendix 8.1 for assembly approaches and some terminologies).

There are some specific genomic assembly tools that have been used in different transcriptome analyses. For instance, SOAPdenovo software has been used for tea and whitefly *de novo* transcriptome assemblies (Wang et al., 2010 and Shi et al., 2011). However, there are some barriers in their application for transcriptome assembly. First, the transcriptome coverage is highly dependent on expression level and regular genome assembly parameters cannot be applied. Second, various splice variants per locus can complicate the assembly. Third, as there are repetitive sequences in a genome shared features in form of protein domains add ambiguity to the process in a transcriptome assembly (Birol et al., 2009). Hence, new assembly tools have been developed specifically for transcriptome assembly such as Trinity (Grabherr et al., 2011) that we used for muscle transcriptome assembly of an Atlantic herring individual. As a result, the objectives of this study are:

- 1- *De novo* transcriptome assembly of an Atlantic herring using Illumina reads from mRNA captured libraries.
- 2- Validation of the Trinity software by comparison to other common assembly tools.
- 3- Find variation in annotated genes.
- 4- Find copies of two specific genes (alpha actin and GPI).

In addition to genome annotation we will provide information about gene models that will be mapped onto the genome. Also, it could assist the genome assembly by providing links between genomic contigs.

## 2 MATERIALS AND METHODS

Figure 11 shows the flowchart followed in the *C. harengus de novo* transcriptome assembly process implemented in this study.

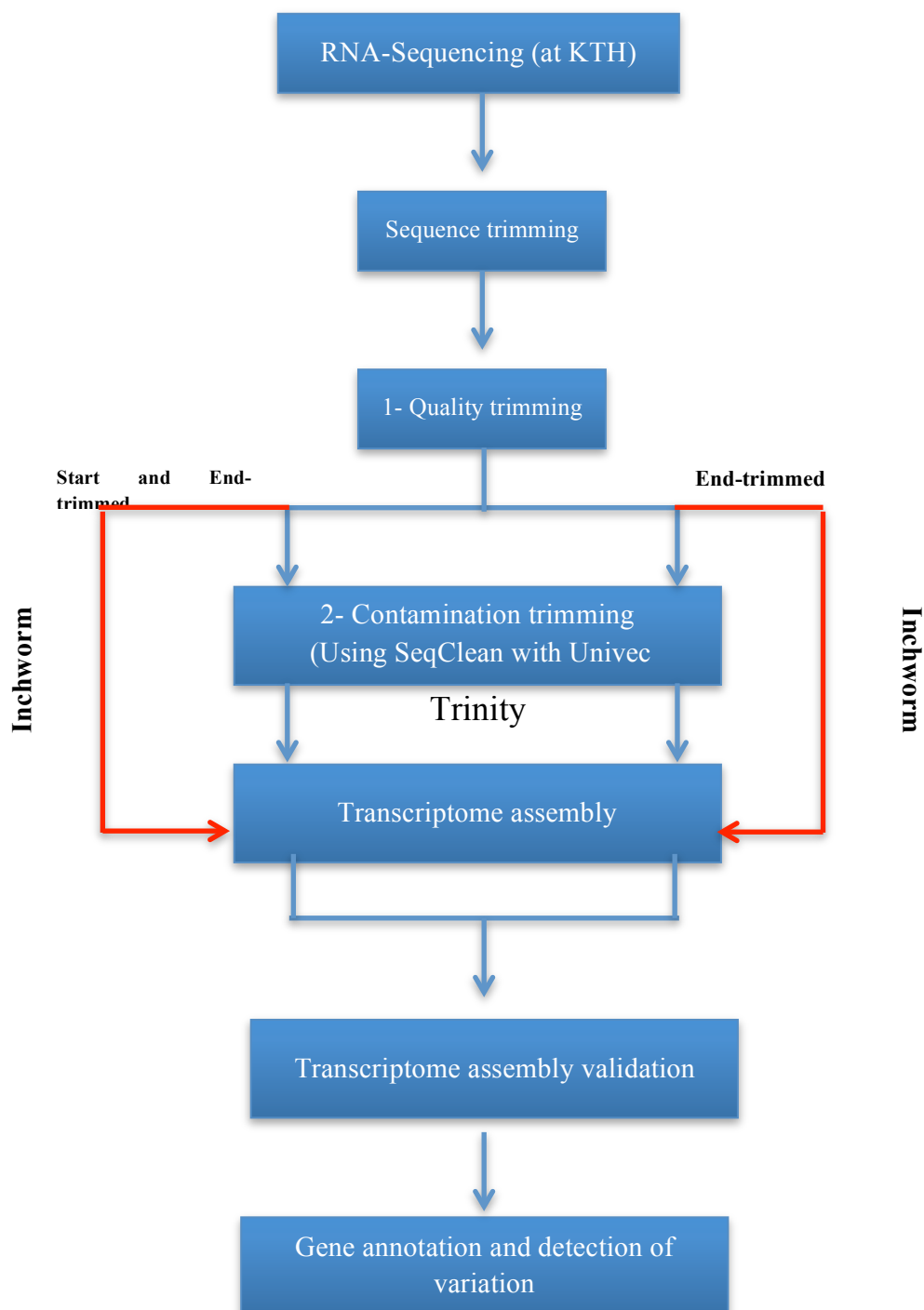


Figure 11- Flowchart of the *C. harengus de novo* transcriptome assembly.

## 2.1 Data

A male spring spawning herring was caught in June 2010, in the archipelago of Stockholm. The RNA was extracted from muscle tissue and rRNAs were depleted. The RNA preparation, library generation, and sequencing were done at the Science for Life laboratory in Stockholm.

For the assembly of the transcriptome, we received raw reads sequenced at an Illumina HiSeq-2000 machine as paired end reads with length of 101 nts in two fastq files: read1 and read2 (see Appendix 8.2 for fastq). The average insert size was 240 nts deviated between 180 and 400 nts. All lanes were spiked with 1% of phiX control library.

Before assembly it is necessary to trim the raw reads for possible errors emerging during either sequencing or cloning that can affect the efficiency and reliability of assembly. The implemented trimming strategy was based on average quality value. We chose a value of 30 representing 0.1% sequencing error (Appendix 8.2) that covers 25 nucleotides at 3' end of reads. After removing these nucleotides we generated a new dataset of reads with a length of 76 nts.

## 2.2 Contamination checking

NGS products must be checked for probable contaminations that may have been introduced during library preparation, cloning, or sequencing. Thus, reads are aligned to the sequences from the vectors, primers, and adaptors that are used in library preparation and sequencing process. UniVec (<ftp://ftp.ncbi.nih.gov/pub/UniVec/> downloaded 28<sup>th</sup> of March 2011) is a database that includes this information and it was used as a reference to check the raw reads with the SeqClean software (released 22<sup>nd</sup> of February 2011) (<http://sourceforge.net/projects/seqclean/>) from The Genome Institute (TGI).

## 2.3 Transcriptome assembly tools

For the transcriptome assembly we used the program Inchworm (released 20<sup>th</sup> of January 2011) (<http://sourceforge.net/projects/inchworm/files/>) individually first and later also combined with Chrysalis and Butterfly in a complete software package called Trinity (released 13<sup>th</sup> of March 2011) (<http://sourceforge.net/projects/trinityrnaseq/files/>). All of the modules in Trinity are based on the *de Bruijn* graph algorithm, one of the most common and efficient approaches for assembly of next generation sequences (Appendix 8.1).

We used Inchworm with k-mer equal to 21 as the optimum value, after comparing assemblies with different k-mers ranging from 19-29 and also based on our analysis for optimal k-mer estimation (Appendix 8.3). Based on the results that we had obtained from Inchworm we used k-mer 21 as the best value to assemble the transcriptome in Trinity.

The Inchworm assembly were further analyzed by merging contigs with overlapping similarities between generated contigs; CAP3 (released in 1999) (<http://seq.cs.iastate.edu/cap3.html>) is a software that does this job using overlap-layout-consensus (OLC) approach (Appendix 8.1). Also, we used a stand-alone program for scaffolding pre-assembled contigs (SSPACE), which extends contigs and converts them to scaffold using paired reads information. SSPACE has been generated for CLC-BIO assembly package by BASECLEAR Institute (released in 2010) (<http://www.baseclear.com/sequencing/data-analysis/bioinformatics-tools/sspace/>). After generating final assembly, we tried to remove redundant sequences by two programs: CD-HIT (version 4.5.4 released on 11<sup>th</sup> of March 2011 [http://www.bioinformatics.org/project/filelist.php?group\\_id=350](http://www.bioinformatics.org/project/filelist.php?group_id=350)) and UCLUSTAL (version 4.0 released in 2010 [http://www.drive5.com/usearch/nonprofit\\_form.html](http://www.drive5.com/usearch/nonprofit_form.html)).

## 2.4 Assembly validation

The number of generated contigs, their length distribution, and total assembled transcriptome such as N50 and N90 (Appendix 8.4) were calculated using in-house scripts (Appendix 8.5.1). Also, we calculated the average read depth of the assembly by this formula:

$$D = \frac{N \cdot L}{T} \quad \text{F 1}$$

where  $N$  is the number of reads,  $L$  is the length of reads, and  $T$  is the size of transcriptome.

Trinity was specifically developed for *de novo* transcriptome assemblies and for its validation we used a genome assembly software, SOAPdenovo (version 1.05, released on the 14<sup>th</sup> of February 2011) (<http://soap.genomics.org.cn/soapdenovo.html>), with the same k-mer of 21 to have a symmetrical comparison. Also, we used jellyfish (version 1.1 released 23<sup>rd</sup> of April 2011) (<http://www.cbc.umd.edu/software/jellyfish/>) a tool for fast and memory-efficient counting of k-mers, to retrieve the number of unique k-mer after the transcriptome assemblies were completed by aforesaid assemblers.

One of the common analyses for assembly validation is realigning the input reads on the generated contigs to quantify the amount of used reads. Also, the results of such alignments can be used to identify genetic variation such as single nucleotide polymorphisms (SNPs) and indels. Software used for these tasks included: for read alignment, Mosaik (version 1.0.1388, released on 1<sup>st</sup> February 2010) (<http://bioinformatics.bc.edu/marthlab/Mosaik>) and SOAPaligner (version 2.2.0, released on 13<sup>th</sup> of August 2009) (<http://soap.genomics.org.cn/>), for detection of genetic variation we used samtools (version 0.1.7a and 0.1.12, released on 16<sup>th</sup> of November 2009 and 5<sup>th</sup> of November 2010, respectively) (<http://sourceforge.net/projects/samtools/files/samtools/>) and VarScan (version 2.2.5, released on 28<sup>th</sup> April 2011) (<http://sourceforge.net/projects/varscan/files/>).

Moreover, we aligned the transcriptome on a draft version of the herring genome that had been assembled by our group using SOAPdenovo. For this alignment we used the regular nucleotide basic local alignment tool (BLAST) (version 2.2.25, released 31<sup>st</sup> of March 2011 2.2.25) (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>).

Furthermore, for some of the statistical analyses we used Excel 2010 and R (version 2.12.2, released 25<sup>th</sup> of February 2011).

## 2.5 Gene annotation

Gene annotation of the assemblies were performed by comparing our transcriptome with other database resources, such as the non-redundant (nr) protein database (set of all non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF), the conserved domain database (CDD), and available mRNA sequences of *C. harengus* released by the National Centre for Biotechnology Information (NCBI) (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/>). The significant hits have been filtered by an E-value threshold of  $10^{-5}$ . Two BLAST programs were used to align contigs to these databases: BLASTX for protein sequences and BLASTN for alignment on nucleotide sequences.

For detailed analysis on some specific genes we retrieved genes/transcripts/ESTs of some organisms (medaka, stickleback, zebrafish, chicken, mouse, and human) from NCBI (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/>) and the Ensembl genome browser (<http://www.ensembl.org/info/data/ftp/index.html>, downloaded 12<sup>th</sup> of April 2011). Moreover, for non-coding RNA analyses we used available eukaryotic tRNAs on Genomic tRNA database (<http://gtrnadb.ucsc.edu/download.html>, downloaded on 17<sup>th</sup> of June 2011). In addition to these

databases, for individual alignment comparisons we used the UCSC genome browser (<http://genome.ucsc.edu/>).

## 2.6 Finding full length sequences

Full-Lengther is a web tool to discover 5' and 3' sites of transcripts by using BLAST against UniProt. It considers a contig as a full-length sequence when there is a start codon before the alignment on protein (Figure 12a). When a contig starts after the protein with a distance of 15 aa from the beginning, it is called a putative full-length sequence (Figure 12b). We used this tool to detect those contigs that did not have significant alignment on other databases.

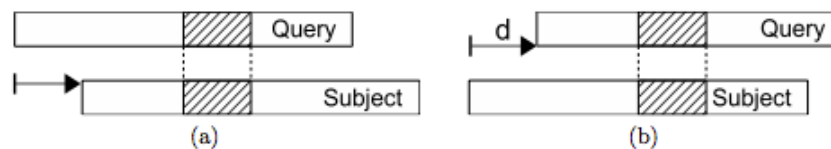


Figure 12- (a) The query (contig/EST) starts before subject. (b) The query (contig/EST) starts after subject. Grey box shows the alignment region (Lara et al.).

## 2.7 Gene expression

For quantification of transcripts expression, we used FPKM and RPKM estimations. Trinity reported the FPKM values in the header of each transcript in final assembly and for RPKM calculation we used this formula:

$$R = \frac{10^9 C}{NL} \quad (\text{Mortazavi et al., 2008}) \quad \text{F 2}$$

where  $C$  is the number of aligned reads on contig,  $N$  is the total number of mappable reads, and  $L$  is the length of the contig.

## 2.8 Multiple sequence alignment (MSA)

Clustalx (version 2.0.10, released 10<sup>th</sup> of January 2011) (<http://www.clustal.org/>) were used to perform MSA for protein sequences. The herring contigs were translated using a web tool called open reading frame finder (ORF-finder; <http://www.ncbi.nlm.nih.gov/projects/gorf/>).

All our analyses were performed at the computer clusters at the Uppsala Multidisciplinary Centre for Advanced Computational Science (UPPMAX) located at Uppsala University. There are three main clusters in UPPMAX and we used **kalkyl** cluster equipped with 2.784 64-bit processor cores, 9.504 GB of memory, and 113 TB of disk space. (<http://www.uppmax.uu.se>).

### 3 RESULTS

#### 3.1 Data trimming

In order to assemble the Herring transcriptome, we performed a preliminary analysis on sequenced reads. Figure S1-4 show the average quality value of sequenced nucleotides and their distribution on raw reads.

**Table 3- Library information before and after trimming.**

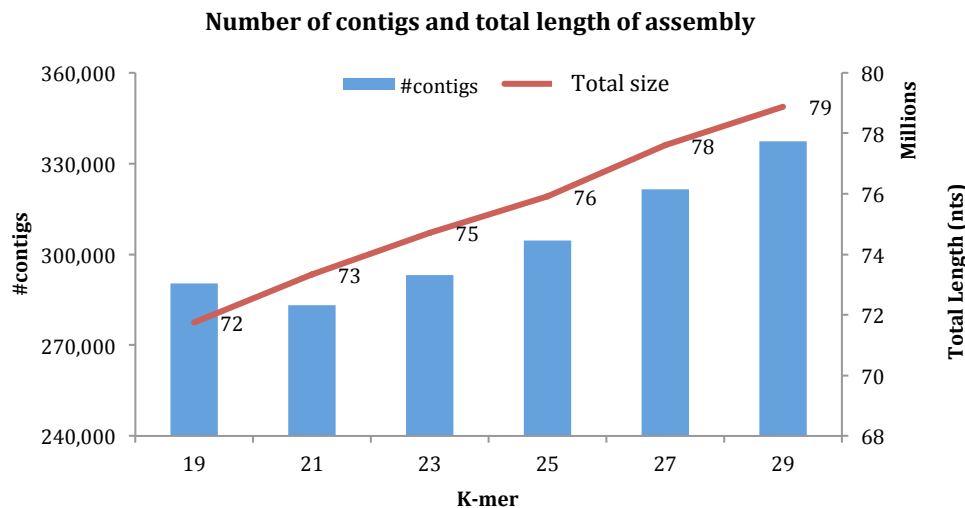
	Original data			End-trimmed data			SeqClean end-trimmed data		
	Length (nts)	Number	Total length (nts)	Length (nts)	Number	Total length (nts)	Length (nts)	Number	Total length (nts)
Read1	101	58,033,226	5,861,355,826	76	58,033,226	4,410,525,176	76	49,059,606	3,728,530,056
Read2	101	58,033,226	5,861,355,826	76	58,033,226	4,410,525,176	76	49,121,812	3,733,257,712
Total		116,066,452	11,722,711,652		116,066,452	8,821,050,352		98,181,418	7,461,787,768

From the plotted distribution of the four nucleotides per position in the reads (Figure S3 and S4) we observed a deviation from the random distribution (25% per nucleotide) at positions 1-12. This region represents the site

that was primed by random hexamers used in the reverse transcription process during sample preparation.

#### 3.2 Transcriptome assembly

The first assembly attempt was done with Inchworm (Appendix 8.6.2) with differently sized k-mers and using end-trimmed reads. By increasing the k-mer value, the number of generated contigs grew and the total size of the assembly showed an upward trend reaching to 79 million nucleotides (Mnts) in the k-mer 29 analyses (Figure 13).



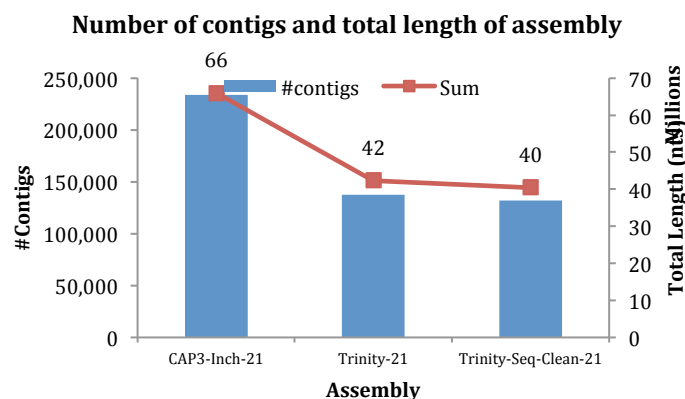
**Figure 13- The number of contigs and total length of assembly using Inchworm with k-mers varied from 19 to 29.**

The highest N50 was observed with the k-mer 21. Also, the maximum length value between the assembled contigs was lowest in the assembly with k-mer 21 (21,413 nts). Based on this statistics, the assembly with k equal to 21 (hereafter called Inch-21) was chosen as the best assembly to improve in following steps. Moreover, our optimal k-mer calculation showed that a k-mer of 21 produces the highest number of unique k-mers -with lower computational costs and time- compared to other k-mers (Appendix 8.3).

In order to improve Inch-21, we implemented two approaches: (1) scaffolding, using SSPACE and (2) merging contigs with similar overlapping ends by CAP3. SSPACE used paired end information to bridge between related contigs (Appendix 8.6.3) but did not show any significant improvement and padded unknown nucleotides (Ns) in the resulting contigs because of the scaffolding process. CAP3 could merge contigs with identical overlapping sequences using default parameters (Appendix 8.6.4) and improve the Inch-21 assembly to a higher N50 value (385 nts) and lowered the number of contigs from 283,113 to 233,732 contigs and singletons<sup>1</sup> together, the total length of the CAP3 assembly decreased by seven Mnts, approximately.

Since CAP3 appeared to improve the Inch-21 assembly, we concatenated all the contigs from different assemblies (Inch-19 to Inch-29) together in a file and used it as an input for CAP3. The results of this new merged assembly were promising because N50 increased to 941 nts and the total size was almost the same size of assembled transcriptome with k-mer 29 (79 Mnts). However, the assembly contained a considerable amount of chimeric structures, putting different genes in one sequence especially in large contigs. Posterior clustering efforts, in order to remove and separate these artefacts, were not successful. Also, in our contigs we detected some unrelated sequences including vectors, primers, and adaptors (Additional file 1). These sequences should be removed before using the reads for the transcriptome assembly. Thus, we decided to trim these sequencing features off the raw reads with SeqClean (Appendix 8.6.1) using the UniVec database (Table 3). While this analysis took place, Trinity was released (released 13<sup>th</sup> of March 2011).

We ran two new assemblies with Trinity and a k-mer size of 21, the first using end-trimmed reads and another with SeqClean end-trimmed reads (Table 3) (see appendix 8.6.5 for Trinity command). The numbers of contigs as well as the total length of the assembly decreased dramatically compared to CAP3-Inch-21 assembly (Figure 14). In addition to number of contigs, The N50 length increased from 385 nts to 421 nts for end-trimmed read assembly and to 419 nts for SeqClean end-trimmed read assembly. Trinity removed the chimeric contigs consisting of two different transcripts but the lengths of the largest contigs were almost halved and the transcriptome size decreased to 40 Mnts with a read depth of approximately 142x. As a result, Trinity produced a better assembly by using paired end information based on the added modules (Chrysalis and Butterfly).



**Figure 14- Number of contigs and total length of different assemblies.**

Before specific transcriptome assembly software had been developed, genome assemblers were used for most transcriptome assembly studies. In order to compare how genome assemblers would perform compared to Inchworm and Trinity, we assembled the herring transcriptome with SOAPdenovo, a genomic assembler (Appendix 8.6.6). Table 4 shows statistics of this assembly

<sup>1</sup> In assemblers with *de Bruijn* graph algorithm, the singletons are referred to contigs that were not merged together to build scaffolds due to lack of supporting paired end information whereas in assemblers with OLC algorithm, the singletons are contigs without any overlapping similarities.

beside all our other assemblies and their improvements after post-processing steps with SSPACE and CAP3.

**Table 4- Transcriptome assemblies with different k-mer size and software.**

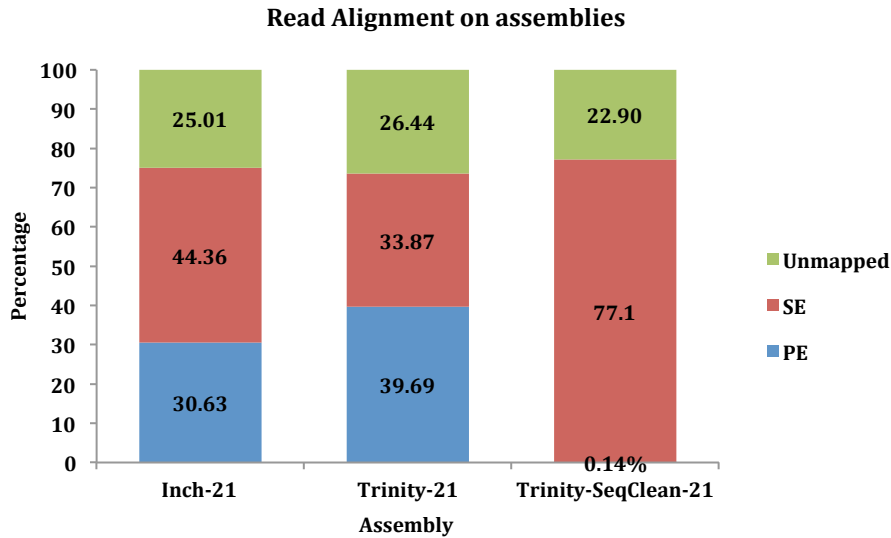
Input	K-mer	#Contigs	Sum	N50	N90	Min	Median	Mean	Max	GC%	#N
End-Trimmed	<b>Inch-19</b>	290,416	71,741,440	287	117	100	151	246	40,461	51.21	<b>0</b>
	<b>CAP3-Inch-19</b>	272,390	69,209,528	304	118	100	153	254	40,461	51.10	0
	<b>Inch-21</b>	283,117	73,343,165	319	117	100	149	259	21,413	51.13	<b>0</b>
	<b>SSPACE-inch-21</b>	280,460	73,344,103	326	117	100	148	261	21,413	51.13	8,335
	<b>CAP3-Inch-21</b>	233,732	66,012,398	385	120	84	155	282	21,413	50.82	<b>0</b>
	<b>SSPACE-CAP3-inch-21</b>	231,365	66,085,706	397	120	100	155	285	21,413	50.82	8,832
	<b>Inch-23</b>	293,123	74,709,284	309	116	100	146	254	35,818	51.13	<b>0</b>
	<b>CAP3-Inch-23</b>	226,666	64,741,183	399	120	92	153	285	35,818	50.70	<b>0</b>
	<b>Inch-25</b>	304,666	75,919,293	293	115	100	144	249	63,566	51.15	<b>0</b>
	<b>CAP3-Inch-25</b>	226,392	64,164,454	396	119	100	152	283	63,566	50.65	<b>0</b>
	<b>Inch-27</b>	321,484	77,610,121	274	114	100	142	241	44,562	51.23	<b>0</b>
	<b>CAP3-Inch-27</b>	225,062	63,375,072	392	119	73	151	281	44,562	50.63	<b>0</b>
	<b>Inch-29</b>	337,270	78,880,555	257	112	100	139	233	59,175	51.26	<b>0</b>
	<b>CAP3-Inch29</b>	225,769	62,731,667	380	118	87	150	277	59,175	50.63	<b>0</b>
	<b>All-cap3</b>	165,676	78,938,639	941	173	75	221	476	23,593	50.52	0
	<b>SOAPdenovo-21</b>	168,311	41,720,868	308	121	100	161	247	5,128	49.42	531,798
<b>Trinity-21</b>	137,708	42,379,085	421	132	100	177	307	9,497	49.77	0	
SeqClean	<b>SOAPdenovo-21</b>	85,575	29,084,678	477	142	100	215	339	7,141	50.70	135,437
	<b>Trinity-21</b>	131,788	40,342,038	414	132	100	178	306	10,762	49.94	0

SOAPdenovo generated a lower number of contigs than Inchworm and higher than Trinity using end-trimmed data. On the other hand, when using SeqClean reads, SOAPdenovo assembled a lower number of contigs. In both cases the assembled transcriptome size was almost half of that obtained with Trinity, including considerable amount of unknown nucleotides (N). Also, the largest contig measured showed a remarkable difference compared to both Inchworm and Trinity outputs. Hence, we conclude that Trinity generated a more reliable assembly.

Also, we compared the assembler performance based on the unique k-mer frequency using jellyfish software (Appendix 8.6.7). The results of this analysis showed that the generated transcriptome by Trinity and Inchworm have the maximum unique k-mer (100%) or the least repetitive sequences in assembly while SOAPdenovo transcriptome included more repetitive sequences.

In order to assess the single-base accuracy of the assembled transcriptome, we realigned all reads on the assembled transcriptome by SOAPaligner (Appendix 8.6.8). Between 73.53% and 77.24% of reads were aligned with the maximum proportion of reads (77.24%) being aligned to the Trinity-SeqClean-21 assembly (Figure 15). In comparison to other assemblies, most aligned reads on Trinity-SeqClean-21 assembly were single end.





**Figure 15-** The proportion of aligned reads on contig using SOAPaligner.

As a control in sequencing, all Illumina lanes were spiked with phiX, a circular genome with the length of 5.4 kb. In a successful assembly, the phiX genome should be assembled in one contig. Trinity assembled the phiX genome in one contig (comp82\_c0) while Inchworm, SSPACE, and CAP3 had partially assembled this genome in different contigs.

A high number of contigs in our assembly (57.10%) were smaller than 200 nts and we needed to remove redundant sequences. For this aim, two different clustering software were used: CD-HIT (Appendix 8.6.9) and UCLUSTAL (Appendix 8.6.10). Trinity also clusters similar sequences based on the *de Bruijn* graph and puts them together into one component. The number of components was 118,411 while the two other programs did not show any clustering performed (131,788 contigs). The multi-contig components represent transcripts from duplicated (paralogous) genes, splice variants, or redundancy. The number of members within clusters of Trinity varied between 1 and 8. However, still the number of components is considerably high and despite the fact that the clustered members should have similarities, comparisons on components with 8 members did not show any significant similarities among them and some of them had hits to different genes. Nonetheless, annotation results could help to validate a proportion of assembled contigs without any redundancy.

### 3.3 Gene annotation

For this part we used BLAST (Appendix 8.6.11) to align assembled contigs from Trinity-SeqClean-21 to various databases.

#### 3.3.1 BLAST on available herring genes

Among 66 non-mitochondrial genes available at the NCBI database, 41 subjects had alignment with transcriptome contigs with average identity of 94.59% showing proper assembly. However, among these subjects, 4 genes were unexpectedly aligned to different rRNA subunits.

One of the transcripts (comp20\_c0) had almost complete alignment to 18s rRNA (X98845.1) with 98.62% identity and another to 28s rRNA (c1, D1, and c2 Domain) with 99% identity covering 98% of this query. In addition, two other contigs had partial alignment to 28s rRNA (c1, D1, D2, and c2 Domain). The rest of hits were related to muscle specific genes such as alpha actin (GQ455648 for fast myotomal muscle and EF495203.2 for slow myotomal muscle) with identity of 99.74% and 87.23%, respectively (Additional file 2).

Also, we checked the quality of our transcriptome assembly by comparing our contigs to mitochondrial protein-coding genes available in GeneBank. There are 13 mitochondrial genes that were assembled in five contigs and this fusion is because of the compact structure of mitochondrial genome (Table 5).

**Table 5- Contigs that covered all 13 mitochondrial protein-coding genes.**

Contig	Gene Name	GeneID Acc_ID	Identity%
comp212_c0_seq1	NADH dehydrogenase subunit 1	gi 148762521 dbj BAF63971.1	96.91
comp212_c0_seq1	NADH dehydrogenase subunit 2	gi 148762522 dbj BAF63972.1	94.10
comp30_c0_seq1	Cytochrome c subunit 1	gi 148762523 dbj BAF63973.1	95.54
comp30_c0_seq1	Cytochrome c subunit 2	gi 148762524 dbj BAF63974.1	95.22
comp38_c0_seq1	NADH dehydrogenase subunit 6	gi 148762532 dbj BAF63982.1	97.63
comp38_c0_seq1	NADH dehydrogenase subunit 5	gi 148762531 dbj BAF63981.1	95.52
comp38_c0_seq1	Cytochrome c subunit b	gi 148762533 dbj BAF63983.1	94.74
comp40_c0_seq1	NADH dehydrogenase subunit 4L	gi 148762529 dbj BAF63979.1	96.94
comp40_c0_seq1	Cytochrome c subunit 3	gi 148762527 dbj BAF63977.1	95.02
comp40_c0_seq1	NADH dehydrogenase subunit 4	gi 148762530 dbj BAF63980.1	94.46
comp40_c0_seq1	NADH dehydrogenase subunit 3	gi 148762528 dbj BAF63978.1	92.55
comp85_c0_seq1	ATPase subunit 6	gi 148762526 dbj BAF63976.1	96.44
comp85_c0_seq1	ATPase subunit 8	gi 148762525 dbj BAF63975.1	90.91

### 3.3.2 BLAST on non redundant (nr) protein database

Transcriptome contigs were aligned to the nr database of protein sequences by BLASTX and returned 43,081 (32.69% of all contigs) unique hits with E-value less than 1.0E-5. 33.88% of the hits had strong homology (E-value < 1.0E-50) and the rest had a homology varying between 1.0E-5 and 1.0E-50. The unique hits, with identity score ranging from 16% to 100%, corresponded to 24,669 subjects in nr database and 13.85% of them had an identity bigger than 90%. 16,750 subjects were represented by single contig and 7,919 had hits with more than one contig. The most frequent hit (67) was related to the Titin (TTN) protein (XP\_001923800) from zebrafish. TTN is a giant protein with 360 exons coding for 32,757 aa and it is responsible for passive elasticity of muscle (Steffen et al., 2007). This protein has more than 300 repeated conserved domains where the immunoglobulin domain is the most frequent and conserved one (Garcia et al., 2009).

Because of the short length of queries and lack of genomic information, most contigs did not have any significant hit. The query length was crucial in alignment on nr subjects. As shown in figure 16 only 26.66% of queries shorter than 500 had significant hit while this value increased considerably to more than 90% in contigs larger than 1,500 nts.

### Effect of query length on BLAST matches

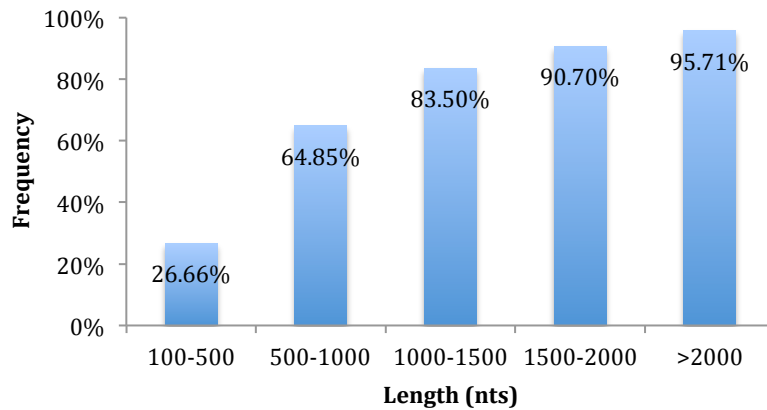


Figure 16- Effect of query length on percentage of matches with cut-off value of 1,0E-5.

Most of the best hits (74%) were from other fish and among them zebrafish had the highest proportion (43.58%) of similar transcripts to Atlantic herring transcripts (Figure 17). On average, aligned regions covered 59.81% of contigs' length consisting of 27.7% with coverage less than 60% and 72.3% with coverage more than 60%.

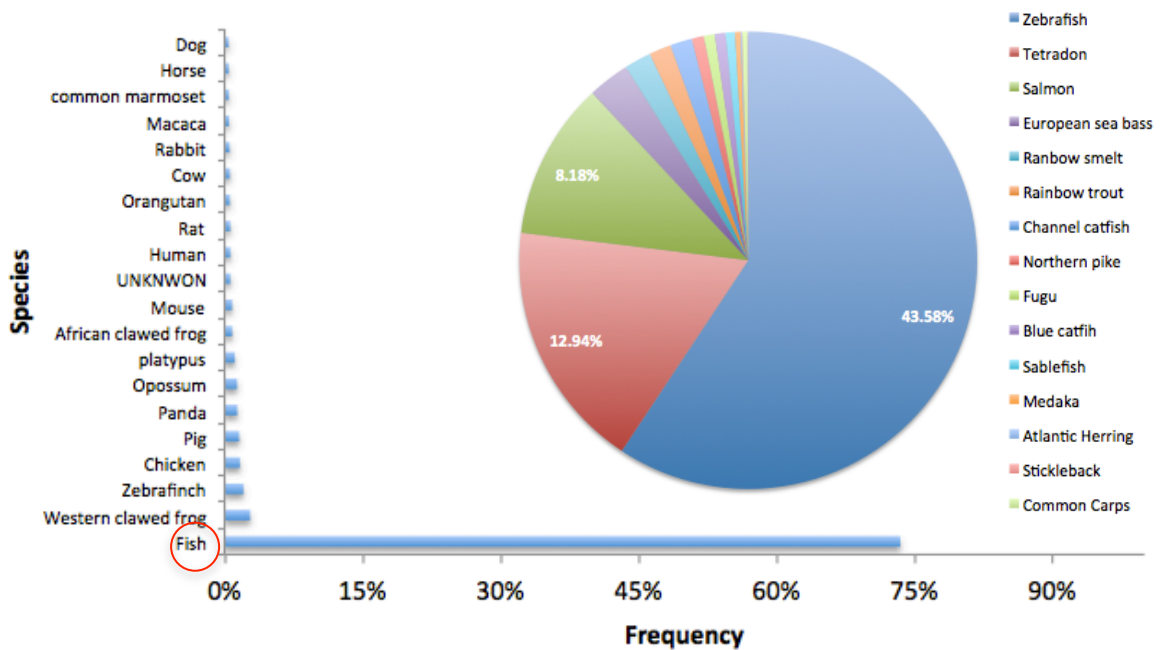


Figure 17- Bar chart) Species distribution of the best hits from nr database. Pie chart) species distribution among fish.

### 3.3.3 BLASTX against conserved domain database (CDD)

Alignment on CDD resulted in 159,320 hits with E-value less than 1.0E-5 from which, 12.90% of our contigs had similarity to 5,988 conserved domains in CDD (Additional file 3). The most frequent domain was cd00204, ankyrin (ANK) repeat. Ankyrin is a 33 amino acids tandem repeat occurring in diverse proteins in eukaryotes and involves in protein-protein interactions (SMART). Muscle ankyrin protein family (MARF) has key role in muscle function and structure, however the activity of this family is not accurately known. MARFs affect gene expression program of skeletal muscle, particularly in response to stimuli and injuries (Barash et al., 2007).

### 3.3.4 BLASTN on *Oryzias latipes* (medaka) UniGene and *Danio rerio* (zebrafish) muscle ESTs

43,081 contigs had an alignment to the nr database but all genes have not been annotated and included in this database. Therefore, we aligned the rest of our contigs (88,707) to medaka ESTs and zebrafish muscle ESTs. In alignment against medaka ESTs, we received 15,002 significant hits with E-value less than 1.0E-5 from which 770 contigs had unique hits with average identity of 90.81% ranged from 79 to 100%. For zebrafish, among 12,542 significant hits, 827 contigs had unique hits with average identity of 91% varied between 80% and 100%. In total, 1,597 contigs had non-redundant hit on these datasets that was not present among any nr hits before.

### 3.3.5 Finding full length sequences

We used a web tool, Full-Lengther, to detect contigs corresponding to 5' and 3' sites of a protein from un-annotated contigs. This analysis resulted in 16.8% full-length and 1.48% putative length from un-annotated contigs. In total, with other unique hits from other sources, 47.09% of all contigs were reliable (Table 6).

**Table 6- Reliable number of contigs and their proportion resulted from BLAST against nr, medaka ESTs, zebrafish ESTs, and Full-Lengther.**

Source	#Contigs	Proportion of contigs (%)
nr (protein)	43,081	32.69
Full-lengther (Full-Length/Putative)	17,391	13.20
medaka (EST)	770	0.58
zebrafish muscle (EST)	827	0.62
Total	62,069	47.09

### 3.3.6 Alignment on eukaryotic tRNAs

In total 62,069 contigs were collected as reliable transcripts. In order to check the efficiency of mRNA purification we aligned remained contigs (61,719) against eukaryotic tRNAs. The outcome of this alignment was 17 contigs with average identity of 96%.

## 3.4 Gene expression

FPKM and RPKM are parameters representing the relative expression of each contig/transcript among all the generated contigs/transcripts in the final assembly. FPKM is based on paired reads alignment but as we only received 0.14% of total reads in paired end alignment when realigning the reads on Trinity-SeqClean-21 assembly. Therefore, we calculated RPKM for each contig by formula 1. Figure 18 shows the contigs length and the RPKM values distribution (in logarithmic scale) that ranged between 0.06 and 43,394.54 with average of 6.66.

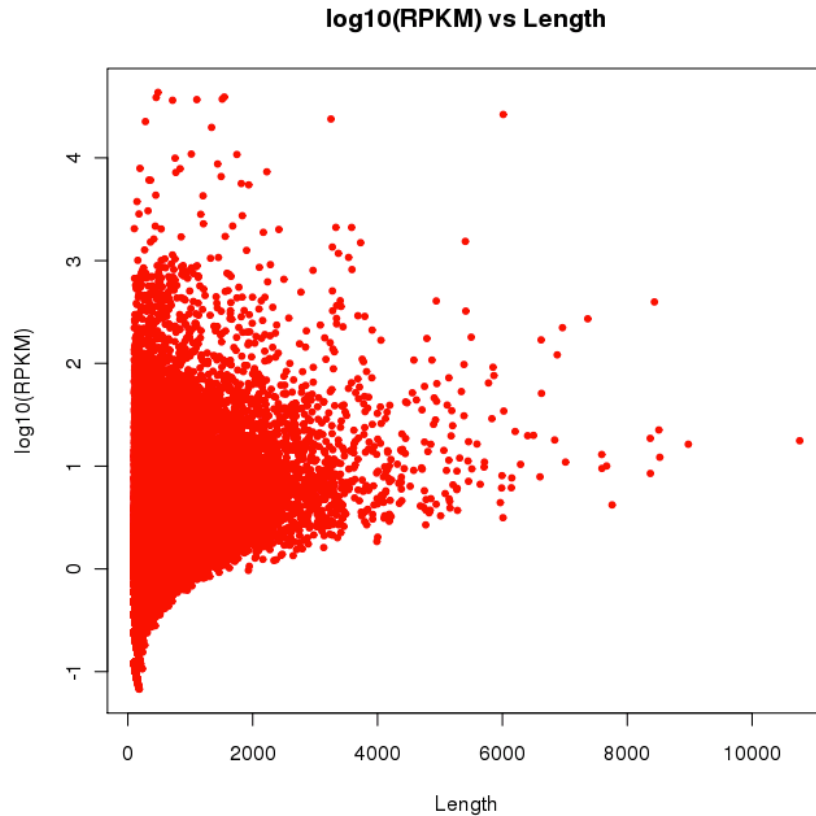


Figure 18- log<sub>10</sub> (RPKM) values versus length of contigs.

The information of highly expressed transcripts ( $\log(\text{RPKM}) \geq 4$ ) along their significant hit on nr and their annotations are summarized in table 7.

Table 7- Highly expressed transcripts and their information from BLASTX on nr.

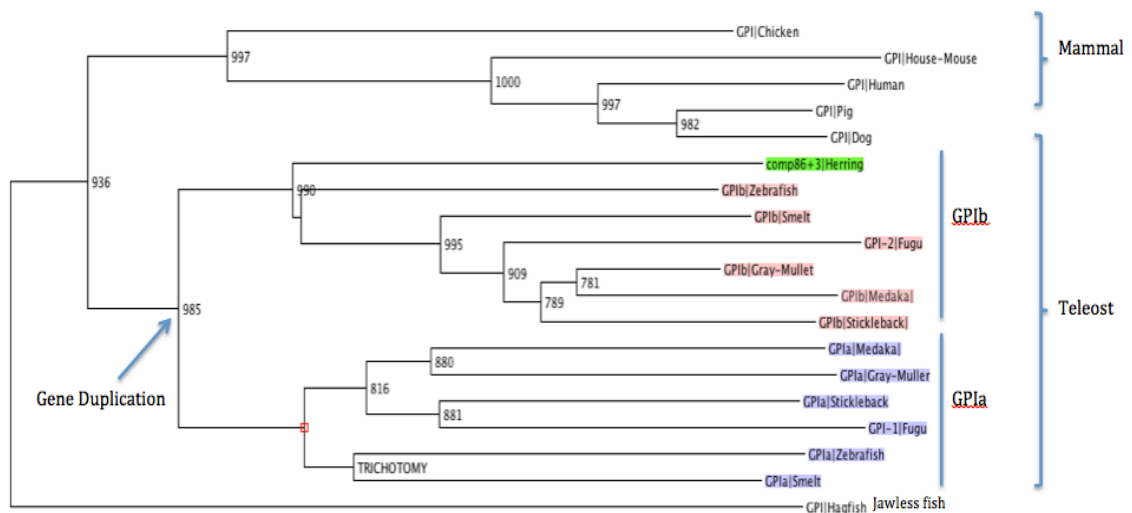
Contig	RPKM	Length (nts)	nr hit
Comp0_c0	43,394.54	483	mCG146230 [Mus musculus]
Comp1_c0	39,100.59	1,545	actin <sup>§</sup> [ <i>Danio rerio</i> ]
Comp2_c1	38,669.11	453	muscle creatine kinase b [ <i>Danio rerio</i> ]
Comp3_c0	37,306.62	1,512	Fructose-bisphosphate aldolase A [ <i>Dicentrarchus labrax</i> ]
Comp0_c3	36,905.84	1,104	muscle-type creatine kinase CKM1 [ <i>Pagrus major</i> ]
Comp4_c0	36,344.59	716	creatine kinase [ <i>Salmo salar</i> ]
Comp5_c1	26,436.65	6,013	myosin heavy chain 4 [ <i>Danio rerio</i> ]
Comp7_c0	23,885.42	3,253	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1 [ <i>Makaira nigricans</i> ]
Comp6_c0	22,567.10	281	Ckmb protein [ <i>Danio rerio</i> ]
Comp11_c0	19,790.14	1,341	glyceraldehyde 3-phosphate dehydrogenase [ <i>Oncorhynchus mykiss</i> ]
Comp15_c0	10,888.38	1,016	fast mytomal muscle troponin T isoform b (TNNTb) [ <i>Salmo salar</i> ]
Comp9_c0	10,782.95	1,746	Tropomyosin [ <i>Epinephelus coioides</i> ]

<sup>§</sup> Find the explanation about actin in section 3.6.

### 3.5 Glucose 6-phosphate isomerase (GPI)

A previous study showed that herrings are polymorphic at GPI locus (Andersson et al., 1981). Therefore, we chose to study this gene in order to detect possible variations. In fish, one to five loci code this gene but mostly two genes are responsible for GPI production (reviewed in Henry and Ferguson 1987). These genes are GPIa (553 aa) and GPIb (553 aa) in teleost fish. In zebrafish, pairwise comparison between GPIa and GPIb showed 77.6% identity on nucleotide level and 82.4% on peptide (Lin et al., 2009). We found a contig (comp86\_c0\_seq1: 2,106 nts long) with 79.67% identity to zebrafish GPIa (ensemble acc-id: [ENSDART00000022437](#)). In order to identify whether our contig is GPIa or GPIb we performed a MSA among teleost fish and some other organisms with Clustalx. We translated comp86\_c0 to amino acid sequence with ORF-finder resulting in a protein sequence with a length of 553 amino acids (ORF+3). We retrieved other proteins from the Ensemble genome browsers (Find accession IDs in additional file 4).

MSA revealed that most teleost fish encode proteins with a length of 553 aa, with the exception of fugu GPIb and European eel GPIa having a deletion at position 553. Lin and colleagues (Lin et al., 2009) had reported this deletion in fugu GPIa and smelt GPIb. Among all amino acid sites, 325 sites (58.77%) were identical and all of active sites were evolutionary conserved across all species including residues interacting with phosphate substrate binding (Ser159, Ser209, Lys210, Thr211, and Thr214) and residues participating in catalytic activity (Glu216, Arg272, His388, and Lys518) (Additional file 5). The phylogenetic analysis showed that GPIa and GPIb were clustered separately and highlights the gene duplication occurred in teleost fish as compared to mammals. The contig comp86\_c0 clustered with GPIbs with high bootstrapping value (990), meaning that herring transcript assembled is evolutionary close to GPIbs (Figure 19).



**Figure 19- Phylogenetic tree of GPI amino acid sequences. Bootstrap values greater than 50% have been indicated next to their branches. Red: GPIb, Purple: GPIa, Green: Comp86\_c0 ORF (+3).**

We realigned reads on comp86\_c0 by MosaicAligner (Appendix 8.6.12) and made a pileup file using samtools (Appendix 8.6.13). Afterward, we called SNPs/indels by VarScan software (Appendix 8.6.14). Since the transcriptome was generated from a single fish we expected to find about 50% of the reads from each allele at heterozygous positions. Therefore, we focused on frequencies close to this value and p-value less than 0.05. Based on these cut-off values, we detected four SNPs on this contig (Table 8). Two of them were located in the ORF region and one of them represented a non-synonymous substitution at position 351 that replaced Asparagine (N)

by Aspartic acid (D). Other variations were located in 5' (position 53) and 3' (position 2,082) UTRs.

**Table 8- Detected SNPs in comp86\_c0 (herring GPIb).**

	Position	Ref	Reads1	Reads2	VarFreq	Pvalue	VarAllele	Original (+3)	Variant (+3)
comp86_c0	53	A	1,630	1,158	41.54%	0	C	NOT IN ORF	NOT IN ORF
comp86_c0	351	A	3,842	3,357	46.63%	0	G	AAC(N)	GAC(D)
comp86_c0	686	C	3,360	2,983	47.03%	0	T	TTC(F)	TTT(F)
comp86_c0	2,082	G	375	274	42.22%	2.95E-99	A	NOT IN ORF	NOT IN ORF

### 3.6 Alpha actin

Contig comp1\_c0 had almost complete identity with the already annotated herring alpha actin in fast myotomal muscle (GQ455648.1) and high identity with slow myotomal alpha actin (EF495203.2). By ORF-finder we found an ORF (-3) with a length of 377 aa. We used this sequence with alpha-skeletal actins (ACTAs) and alpha-cardiac actins (ACTCs) from human, zebrafish, nematode, snake, mouse, and chicken for MSA. Surprisingly, the MSA analyses with zebrafish actins showed identical alignment with alpha-cardiac actin (cfk, AAO38846.1 and ACTC, CAQ15402.1). We compared these genes in Ensembl and NCBI and observed that they differ in two residues, 239 and 351. Therefore, we performed another MSA with a new set of sequences (for accession IDs see additional file 4) resulted in identical alignment with zebrafish ACTA1b.

Alpha actin has three conserved domains: ATP binding with six active sites, Gelsolin with nine active sites, and Profilin with 11 active sites. The MSA showed that all of these active sites have been conserved across all species.

As GPI we realigned reads on contig comp1\_c0 with the same pipeline and parameters to find SNPs. Results showed five synonymous mutations in the ORF region and a three nucleotides deletion in the beginning of the contig corresponding to 3' UTR region of this gene (Table 9).

**Table 9- Detected SNPs/indel in comp1\_c0.**

Chrom	Position	Ref	Reads1	Reads2	VarFreq	Pvalue	VarAllele	Original (-3)	Variant (-3)
Comp1_C0	371	G	3,999	3,995	49.97%	0	A	GAC(D)	GAT(D)
Comp1_C0	374	G	3,993	3,992	49.99%	0	A	TAC(Y)	TAT(Y)
Comp1_C0	893	C	4,618	3,377	42.24%	0	G	CTG(L)	CTC(L)
Comp1_C0	1,130	C	3,813	4,180	52.30%	0	G	CTG(L)	CTC(L)
Comp1_C0	1,133	G	3,745	4,242	53.11%	0	T	CCC(P)	CCA(P)
Comp1_C0	157	T	7,032	2,672	38.00%	0	-CAA	NOT IN ORF	NOT IN ORF

## 4 DISCUSSION

RNA-Seq is an efficient and cost-effective approach to characterize a transcriptome, especially in non-model organisms. This is the first attempt to study the transcriptome of Atlantic herring fish using NGS technology. Previous studies have been mainly based on a set of genes to characterize the population structure in different stocks of herring. In this study the herring transcriptome was assembled in parallel to the genome assembly, with the purpose of obtaining concise information about this sequenced individual. These results will be used for quantification of variation in pools of herring from different regions, based on former studies by Andersson and colleagues (1981) and Ryman and colleagues (1984).

We received 116,066,452 reads with total length of 11,722,711,652 nts produced by Illumina Hiseq 2000. 25 nucleotides in the end of the reads had low quality was therefore trimmed. Furthermore, the first 12 nucleotides at 5' ends indicated a biased distribution compared to



other positions along reads (Figure S3 and S4). There are two possible reasons for this feature: first, this length corresponds to the random primer that is used in sequencing and these primers are either pseudorandom or have some favoured sites. Second, RNAs form a secondary structure at room temperature inhibiting the priming (Mortazavi et al., 2008). Although this distribution deviated from the random distribution, it contained some biological information of transcripts. Therefore, we used reads without trimming their 5' ends.

Our data had some contaminations of vectors, adaptors, and primers that are possibly due to introduced errors during library preparation or sequencing. The presence of primers in raw data is because of primer dimer formation. Consequently, suspicious reads were filtered out and we used cleaned reads to assembly.

We applied various strategies using new tools to generate transcriptome assembly. Initial results from Inchworm and the study on optimal k-mer showed that the best size is 21; hence posterior assemblies were performed by this value<sup>1</sup>. However, large contigs in Inchworm assemblies was found to consist of several unrelated transcripts. Two new modules in Trinity (Chrysalis and Butterfly) resolved afterwards these chimeric contigs by using paired end information.

The number of unique k-mer is an indicative of the quality of the assembly; the higher proportion the better assembly. In other word, this value shows how many times a k-mer has occurred in transcriptome/genome. Trinity and Inchworm showed that all of the possible k-mers in assembly are unique while SOAPdenovo had lower proportion of unique k-mers.

There are some evidences that highly support the good quality of our transcriptome assembly, including assembly of all mitochondrial genes, contigs alignment on 70% of previously annotated herring genes in NCBI with average identity of 94%, and assembly of phiX genome in one contig. Additionally, 32.7% of contigs had unique hit to nr database that is close to annotated proportion of other fish transcriptome studies (European eel, Southern platyfish and guppy) (Coppe et al., 2010, Zhang et al., 2001, and Fraser et al., 2011). The coverage that we have in this study is remarkably higher than other *de novo* transcriptome studies such as those for viviparous eelpout, lake sturgeon, rainbow trout, guppy, and European eel (Kristiansson et al., 2009, Coppe et al., 2010, salem et al., 2010, and Fraser et al., 2011). One reason for this difference is that Illumina sequencing technology generates high number of reads while in aforesaid studies 454 technology were used that generates longer reads but less number compared to Illumina.

The result of alignments on eukaryotic tRNA resulted in a few number of contigs but with higher average read depth than the whole transcriptome. Most of these hits were mitochondrial tRNAs and assembly of them together with other mitochondrial transcripts is because of the compact structure of mitochondrial genome. To confirm the efficiency of ploy-A selection of our sample, further studies are required for other kinds of non-coding RNAs.

In general, 47.09% of our contigs have been annotated or validated, based on their comparison to nr, medaka ESTs, zebrafish muscle ESTs, and contigs that were reported by Full-Lengther (this set matching is hereafter called known). Table 10 lists the information about these contigs and the other group without any reliable result from alignment against aforesaid databases (so-called unknown).

---

<sup>1</sup> Study on optimal K-mer showed that k-mer size is correlated to genome/transcriptome size. However, following studies is demanding by taking sequencing depth into account.



**Table 10- validated and non-validated contigs from Trinity-SeqClean-21 assembly.**

Subset	#Contigs	Sum	N50	N90	Min	Median	Mean	Max	GC%	Coverage
<b>Known</b>	62,069	27,665,725	685	183	100	264	445	10,762	51.50	201.5
<b>Unknown</b>	69,719	12,676,313	178	116	100	147	181	3,648	46.58	15.5

The known contigs subset has higher N50, N90, mean, and total size compared to the unknown subset. Also, the GC content is significantly lower in the unknown subset indicating these contigs possibly have a higher proportion of non-coding sequences. Comparison between cold-blooded and warm-blooded vertebrate on mosaic structure revealed that GC content in introns is significantly lower than exons (Melodelima and Gautier 2008 and Fortes et al., 2007). As a result, these contigs can be introns due to either retained intron features or they are from pre-mRNA before removing introns by spliceosome (Brown, 2002). Also, all the genes have not been annotated and some of these contigs could be novel transcripts.

In addition, unknown contigs can be UTRs of transcripts since the alignment between unknown and known subsets resulted in some significant hits showing some overlaps between contigs. There are two explanations why UTRs may have not been combined with the rest of related transcripts in known subset. First, after trimming we lost some part of our data containing paired information between reads and due to lack of this information, related contigs could possibly not combine together. Also, this is supported by the low proportion of paired end alignments contained in Trinity-SeqClean-21 compared to other assemblies (Figure 14). Second, although we modified the k-mer size from 25 to 21 for Inchworm, it seems that next modules (Chrysalis and Butterfly) continue by k-mer size of 25 as a default value and in these steps a minimum alignment size of 25 is required while the assembly was performed by a k-mer size of 21.

Nonetheless, the improvement in transcriptome assembly by Trinity was remarkable, in terms of better N50 (414 nts) and total size (40 Mnts). All available teleost fish transcriptome are whole body transcriptome and their size is larger compared to our assembly but the total size of herring muscle transcriptome is in vicinity to medaka and stickleback and almost half compared to zebrafish and fugu (Table 11). The transcriptome is larger in the highly duplicated genome of zebrafish and the alternative splicing events (AS) are less than other teleost fish, i.e. AS number show an inverse correlation to the zebrafish genome size. For instance, alternative splices per gene is 17% in zebrafish while in medaka and fugu it is 31,2% and 43,2%, respectively (Lu et al., 2010).

**Table 11- Transcriptome assembly of some teleost fish.**

	#Contigs	Transcriptome size (nts)	N50 (nts)	Maximum (nts)	GC%
Zebrafish <sup>a</sup>	47,138	87,699,104	2,518	93,957	46.26
Medaka <sup>a</sup>	24,662	38,241,640	1,883	78,426	49.63
Stickleback <sup>a</sup>	27,628	45,677,827	1,933	78,390	55.03
Fugu <sup>a</sup>	48,003	91,869,840	2,370	65,562	53.95
Guppy <sup>a</sup>	54,987	25,534,864	846	3,571	NA <sup>b</sup>
Atlantic herring <sup>c</sup>	131,788	40,342,038	414	10,762	51.50

<sup>a</sup> Whole body transcriptome

<sup>b</sup> Not available

<sup>c</sup> Muscle transcriptome

Butterfly can identify isoforms of transcripts but in our assembly it failed to detect any of them. It can be due to the k-mer setting previously described. One of the complementary modules developed by M. Grabherr (Broad Institute) can detect these features, though it has not been released till now.

RNA-seq enable us to quantify the expression of a gene more precisely than microarrays. In this study we used RPKM parameter for this quantification resulted in some genes that are highly expressed in muscles (Table 7). Among these transcripts we had hit to mCG146230, a sequence from the mouse whole genome shotgun sequencing project that had partial similarity to human creatine kinase, one of the key enzymes in skeletal muscle (CKM) and other tissues that consume ATP rapidly. This gene family is responsible for transferring phosphate from creatine phosphokinase to ADP (Wallimann et al., 1992). There are at least four genes in this family: cytoplasmic, skeletal muscle, cardiac muscle, and brain specific (reviewed in Harder and McGowan 2001). In total, five contigs (comp0\_c0, comp0\_c3, comp2\_c1, comp4\_c0, and comp6\_c0) had significant alignment on CKM (Table 7). We aligned all of these contigs on zebrafish genome at UCSC genome browser and found hits for CKMa (comp0\_c0, comp0\_c3, comp2\_c1, comp4\_c0, and comp6\_c0) located on chromosome five and CKMb (comp0\_c0, comp0\_c3, comp4\_c0, and comp6\_c0) located on chromosome 15 (Figure S5).

The reason that these contigs were not combined in one contig might be due to different copies of this gene and that they are possibly representing different isoforms of CKM. Another explanation for the broad distributed sequence of CKM in multiple contigs is the lack of paired end information after trimming by SeqClean, since contaminated reads were removed from raw data. In other word, reads that built these contigs are mainly single rather than paired and it was also indicated in read realigning on all the contigs (Figure 14).

Fructose-bisphosphate aldolase was also among nr subjects that had high similarity (88%) with a transcript (comp3\_c0) with high RPKM value (37,306.62). This gene is a class of three aldolase classes and is responsible of converting fructose into dihydroxyacetone-phosphate and is one of the key enzymes in glycolysis and actin dynamics (St-Jean et al., 2007).

Likewise, myosin heavy chain belongs to a conserved multigene family and it provides energy sources for diverse motile activities including muscle contraction. It is found abundantly in skeletal muscle generating energy for muscle movement by ATP hydrolysis and its ATPase activity depends on speed of contraction (reviewed in Weiss et al., 1999). Contig comp5\_c1 had high similarity (83%) with this gene.

Ca<sup>2+</sup> is a crucial ion for muscle activity and many other cellular processes. Sarcoplasmic/endoplasmic reticulum calcium ATPase 1 mediates the sequestration of calcium for skeletal muscle relaxation (Adachi et al., 2004) and contig comp7\_c0 with high RPKM value (23,885.42) showed high similarity (85%) to this gene.

Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) is a housekeeping gene involved in basic cell maintenance biological function and contig comp7\_c0 was highly similar (92%) to this gene. This enzyme participates in glycolysis path that breaks down glucose for energy (Casadei et al., 2011).

Actin, tropomyosin, and troponin are core components of muscle contraction. Troponin t (TNNT) is one of the troponin complex subunits, which regulates muscle contraction by mediating Ca<sup>2+</sup>-dependent interaction between actin and myosin (Ferrante et al., 2011). TNNT is differentially expressed in cardiac, slow, and fast skeletal muscle but its main role is anchoring the complex to actin filament by tropomyosin (Fu et al., 2009 and Ferrante et al., 2011). Three transcripts (comp1\_c0, comp15\_c0, and comp9\_c0) with large RPKM values showed high similarities to these genes (Table 7).

To summarize, all these transcripts with high RPKM values, corresponds to specific genes expressed in muscles makes perfect sense for a muscle transcriptome.

GPI is a key enzyme in glycolysis catalysing the conversion of glucose-6-phosphate into fructose-6-phosphate. Beside this intracellular metabolic role it also has extracellular activities (reviewed in Lin et al., 2009). For instance, it acts as neurotropic factor when secreted by T-cells during embryonic development and it is a vital enzyme for embryo survival (Gurney 1986, Chaput et al., 1988). In phylogenetic analysis the herring transcript clustered with GPIbs. Also, GPIb is highly expressed in skeletal and cardiac muscle. It explains that our transcript is GPIb while GPIa is mainly expressed during embryo development (Lin et al., 2009).

One of the detected SNPs in GPIb indicated a non-synonymous mutation caused substitution of Asparagine by Aspartic acid. This mutation affects the isoelectric point (pI) of the protein, a parameter that had been previously used by Andersson et al. (1981) to detect different variant of GPI in herring fish by starch gel electrophoresis of tissue extracts.

We expect to find gene paralogs on different locations of genome. Although the genome assembly is not complete yet, we aligned transcripts corresponding to GPIb and actin on draft version of genome assembly by BLASTN. For GPIb, we found significant hits on two scaffolds (scaffold176474 and scaffold69157) and partial alignments on some singletons. The aligned region on scaffolds covered 33% and 41.12% of transcriptome contig, respectively. This low coverage is because of the high number of Ns in genomic scaffolds: 34.7% in scaffold176474 and 61.41% in scaffold69157. The alignment on some singletons were identical, in regions that scaffolds had Ns. Therefore, it indicates that we not only can identify the gene copies on the genome but also this kind of alignments may assist the genome assembly. As a result, the alignment on two different scaffolds corresponds to copies of this gene. We did the same for other transcript that had a high similarity to alpha actin but the result was not as informative as for GPI. After the genome assembly is completed we can address this issue about actin and other gene families.

Actin is a multigene family with six members: alpha-skeletal, alpha-cardiac, alpha-smooth muscle, gamma-smooth muscle, beta-cytoplasmic, and gamma-cytoplasmic. Actins are important for cell motility, cytoskeleton and muscle contraction (Bertola et al., 2008) and this family is one of the most well conserved families in vertebrates (Vandekerckhove et al., 1981). Our comparisons between zebrafish ACTA1bs annotations from Ensembl and NCBI indicated two mutations at positions 239 E>G and 351 L>P. The E substitution by G can affect the structure and isoelectric charge of the protein. Since G is the smallest amino acid a change to that amino acid may cause turns and breaks in the secondary structure. It is also a neutral amino acid while E is a hydrophilic amino acid and sits in helices (Murray et al., 2003). Therefore, based on this explanation and description of this entry on NCBI, we think it might be an erroneous annotation but we need to do further investigations.

The contig comp1\_c0 representing the alpha actin had two significant hits in the NCBI database on two annotated alpha actins in herring that both of them had a length of 377 aa but their nucleotide sequence were different. This transcript was identical with GQ455648 (fast myotomal muscle) and MSA analysis showed that this contig is identical with zebrafish skeletal alpha actin isomerase b (ACTA1b).

The spawning time affects the muscle development in embryonic stage and the temperature regime during this period can have a long-term effect on skeletal muscle growth in adults (Johnston 2006). Spring spawning embryos experience higher temperature and studies have shown that the density of myofibrils in the fast fibers is greater in these fish (Johnston et al.,

1997). Our sample is a spring spawning herring and therefore the transcript of alpha actin is identical to annotated fast muscle alpha actin.

## 5 CONCLUSION

The growing of genomic and transcriptomic resources of various species has opened doors to address fundamental questions about population structure, evolution, and adaptation. In this study we used high throughput Illumina Hiseq 2000 sequencing to assemble the Atlantic herring transcriptome by using Trinity, a specific transcriptome assembler. The results are not equally good to other *de novo* assembled transcriptomes by Trinity reported by Grabherr and colleagues (2011). There are some doubts about efficiency of mRNA preparation and sequencing but for at least 47% of the contigs we have high reliability based on comparison with available databases. In comparison to other *de novo* assemblies of fish transcriptomes, we found almost the same proportion of contigs with significant hits on annotated proteins. To obtain better assembly we suggest reassembling the transcriptome with Trinity and the default k-mer (25) or using the latest released version of Trinity (released 13<sup>th</sup> May 2011) in which we can run Trinity steps separately by setting the k-mer size independently.

In order to complete the annotation, we suggest performing functional annotation and detecting alternative splicing events. For functional annotation we recommend BLAST2GO and for AS detection there is a specific teleost fish AS database that can be used called ASpipe. We believe that after removing bottlenecks in this study, it has the potential to be considered as a unique approach for transcriptome assembly and genome annotation using NGS and finally setup a standard protocol for experiments dealing with Illumina data.

The result of this study would help in the further investigations of the following issues. First, revealing the true structure of the Atlantic herring population in the Baltic Sea. The primary analysis in 1980s by Ryman and colleagues presented differences between morphological and genetic classification (Ryman et al., 1984). Discovering the population structure of herring and associated markers to the different schools in the Baltic Sea would be helpful in governing herring fishing to control the population size and avoid overfishing.

## 6 ACKNOWLEDGMENTS

This thesis was a part of a big project and I presented just a small section of this study. Thus, it was teamwork and I would like to thank people in this group whom I worked with them at Uppsala University, KTH, and Science for life laboratory. Special thanks go to Prof L. Andersson who gave me this opportunity to work and research under his supervision, my co-supervisors, Dr. A. Martinez, and Dr. CJ Rubin. Additionally, I would like to thank Dr. G. Sundström who helped me in phylogenetic analyses. I also benefited by working with my colleague S. Lamichhaney, it was a nice experience to work in this group. I express my thanks to Aida Elhamian and other friends who helped me in this project. Finally, I would like to thank EM-ABG staff at SLU University Dr. B. Malmfors, D. Lont, and Prof. J. Van Arendonk the coordinator of EM-ABG at Wageningen University.

## 7 REFERENCES

- Adachi, T., Weisbrod, R.M., Pimentel, D.R., Ying, J., Sharov, V.S., Schöneich, C., Cohen, R.A., (2004). S-glutathiolation by peroxynitrite activates SERCA during arterial relaxation by nitric oxide. *Nat. Med.* **10**:1200–120.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P., (2007). *Molecular Biology of the Cell - Reference Ed.* Garland Science Publisher.
- Andersson, L., Ryman, N., Rosenberg, R., and Ståhl, G., (1981). Genetic variability in Atlantic herring (*Clupea harengus harengus*): description of protein loci and population data. *Hereditas* **95**:69-78.
- Barash, I.A., Bang, M., Mathew, L., Greaser, M.L., Chen, J., and Lieber, R.L. (2007). Structural and regulatory roles of muscle ankyrin repeat protein family in skeletal muscle. *Am J. Physiol Cell Physiol* **293**: C218–C227.
- Bertola, L.D., Ott, E.B., Griepsma, S., Vonk, F.J., and Bagowski, C.P., (2008). Developmental expression of the alpha-skeletal actin gene. *BMC Bioinformatics* **8**:166.
- Birol, I, Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E., Horsman, D.E., Connors, J.M, Gascoyne, R.D., Marra, M.A., Jones, S.J., (2009). De novo transcriptome assembly with ABySS. *Bioinformatics* **25(21)**:2872-7.
- Black, D.L., (2000). Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell.* **103(3)**:367-70
- Brown, T.A.,( 2002). *Genomes 2. Second Ed.* BIOS Scientific Publisher.
- Carlson, R., (2010). <http://www.synthesis.cc/2010/08/recent-dna-cost-and-productivity-figures-from-the-economist.html>. Accessed 2011 May 3.
- CBCB. [http://www.cbc.umd.edu/research/assembly\\_primer.shtml](http://www.cbc.umd.edu/research/assembly_primer.shtml). Accessed 2011 April 13.
- Chaput, M., Claes, V., Portetelle, D., Cludts, I., Cravador, A., Burny, A., Grast, H., Tartar, A., (1988). The neurotrophic factor neuroleukin is 90% homologous with phosphohexose isomerase. *Nature* **332**: 154–155.
- Chaisson, et al., (2009) Chaisson, M. J., D., Brinza, and P. A., Pevzner. 2009, De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res.* **19**: 336-346.
- Coppe, A, Pujolar, J.M., Maes, G.E., Larsen, P.F., Hansen, M.M., Bernatchez, L, Zane, L., Bortoluzzi, S., (2010). Sequencing, de novo annotation and analysis of the first *Anguilla anguilla* transcriptome: EelBase opens new perspectives for the study of the critically endangered European eel. *BMC Genomics*.**11**:635.
- Doucette, A. Jr., Fitzsimons, J. (1982). Karyology of the Ladyfish *Elops saurus*. *Jap. J. Ich.* **29(2)**: 223-226.
- Encyclopædia Britannica. <http://www.britannica.com/EBchecked/topic/127892/commercial-fishing>. Accessed 2011 April 27.
- Ensembl. <http://www.ensembl.org/index.html>.
- FAO Fishery Statistics. <http://faostat.fao.org/> Accessed 2011 April 30.
- Ferrante, M. I., Kiff, R. M., Goulding, D. A. and Stemple, D. L., (2011). Troponin T is essential for sarcomere assembly in zebrafish skeletal muscle. *J. Cell Sci.* **124**:565-577.
- FishBase. <http://www.fishbase.org>. Accessed 2011 April 30.
- Fortes, G. G., Bouza, Æ., C., Martí'nez, Æ. P., Sa' nchez, Æ. L., (2007). Diversity in isochore structure among cold-blooded vertebrates based on GC content of coding and non-coding sequences. *Genetica* **129**:281-289.
- Fraser, B. A., Weadick, C. J., Janowitz I., Rodd, H., Hughes, K. A., (2011). Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome. *BMC Genomics* **12**:202.
- Garcia, T. I., Oberhauser, A. F., and Braun, W., (2009). Mechanical Stability and Differentially Conserved Physical-chemical Properties of Titin Ig-domains. *Proteins* **75** (3): 706-716.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, Dawn A., Amit, I., Adiconis, X., et al, (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biol.* Online publication: doi:10.1038/nbt.1883

Gurney, M.E., (1986). Molecular cloning and expression of neuroleukin, a neurotrophic factor for spinal and sensory neurons. *Science* **234**: 566–574.

Harder, G. and McGowan, R., (2001). Isolation and characterization of the muscle-specific isoform of creatine kinase from the zebrafish, *Danio rerio*. *Biochem. & Cell Biol.* **79**:779:782.

He, L. and Hannon, G. J., (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews*, **5**:631.

Henry, T., Ferguson, A., (1987). Phosphoglucose isomerase isozymes and allozymes of the brown Trout, *Salmo Trutta L.* *Comp. Biochem. Physiology* **88**:751-756.

HerringNetwork. <http://www.clupea.net>. Accessed 2011 April 30.

Inchworm. [http://inchworm.sourceforge.net/inchworm\\_denovo\\_assembly.html](http://inchworm.sourceforge.net/inchworm_denovo_assembly.html).

Johnston, I. A., (2006). Review Environment and plasticity of myogenesis in teleost fish. *J. Exp. Biol.* **209**:2249-2264.

Johnston, I.A., Cole, N.J., Vieira, V.L.A., Davidson I., (1997). Temperature and developmental plasticity of muscle phenotype in herring larvae. *J Exp Biol* **200**:849–868.

Kasahara, M., and Morishita, S., (2006). Large-Scale Genome Sequencing Processing, *Imperial College Press, Japan*.

Kristiansson, E., Asker, N., Förlin, L., and Larsson, DG. JK. (2009). Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. *BMC Genomics* **10**:345.

Lara, A. J., Pérez-Trabado, G., Villalobos, D. P., Díaz-Moreno, S., Cantón, F., and Gonzalo, Claros, M. A Web Tool to Discover Full-Length Sequences: *Full-Lengther*.

Li, R., Hongmei, Jue, Z., Wubin, R., et al., (2010). De novo assembly of human genomes with massively parallel short reads sequencing. *Genome Res.* **20**:265-272

Lin, W., Chen, L., Chen, M., Kao, H., (2009). Differential expression of zebrafish *gpiA* and *gpiB* during development. *Gene Expr. Pattern* **9**:238-245.

Lister, R., B. D., Gregory, and J. R., Ecker, (2009). Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol* **12**:107-118.

Marguerat, S., and J., Bähle, (2010) RNA-seq: from technology to biology. *Cell Mol. Life Sci.* **67(4)**: 569–579.

Marguerat, S., B. T., Wilhelm, and J., Bähler, (2008). Next-generation sequencing: applications beyond genomes. *Biochem. Soc. Trans.* **36(Pt 5)**: 1091–1096.

Melodelima, C., and Gautier, C., (2008). The GC-heterogeneity of teleost fish. *BMC genomics* **9**:632-642.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B., (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621 – 628.

Moyle, P.B., and Cech, J.J., (1992) *Fish, An introduction to Ichthyology, 2nd Edition. Prentice Hall, New Jersey*.

Murray, R. K., Granner, D. K., Mayes, P. A., Rodwell, V. W., (2003). *Harper's Illustrated Biochemistry (26th Edition). McGraw-Hill.* ([http://www.knovel.com/web/portal/browse/display?EXT\\_KNOVEL\\_DISPLAY\\_bookid=955&VerticalID=0](http://www.knovel.com/web/portal/browse/display?EXT_KNOVEL_DISPLAY_bookid=955&VerticalID=0))

Muto, A, Ushida, C., Himeno, H., (1998). A bacterial RNA that functions as both a tRNA and an mRNA. *Trends Biochem. Sci.*, **23**:25–29.

NCBI. <http://www.ncbi.nlm.nih.gov/>

Peng, Z., Diogo, R., and He S. (2009) Teleost fish (Teleostei) *The Timetree of Life*, Oxford University Press: 335-338.

Pop, M., (2004). Shotgun Sequence Assembly. *ADVANCES IN COMPUTERS, VOL. 60* (ISSN: 0065-2458/DOI 10.1016/S0065-2458(03)60006-9).



- Pop, M., (2009). Genome assembly reborn: recent computational challenges. *Brief Bioinfo.* **10(4)**: 354 -366.
- Ryman, N., Lagercrantz, U., Andersson, L., Chakraborty, R., and Rosenberge, R., (1984). Lack of correspondence between genetic and morphologic variability patterns in atlantic herring (*Clupea harengus*). *Heredity* **53(3)**: 687-704.
- Salem, M., Rexroad, C., Wang, J., Thorgaard, G., Yao, J., (2010). Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches. *BMC Genomics* **11(1)**:564.
- Scheibye-Alsinga, K., Hoffmannb, K., Frankel, A., Jensena, P., Stadlerb, P.F., Mangg, P.F., Tommerupg, N., Gilchrist, M.J., Nygård, A.-B., Cireraa, S., Jørgensena, C.B., Fredholma, M., and Gorodkina, J., (2009). Review Sequence assembly. *Computational bio. and chem.* **33**:121-136.
- Shi, C., Yang H., Chao-Ling, W., Yu, O., Zhang, Z., 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.
- SMART. [http://smart.embl-heidelberg.de/smart/do\\_annotation.pl?BLAST=DUMMY&DOMAIN=ANK](http://smart.embl-heidelberg.de/smart/do_annotation.pl?BLAST=DUMMY&DOMAIN=ANK).
- SOAP. <http://soap.genomics.org.cn/soapdenovo.html>
- St-Jean, M., Izard, T., and Sygusch, J., (2007). A Hydrophobic Pocket in the Ac. *The J. Biol. Chem.* **282(19)**: 14309-14315.
- Steffen, L. S., Guyon, J. R., Vogel, D., Howell, M. H., Zhou, Y., Weber, G. J., Zon, L. I., and Kunkel, L. M.,( 2007). The Zebrafish runzel Muscular Dystrophy is Linked to the Titin Gene. *Dev. Biol.* **309**:180-192.
- Sundström, G., (2010). Ph. D. thesis: Evolution of the neuropeptide Y and opioid systems and their genomic regions. Uppsala University, Disciplinary Domain of Medicine and Pharmacy, Faculty of Medicine, Department of Neuroscience, Pharmacology.
- Swedish fish database. <http://fiskbasen.se/laxtobis.html>. Accessed 2011 May 2.
- Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biol.* **28(5)**: 511-15.
- UCSC. <http://genome.ucsc.edu/>.
- Vandekerckhove, J, Weber, K, (1981). Actin Typing on Total Cellular Extracts. A Highly Sensitive Protein-Chemical Procedure Able to Distinguish Different Actins. *Eur. J. Bio-chem.* **113(3)**:595-603.
- Volff, J. N., (2002). Sex determination in fish. *Genome Biology* **3**:reports0052 doi:10.1186/gb-2002-3-9-reports0052.
- Wallimann, T., Wyss, M., Brdiczka, D., Nicolay, K., Eppenberger, H.M., (1992). Intracellular compartmentation, structure and function of creatine kinase isoenzymes in tissues with high and fluctuating energy demands: the 'phosphocreatine circuit' for cellular energy homeostasis. *The Biochem J.* **281 (1)**: 21–40.
- Wang X., Luan, J., Li, J., Bao, Y., Zhang, C., and Liu, S., (2010). De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics* **11**:400.
- Weiss, A., McDonough, D., Wertman, B., Acakpo-Satchivi, L., Montgomery, K., Kucherlapati, R., Leinwand, L., and Krauter, K., (1999). Organization of human and mouse skeletal myosin heavy chain gene clusters is highly conserved. *Proc. Natl. Acad. Sci. USA* **96**:2958–2963.
- Wilhelm, B.T., Landry, J.R., (2009). RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48(3)**:249-57.
- Xu, Y., He, J., Wang, X., Lim, T. M. and Gong, Z., (2000). Asynchronous activation of 10 muscle-specific protein (MSP) genes during zebrafish somitogenesis. *Dev. Dyn.* **219**: 201–215.

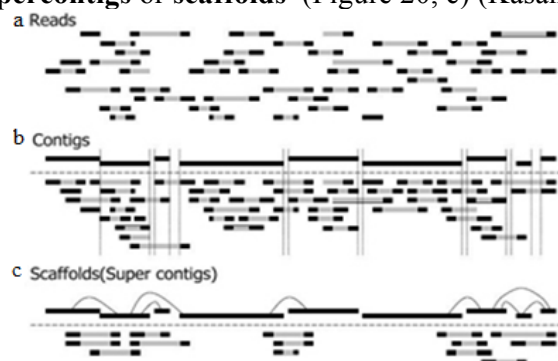
Zhang, Z., Wang, Y., Wang, S., Liu, J., Warren, W., et al. (2011) Transcriptome Analysis of Female and Male *Xiphophorus maculatus* Jp 163 A. *PLoS ONE* **6(4)**: e18379. doi:10.1371/journal.pone.0018379.



## 8 APPENDIX

### 8.1 Genome/Transcriptome assembly

Assembly is the process in which the original genome/transcriptome is reconstructed by merging reads that overlap. Overlapping reads are merged together to form several islands, or contiguous sequences termed **contigs** (Figure 20, b). Then paired end information is used to link contigs together and generate **supercontigs** or **scaffolds** (Figure 20, c) (Kasahara and Morishita, 2006).



**Figure 20- Genome/Transcriptome assembly process. a) Paired-end reads. b) Merged reads that made islands known as contigs. c) Scaffolding by paired-end information.**

In assembling, challenges are sequencing errors and other errors and biases introduced by cloning technologies. Before a genome assembly is conducted it is necessary to trim low-quality reads to avoid introducing possible errors in the process.

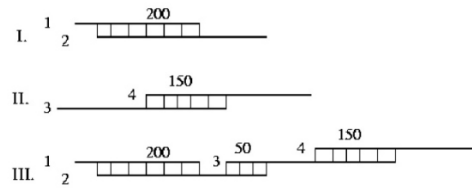
There are different assembly algorithms: greedy, overlap-layout-consensus (OLC), and *de Bruijn* graph (Pop 2004 and Pop 2009). These approaches are applied according to sequencing technologies (Table 12). The first two algorithms are based on reads overlap.

**Table 12- Genome assembly software, their algorithm and the derived data from different sequencing technologies**

Software	Algorithm	Sequencing Technology
Arachne	OLC	Sanger
Celera Assembler	OLC	Illumina, 454, SOLiD
Newbler assembler	OLC	Reads with average length of 300nts
SHARCGSs	Modified-Greedy	Sanger, SOLiD, Illumina, 454
SSAKE	Modified-Greedy	Illumina
VCAKE	Modified-Greedy	Illumina
CAP3	Greedy	Sanger
Edena	Greedy	Illumina
Phrap	Greedy	Sanger, 454
TIGR	Greedy	Sanger
ABYSS	<i>de Bruijn</i> graph	Illumina, SOLiD
SOAPdenovo	<i>de Bruijn</i> graph	Illumina
Velvet	<i>de Bruijn</i> graph	Sanger, SOLiD, Illumina, 454
ALLPATHS	<i>de Bruijn</i> graph	Illumina, SOLiD
EULER-SR	<i>de Bruijn</i> graph	454, Illumina
Trinity (Inchworm)	<i>de Bruijn</i> graph	Illumina GA

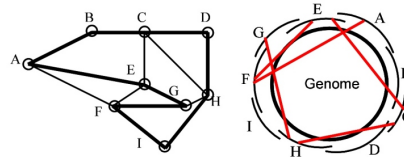
Greedy algorithms typically perform a pairwise alignment between overlapping reads. Multiple alignments are built up iteratively by adding new reads at a time (Figure 21). After the multiple alignment step, a consensus sequence is generated (Pop, 2009 and Scheibye-Alsinga et al., 2009).

The quality of overlap depends on the length of the overlap and the level of identity in the overlapping region (Pop, 2009).



**Figure 21- Greedy assembly for four reads. Numbers show the length of alignment (Pop, 2009).**

OLC is the most widely used approach, pioneered by Pletola et al. in 1984 (Pop, 2009). This approach breaks the assembly into three distinct steps in order to enable a global analysis for all relationships between reads rather than localised approach in greedy algorithm. The first step (**overlap**) is the same as for greedy approach. This information is then used to construct an overlap graph. This graph contains nodes that represent reads and an edge that connects two reads (nodes) with an overlap (Pop, 2009). In the **layout** stage, the overlap graph is analyzed in order to find the paths through the graph that correspond to segments of the genome. The final step is to find a unique path that covers each node in the overlap graph once, this path is known as the **Hamiltonian path** (Pop, 2009 and Scheibye-Alsing et al, 2009). Most *de novo* assemblers for Sanger reads follow the OLC algorithm (Chaisson et al., 2009). However, OLC is an unsuitable approach for very short reads, because it is hard to distinguish the correct assembly due to repetitive sequences and the very short overlaps between these short reads (Li et al., 2010b).



**Figure 22- Overlap graph for a bacterial genome. The right figure shows the correct layout of the reads. The remaining edges represent false overlaps induced by repeats (exemplified by the red lines in the figure on the right) (CBCB).**

The *de Bruijn* graph is a solution to overcome the difficulties of short reads assembly. This method generates a hash table of k-mers (equal short fragments of a read) in reads and produces a graph with the following characteristics. The graph contains nodes that represent k-mers in the shotgun reads, where unit of k varies according to the software limitation. For instance, k-mer limitation in SOAP<sup>3</sup> is 13 to 31 and in Inchworm is 19 to 29. For every k-mer an edge from its (k-1)-mer prefix to its (k-1)-mer suffix is added to a node (Figure 23, A) (Kasahara and Morishita, 2006). The condensed *de Bruijn* graph replaces all paths containing non-branching edges by a single edge labelled by the sequence that generated the path (Figure 23, B).

<sup>3</sup> Short Oligo-nucleotide Analysis Package. It is noteworthy that the new version of this software only accepts K-mers between 13-61 (odds number).

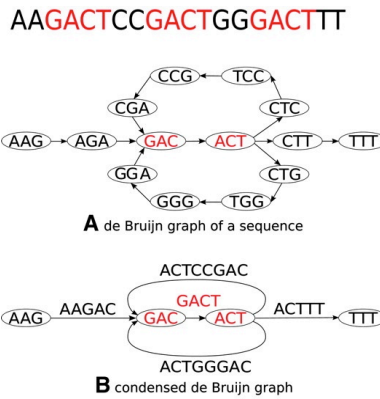


Figure 23- *de Bruijn* graph structure with nodes and edges that have been produced from a short read with the length of 20 nts. A) The *de Bruijn* graph that contain nodes and edges for every K-mer. B) Condensed *de Bruijn* graph that contain a labelled sequence after finding the path (Chaisson et al., 2009).

## 8.2 Quality of sequenced reads

DNA sequencers determine bases according to either waves on electropherogram (Sanger method) or images (Next-generation sequencing technologies) and convert them to one of the nucleotides, i.e., A, C, G, and T. This conversion is called **base calling**, which is often accomplished by computer programs (PHRED for Sanger, Bustard for Illumina, and etc.) (Scheibye-Alsing et al, 2009).

The outputs of the software are nucleotide sequences and quality values (QV). A logarithmic scale is used to calculate QVs with the following formula where P is the error probability per base:

$$QV = -10 \log_{10} P$$

QV is assigned to each base, indicating the probability of an error at that base. In table 13 the accuracy of base calling and corresponding QV is shown.

Table 13- The error probability of base calling and the QV of Illumina and Phred.

Phred Quality Score	Illumina Quality Score <sup>§</sup>	Probability of incorrect base call	Base call accuracy
10	10.41	1 in 10	90 %
20	20.04	1 in 100	99 %
30	30.00	1 in 1000	99.9 %
40	40.00	1 in 10000	99.99 %
50	50.00	1 in 100000	99.999 %

$$\text{§Q}_{\text{Solexa / Illumina}} = 10 \times \log_{10} \left( 10^{\frac{Q_{\text{PHRED}}}{10}} + 1 \right)$$

As final output, sequencers create a file with the nucleotide sequence and the ASCII encoded quality values in a format called fastq (Figure 24).

```
@HWUSI-EAS526_0001:8:1:8:1893#0/1
AGGAACNTGTGAGTGAGGTACAAAAAGAGACAACCTTTTTTCAGTAACGAGTACGATGCCAAAACATTATCACT
+HWUSI-EAS526_0001:8:1:8:1893#0/1
`_UZaaB`OWZSWY``]Za_`aaa`^U^X^]`aTQY_XBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

Figure 24- fastq format.

The QVs differ depending on sequencing technology; in table 14 you can see the difference in QVs within different sequencing technologies.

**Table 14- ASCII characters and the range of quality values in different sequencing technologies. (Cock et al., 2009)**

Description, OBF name	ASCII characters	Quality score	
	Range	Type	Range
Sanger standard <i>fastq-sanger</i>	33–126	PHRED	0 to 93
Solexa/early Illumina <i>fastq-solexa</i>	59–126	Solexa	–5 to 62
Illumina 1.3+ <i>fastq-illumina</i>	64–126	PHRED	0 to 62

### 8.3 Determining optimal k-mer

In order to find optimal k-mer for transcriptome assembly, we used jellyfish software that counts the number of unique k-mers in transcriptome/genome assemblies. We performed this analysis on three fish species: medaka, zebrafish, and fugu. Figure 25 shows the trend of unique k-mer frequencies by increasing the value but it comes to a plateau on value of 20/21 in zebrafish and medaka while the plateau appeared in k-mer size of 18 for fugu. As a result based on this analysis and individual assemblies that we did by Inchworm and different k-mer size we chose k-mer 21 as optimal value.

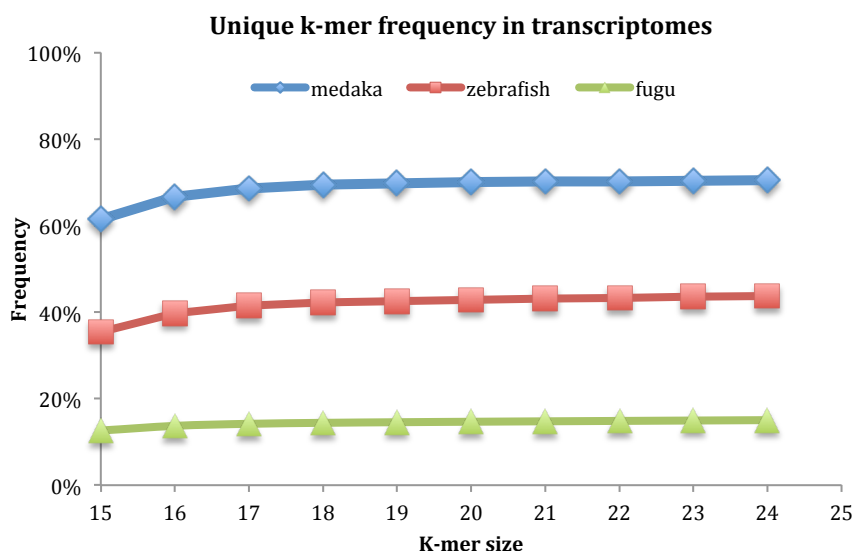


Figure 25- Unique k-mer frequency in medaka (blue), zebrafish (red), and fugu (green) transcriptome.

### 8.4 N50 and N90 calculation

N50 and N90 are most common statistics for comparison of assembled genomes/transcriptome. N50 is the shortest sequence in a set of sequences that together cover 50% of the total length of all assembled sequences. The calculation starts by ranking sequences according to their length in descendant order. If the cumulative length of sequences covers 50% of the total length of the transcriptome, then the order and the length of the last contig sequence in that group shows the result of the N50 statistic. The N90 calculation is the same but for contigs that cover 90% of total length of all sequences. Table 15 shows an example of N50 and N90 calculation. I generated a script in the programming language Perl for calculation of N50/N90 (Appendix 5.5.1).

**Table 15- N50 and N90 calculation.**

Length	Sorted	Cumulative length
4	9	9

8	8	17
6	8	25
5	7	32
5	6	38
1	5	43
1	5	48
3	5	53
4	5	58
3	4	62
3	4	66
5	4	70
9	4	74
5	3	77
4	3	80
7	3	83
4	2	85
1	1	86
8	1	87
2	1	88
<b>Total</b>	<b>88</b>	
50%	<b>44</b>	
90%	<b>79</b>	
N50	<b>5</b>	
N90	<b>3</b>	

## 8.5 Scripts

### 8.5.1 N50 and N90

```
#!/usr/bin/perl
use strict;
#this module calculates the N50 and N90 of a set of contigs.
# ./contig-summary FILENAME (in fasta format)
#Nima Rafati

my $fileName="";
my $input="";
my $contigLn=0;
my $cntr=0;
my $cntrTotal=0;
my $contigSum=0;
my $min=0;
my $max=0;
my $mean=0;
my $nN50;
my @N50;
my @N90;
my $nN90;
my @contigArr;
my $contig=0;
my $matchS;
my $N50Sum;
my $N90Sum;
my $median;
my @scafflength;
my $seq;
```

```

my $seqLength;
$fileName= $ARGV[0];
if ($fileName eq "")
{
    print " please insert the name of the file after the script\n";
    exit;
}
open ("infile", $fileName) || die " I couldn't find the $fileName \n";
while (<infile>)
{
    $input=$_;
    if (/^\s>/)
    {
        chomp($seq);
        $seqLength=length($seq);
        $seq="";
        if ( $seqLength != 0)
        {
            push @contigArr, $seqLength;
            $contigSum+=$seqLength;
            $seq="";
            $seqLength="";
            $cntrTotal++;
        }
    }
    else
    {
        $input=~ s/\n//g;
        $seq.=$input;
    }
}
$seqLength=length($seq);
push @contigArr, $seqLength;
$contigSum+=$seqLength;
$seqLength=length($seq);

close ("infile");
@contigArr= sort {$a <=> $b} @contigArr;
@N50=@contigArr;
@N90=@contigArr;
$contig=@contigArr;
$min=$contigArr[0];
if ($contig % 2)
{
    $median= $contigArr[$contig/2];
}
else
{
    $median=( $contigArr[$contig/2]+$contigArr [ $contig/2-1])/2;
}
$mean=int($contigSum/$contig);
$max=$contigArr[$contig-1];
print "min: $min\t max: $max\t median: $median\t Number: $contig\t sum: $contigSum\t Average:
$mean \n";
while ( $N50Sum < $contigSum * 0.5 )
{
    $contigLn = pop @contigArr;
    $nN50++;
    $N50Sum+= $contigLn;
}
print "N50: $contigLn\t nN50: $nN50\t Sum of N50: $N50Sum\n";
$contigLn=0;
while ( $N90Sum < $contigSum * 0.9 )
{
    $contigLn = pop @N90;
}

```

```

    $nN90++;
    $N90Sum += $contigLn;
}
print "N90: $contigLn\tN90: $nN90\tSum of N90: $N90Sum\n";
exit 0;

```

## 8.6 Software commands

### 8.6.1 SeqClean

The seqclean program trims reads/contigs by a database like UniVec:

```

>seqclean /bubo/proj/b2010051/private/INBOX/20110126A_hiseq2000/end-
trimming/End_2_110126_A816HRABXX_1.fasta -c 8 -o cleaned-End-1.fasta -l 76 -s
UniVec.fasta

```

### 8.6.2 Inchworm

Before running Inchworm it is necessary to modify reads. First, reads should be in fasta format. Second, since our data was not strand-specific we had to convert sequences in one of the files into reverse complementary format. Third, all of the files should be merged in one unique file (Inchworm). There are some handy scripts in Inchworm package that do these modifications:

```

>util/fastQ_to_fastA.pl -I left.fq -a 1 --rev > left.fq.fa
>util/fastQ_to_fastA.pl -I right.fq -a 2 > right.fq.fa
>cat left.fq.fa right.fq.fa > both.senseOriented.fa

```

After input file preparation we used following command to assemble our data with a k-mer size of 21 and minimum length of 100 nts:

```

>inchworm --reads /bubo/proj/b2010051/private/INBOX/20110126A_hiseq2000/end-
trimming/End_1_and_Rev.fasta --run_inchworm --DS monitor 2 --coverage_outfile coverage.txt -
-K 21 -L 100

```

### 8.6.3 SSPACE

SSPACE needs a library file describing the input data.

**Library file example:**

```

lib1
/proj/end-trimming/End_2_110126_A816HRABXX_1.fastq
/proj/end-trimming/End_2_110126_A816HRABXX_2.fastq 240 0,5 0
240: insert size.
0,5: deviation of insert size.

```

For extension of our contigs we used this command:

```

>perl (path_to_SSPACE)/SSPACE_v1-1.pl -l libraries.txt -s
/bubo/proj/b2010051/private/UserDirectories/nima/Herring-Transcriptome/Results/inchworm-K-
21/inchworm-k-21.fasta -x 1 -b SSPACE-inchworm-k-21

```

### 8.6.4 CAP3

We used default values of CAP3 with following command:

```

>Cap3 /bubo/proj/b2010051/private/UserDirectories/nima/Herring-
Transcriptome/Results/inchworm-K-21/inchworm-k-21.fasta

```

### 8.6.5 Trinity

Despite of Inchworm, Trinity does not need any modification of files before running the program.

```
>../Trinity.pl --seqType fa --left
/bubo/proj/b2010051/private/INBOX/20110126A_hiseq2000/end-trimming/cleaned-
End_2_110126_A816HRABXX_1.fasta --right
/bubo/proj/b2010051/private/INBOX/20110126A_hiseq2000/end-trimming/cleaned-
End_2_110126_A816HRABXX_2.fasta --min_contig_length 100 --paired_fragment_length 240 -
-run_butterfly --num_butterfly_CPU 2 --output SeqClean-21
```

Final assembly paths:

1- Trinity-SeqClean-21-All:

/bubo/proj/b2010051/private/UserDirectories/nima/Herring-Transcriptome/Results/Trinity-Seq-Clean-21/FASTA/Trinity.fasta

2- Known subset:

/bubo/proj/b2010051/private/UserDirectories/nima/Herring-Transcriptome/Results/Trinity-Seq-Clean-21/FASTA/Trinity-ORF-Putative-medaka-muscle-zebrafish-nr/Trinity-ORF-Putative-medaka-muscle-zebrafish-nr.fasta

3- Unknown subset:

/bubo/proj/b2010051/private/UserDirectories/nima/Herring-Transcriptome/Results/Trinity-Seq-Clean-21/FASTA/Trinity-NON/Trinity-NON-ORF-Putative-medaka-muscle-zebrafish-nr.fasta

### 8.6.6 SOAPdenovo and GapCloser

#### SOAPdenovo

Before running the software we have to prepare a configuration file. In this file you can set the settings for (SOAP):

- 1- Generating contigs.
- 2- Generating scaffolds.
- 3- Generating contigs and scaffold together.
- 4- Intra-scaffolding.

Below you can find a sample of configure file

**#maximum read length.**

max\_rd\_len=76

**#First library**

[LIB]

**#average insert size.**

avg\_ins=240

**#For paired end reverse\_seq=0 and for mate-pair this value is equal to 1.**

reverse\_seq=0

**# for contig generation (1), scaffold (2), contig and scaffold (3), and intra-scaffolding (4).**

asm\_flags=3

**#rank of the library.**

rank=1

**#input file path: for FASTA format (f1/2) and for FASTQ format (q1/2).**

f1=/bubo/proj/b2010051/private/INBOX/20110126A\_hiseq2000/end-trimming/End\_2\_110126\_A816HRABXX\_1.fasta

f2=/bubo/proj/b2010051/private/INBOX/20110126A\_hiseq2000/end-trimming/End\_2\_110126\_A816HRABXX\_1.fasta



.....( and for other libraries).

- ✓ **Please note that you have to make the configure file according to your assembly strategy.**

SOAPdenovo can be performed either step by step or all steps together:

For making contigs and scaffolds at one run we can use:

```
>SOAPdenovo all -s Config.txt -K 314 -o graph_prefix -R yes -p 15
```

For step by step we have to follow commands in below:

```
>soapdenovo pregraph -s config_file -K 31 -p 15 -o graph_prefix
```

```
>soapdenovo contig -g graph_prefix
```

```
>soapdenovo map -s config_file -g graph_prefix -p 15
```

```
>soapdenovo scaff -g graph_prefix -F yes
```

### GapCloser (Intra-Scaffolding)

```
>GapCloser -b config_file -a graph_prefix -o output_file -p 31 -t 15
```

Input files are:

Generated scaffold file with the extension of scafq by SOAPdenovo.

Config.txt file that was used as an input for SOAPdenovo

Output files:

There are two files: Scaffolds and singletons' sequences after filling the sequence gaps<sup>5</sup> with the extension of Gap.

The descriptions of filled gaps are saved in a file with the extension of fill.

### 8.6.7 Jellyfish

Jellyfish software count the number of k-mers by this command:

```
>jellyfish count -m $j -o $dirName -s 100000000 $genomeSeqfile
```

we used different values for k-mer (\$j=15 to 25).

### 8.6.8 SOAPaligner

By this program we can align either one read library or paired-end libraries on contigs/scaffolds. In first case there is only one output while in second case there are two outputs including paired-end aligned reads and unpaired aligned reads. The input read files can be in fasta or fastq format (SOAP).

```
>soap -a /bubo/proj/b2010051/private/INBOX/20110126A_hiseq2000/end-trimming/End_2_110126_A816HRABXX_1.fasta -b /bubo/proj/b2010051/private/INBOX/20110126A_hiseq2000/end-trimming/End_2_110126_A816HRABXX_2.fasta -o output.soap -2 Unpaired.soap -u Unmapped.soap -m 180 -x 400 -p 8
```

### 8.6.9 CD-HIT

```
>cd-hit -i ../FASTA/Trinity.fasta -o Trinity100.fasta.node -c 0.9 -n 5 -T 8
```

### 8.6.10 UCLUSTAL

UCLUSTAL remove artefacts in two steps:

---

<sup>4</sup> This number must be odd number. New version of this software accepts K-mer to 61.

<sup>5</sup> These gaps are between contigs in one scaffold. These gaps are different from physical gaps that exist between scaffolds and the length of them is unknown.

- 1- Sorting contigs by their length:  
`>usearch --sort /bubo/proj/b2010051/private/UserDirectories/nima/Herring-Transcriptome/Results/Trinity-Seq-Clean-21/FASTA/Trinity.fasta --out Sorted-Trinity.fasta --maxlen 12000 --log sort.log`
- 2- Removing artefacts:  
`>usearch --cluster Sorted-Trinity.fasta --id 0.97 --seedsout clustered.fasta --uc results.uc --log cluster.log --maxlen 12000`

### 8.6.11 BLAST

We used BLAST to align our transcriptome contigs against available databases. For proteins we used BLASTX and for nucleotides we used BLASTN. BLAST uses specific format of database that is prepared by **formatdb**.

For protein:

```
>formatdb -p T -n nr-formatdb -i nr.fasta
```

For nucleotide:

```
>formatdb -p F -n nt-formatdb -I nt.fasta
```

Transcriptome alignment on protein database:

```
>blastall -p blastx -d nr-formatdb -e 0.00001 -m 8 -a 8 -W 3 -i Trinity.fasta -out Trinity-nr.blast
```

Transcriptome alignment on nucleotide database:

```
>blastall -p blast -d nt-formatdb -e 0.00001 -m 8 -a 8 -W 11 -i Trinity.fasta -out Trinity-nt.blast
```

### 8.6.12 Mosaik

In first step, reads and reference sequences are converted to binary files:

```
>MosaikBuild -fr Trinity-SeqClean-21.fasta -oa Trinity-SeqClean-21.dat
```

```
>MosaikBuild -f /bubo/proj/b2010051/private/INBOX/20110126A_hiseq2000/end-trimming/cleaned-End_2_110126_A816HRABXX_1.fasta -f2 /bubo/proj/b2010051/private/INBOX/20110126A_hiseq2000/end-trimming/cleaned-End_2_110126_A816HRABXX_2.fasta -out reads.dat -st illumina -p 8
```

After making binary files we performed the alignment by this command:

```
>MosaikAligner -in reads.dat -ia Trinity-SeqClean-21.dat -out reads-on-Trinity-SeqClean-21.dat -p 8
```

Then we sorted the output of MosaikAligner by:

```
>MosaikSort -in reads-on-Trinity-SeqClean-21.dat -out reads-on-Trinity-SeqClean-21-sorted.dat
```

By MosaikSort we converted our outputs from MosaikSort to SAM/BAM file by MosaikText:

```
>MosaikText -in reads-on-Trinity-SeqClean-21-sorted.dat [-bam reads-on-Trinity-SeqClean-21.bam] -sam reads-on-Trinity-SeqClean-21-sorted.sam]
```

### 8.6.13 Samtools

In order to find SNPs and indels we prepared pileup file by using samtools with following commands:

```
>samtools view -bT Trinity.fasta Trinity-SeqClean-21.sam >Trinity-SeqClean-21.bam
```

```
>samtools sort Trinity-SeqClean-21.bam Trinity-SeqClean-21.bam.sort
```

```
>samtools index Trinity-SeqClean-21.bam.sort.bam
```

```
>samtools pileup -f Trinity.fasta Trinity-SeqClean-21.bam.sort.bam
```

### 8.6.14 VarScan

VarScan is a program that calls SNPs and indels by using pileup files:

```
>java -jar VarScan.v2.2.3.jar pileup2cns  
/bubo/proj/b2010051/private/UserDirectories/nima/Herring-Transcriptome/Results/Trinity-Seq-  
Clean-21/FASTA/Trinity-ORF-Putative-medaka-muscle-zebrafish-nr/PE-SE-ORF-Putative-  
medaka-muscle-zebrafish-nr.bam.sort.bam.f --min-coverage 8 --min-avg-qual 20 --minvar-freq  
0.2 --p-value 0.05
```