# Bioinformatics Tools and Applications for Rainbow Trout

By

Rafet Al-Tobasei

A Dissertation Submitted in Partial Fulfillment

Of the Requirements for Degree of

Doctor of Philosophy in Computational Science

Middle Tennessee State University

May 2017

Dissertation Committee:

Dr. Mohamed Salem (Chair)

Dr. Hyrum D. Carroll

Dr. Cen Li

Dr. Rebecca Seipelt -Thiemann

Dr. Joshua L. Phillips

*Dedicated to*

*My children (Ahmad, Zayna, Ruba, Leena), my lovely wife (Nour Kattih), and my Mother (Sadeqah Alkatib)*

# ACKNOWLEDGMENTS

During my study, I was blessed with many people that supported me and encouraged me to make this PhD a reality.

My Advisor, Dr. Mohamed Salem, I do not know how to thank him enough for his support and guidance through this journey. Dr. Salem went far and beyond what a dissertation chair does. He taught me about a subject that was completely out of my radar. He helped me drafting my research with patience and excellent enlightenment. Working with Dr. Salem has been a pleasure that will influence me forever. Computational Science Director, Dr. John Wallin, thanks for all the supports and funds you provided me. Your support and encouragements will always be appreciated. I would like to thank Dr. Chrisila Pettey, Computer Science department chair, for her support. I always enjoy working with Dr. Hyrum Carroll sharing ideas and thoughts. Dr. Cen Li; I always enjoyed your classes as a student and as a lab assistant. I learned from you how to be a better instructor. I would like to thank my doctoral committee members: Dr. Rebecca Seipelt - Thiemann and Dr. Joshua Philips for their support, time, feedback, and assistance reviewing this dissertation. Thank you all for your support and guidance.

I would like to thank my colleagues and collaborator, from MOBI program in Middle Tennessee State University Bam Paneru and Ali Ali, from USDA Dr. Yniv Palti and Dr. Timothy Leeds, and from West Virginia University Dr. Brett Kenney.

I would like to thank my family for their support and encouragement during this journey. My wife, Nour Kattih, my children, Ahmad, Zayna, Ruba, and Leena who stood by my

side all the times especially the days when I had to work late. I would like to thank my Mom for her prayer for me.

Thank you Middle Tennessee State University for providing me the education and knowledge that helped me to achieve my PhD.

# ABSTRACT

Rainbow trout is one of the widely used aquaculture species for food worldwide. Due to its commercial importance, various genomic resources are available for the trout including a draft reference genome, microRNA repertoire, quantitative trait loci and single nucleotide polymorphisms (SNPs) associated with different production traits. However, many of these genomic resources still need improvement in terms of quality and quantity. The only available genome draft is not completely annotated, and lacks non-coding RNA and some protein coding genes. Similarly, majority of the previous work aimed at identification of trait-associated genetic markers were not robust due to limitation of genomic resources that were previously available.

In this study, we used genomics and transcriptomic approaches to identify missing genetic elements including long non-coding RNA and protein coding genes in the reference genome. In addition, we utilized these genomics resources to identify genes and genetic variations, especially SNPs, associated with growth and muscle quality traits in rainbow trout. In order to facilitate gene discovery and to improve the draft genome reference, we used deep transcriptome sequencing from 13 vital tissues. De novo assembly of ~1.167 billion paired-end reads from those 13 tissues identified a total of 474,524 protein coding transcripts, of them 11,843 transcripts were not previously reported in the genome reference. In order to discover long non-coding RNA repertoire, we used the same ~1.167 billion RNA sequencing reads in addition to RNA sequence data from 3 other published sources. Transcriptome assembly followed by various filtration steps identified 54,503 long non-coding RNA transcripts, which provided the first long non-coding RNA draft reference in rainbow trout. These long non-coding RNAs exhibited less sequence conservation, one exon biased structure and overall lower expression level compared to protein coding genes. The newly identified long non-coding RNAs showed differential expression in response to *Flavobacterium psychrophilum* infection, and their expression level strongly correlated with body bacterial load in selectively bred, resistant-, control-, and susceptible- genetic lines of rainbow trout. These findings suggest that the lncRNAs have importance roles in antibacterial immune response and disease resistance in rainbow trout. In addition, multiple bioinformatics algorithms were tested and successfully utilized to identify SNPs in protein coding genes and long non-coding RNAs that are associated with 5 important production traits: whole body weight (WBW), muscle yield, muscle crude fat content, muscle shear force (tenderness) and fillet whiteness. A total of 7,930 SNPs identified in protein coding genes and non-coding RNAs showed allelic imbalances (>2.0 as an amplification and <0.5 as loss of heterozygosity) between fish families showing contrasting phenotypes for above-mentioned traits suggesting their importance in the phenotypes. Validation of a small subset of the SNPs with allelic imbalances showed ~93% success rate of the pipelines in calling SNPs suggesting reliability of the algorithms.

This study provides new genomic resources to complement the genome annotation and facilitate functional genomics research in addition to genome-wide studies and selection in rainbow trout.

## DECLARATION

I declare that all the contents of this dissertation are my original work and has been done with collaboration with other colleague. Some of this work has been done with students from MOBI program who may use it in their dissertation. My major contributions were on the second and third publications and I significantly contributed to the first and the fourth publications.

Work presented in this dissertation are based on the following works:

Published/Submitted Papers:

1. Salem M, Paneru B, Al-Tobasei R, Abdouni F, Thorgaard GH, et al. (2015) Transcriptome Assembly, Gene Annotation and Tissue Gene Expression Atlas of the Rainbow Trout. PLOS ONE 10(3): e0121778. doi: 10.1371/journal.pone.0121778
2. Al-Tobasei R, Paneru B, Salem M (2016) Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout. PLOS ONE 11(2): e0148940. doi: 10.1371/journal.pone.0148940
3. Al-Tobasei R, Ali Ali, Timothy D. Leeds, Sixin Liu, Yniv Palti, Brett Kenney, and Mohamed Salem (2017) Identification of SNPs Associated with Muscle Yield and Quality Traits Using Allelic-Imbalance Analysis in Pooled RNA-Seq Samples in Rainbow Trout (submitted)
4. Paneru B, Al-Tobasei R, Palti Y, Wiens GD, Salem M (2106) Differential expression of long non-coding RNAs in three genetic lines of rainbow trout in

response to infection with *Flavobacterium psychrophilum*. Scientific Reports 2016, 6:36032.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER I: INTRODUCTION

The 21st century will be the century of biology, where the advancement of computing technology and sequencing technology produce a tremendous amount of data that needs to be analyzed [1]. Being able to collect biological data and process it using supercomputing power can help society find answers to multiple problems, such as finding diagnostics for diseases -for example The Cancer Genome Atlas project- and solving famine problems by generating crops and stocks that can handle diseases, rough weather and harsh environments.

In 1990, one of the biggest adventures in science was started. This was the launching of Human Genome Project, led by the United States, represented by the National Institutes of Health (NIH) and the US Department of Energy (DOE), and other countries that formed the International Human Genome Sequencing Consortium. The project had a budget of 3 billion dollars and a 15 year deadline [2].

Two main principles were adopted by the Human Genome Project. First was to welcome collaborations from any nation, since the human genome is a common heritage of all humans. The second principle was the release of data by making sure all sequence data and results were released to the public within 24 hours [2].

Releasing the data and making it available to other scientists so they can use it for research opens the door for all scientists from all over the world to be part of this ongoing research. One of the main goals of the Human Genome Project was, the construction of genetic and physical maps of the human and mouse genomes [2]. To achieve the goals of the Human

Genome Project, 200 labs in the United States were funded by DOE and NIH; in addition, by the end of the project, 18 different countries were contributing toward this project.

The first draft of the Human Genome was announced from the White House by President Bill Clinton on June 26, 2000. Three years later, in April 2003, a finished version of human genome was announced by the Human Genome Sequencing Consortium. The work did not stop after this announcement. On the contrary, the adventure and the result of this important work had just begun. Today, scientists and researchers use this tool to better understand how a human being is developed. An important aspect is that it allows scientists to learn more about genetic diseases and how to treat them.

It took almost 13 years, in addition to more than 3 billion dollars, to come up with the first draft of the human genome, but with advancements in supercomputing and sequencing technologies, such as the next-generation sequencing technology by different private companies like Illumina, the goals of having the human genome assembly within one week and with a cost of less than a $1000 dollars is now a reality [3]. Future technologies can further bring the cost down of obtaining one’s genome sequence to less than $100, and only require about 24 hours to complete (Figure 1 ). This advancement will open the doors for scientists to be able to sequence multiple genomes for different people and species and do comparisons for better understanding of the genome biology among different taxa.

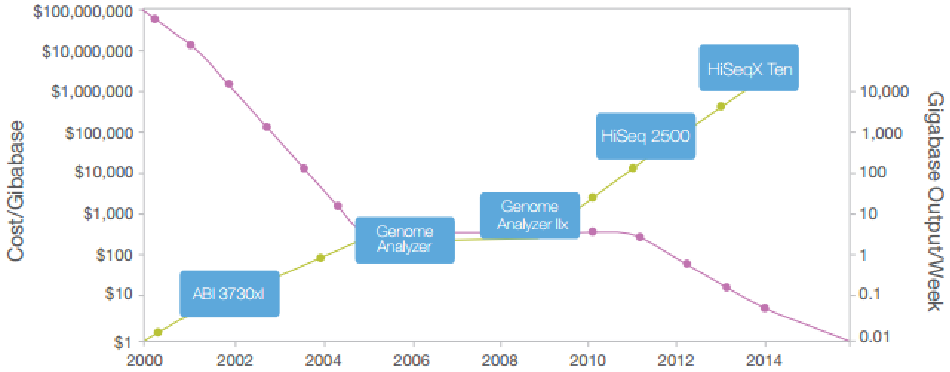


Figure 1: DNA Sequencing Cost and Data Output since 2000 [3].

The Human Genome Sequencing Project led the way for genetic research for other species. Almost all agricultural species, such as cow [4-7], chicken [8-10], pig, wheat [11-13], goat, fish etc. are being subject to this genetic research. The goal of this research is to improve the quality of life by improving yield, disease resistance, and quality of crops [13-16].

## Aquaculture

This dissertation research concentrates on one agricultural species: aquaculture species rainbow trout. Aquaculture is the farming of aquatic species, whether these are used for food, sports, or ornament. Until 1980, most sources of aquatic organisms were seas, oceans, and the rivers. However, due to increased demand and economic reasons, people started raising aquatic organisms in farms. In 1980 almost 97% of the seafood came from natural fishing,  and only 3% came from aquaculture production (Figure 2) [17]. The aquaculture percentage is steadily rising as shown (Figure 2, Table 1); in 2014 more than 44% of seafood came from aquaculture farms.

.

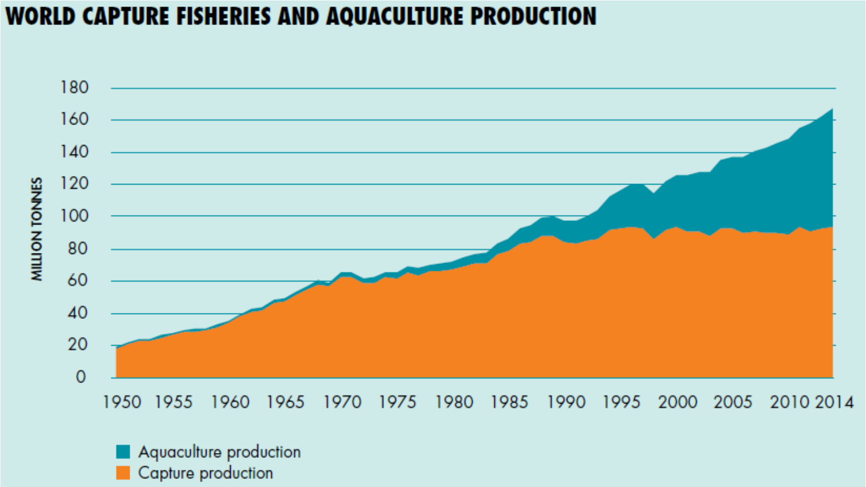


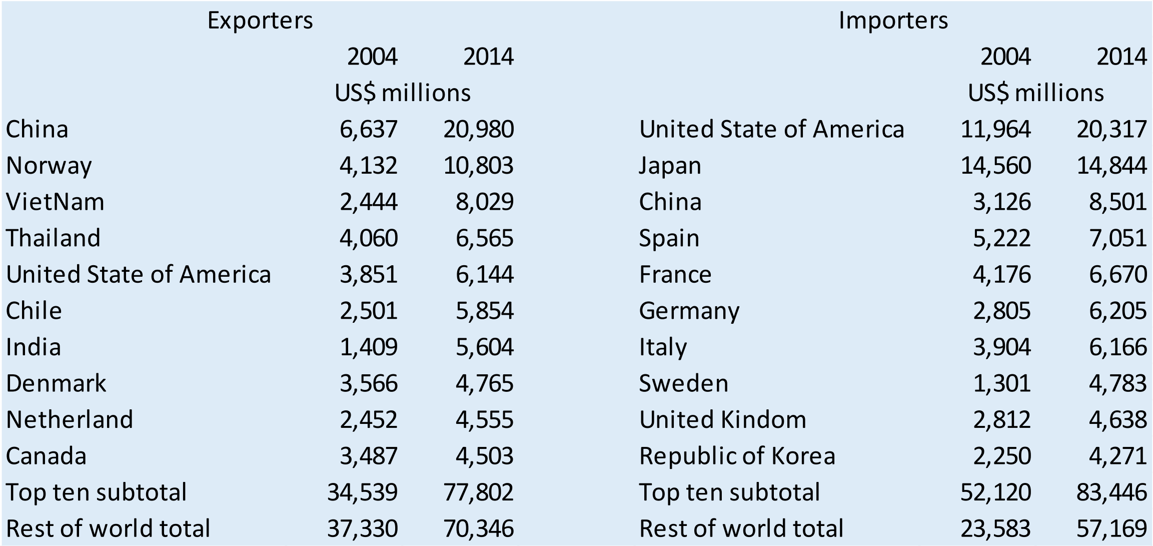
Figure 2: World Capture Fisheries and Aquaculture Production [17].

Table 1: World Fishery and Aquaculture Production data used in this table were adapted from Food and Agriculture Organization of the United Nation FAO [17].

| | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|
| Production | | | | | | |
| Capture | | | | | | |
| Inland | 10.5 | 11.3 | 11.1 | 11.6 | 11.7 | 11.9 |
| Marine | 79.7 | 77.9 | 82.6 | 79.7 | 81 | 81.5 |
| Total Capture | 90.2 | 89.2 | 93.7 | 91.3 | 92.7 | 93.4 |
| Aquaculture | | | | | | |
| Inland | 34.3 | 36.9 | 38.6 | 42 | 44.8 | 47.1 |
| Marine | 21.5 | 22.1 | 23.2 | 24.4 | 21.4 | 26.7 |
| Total aquaculture | 55.8 | 59 | 61.8 | 66.4 | 66.2 | 73.8 |

According to Food and Agriculture Organization of the United Nation (FAO) [17], aquaculture is considered the fastest growing agricultural economic product worldwide at a rate of 6.1%.  According to the United Nations statistics, in 2016 the US considered the largest single importer of fish and fish products, with total imports of $20.13 billion and with total exports of $6.14 billion (Table 2). This big difference between imports and exports produces a deficit of $14 billion, which shows a need for improving aquaculture production in the US.

Table 2: Exporters and Importers of Fish and Fishery Products data used in this table were adapted from FAO [17].



| Exporters | | | Importers | | |
|---|---|---|---|---|---|
| | 2004 | 2014 | | 2004 | 2014 |
| | US$ millions | | | US$ millions | |
| China | 6,637 | 20,980 | United State of America | 11,964 | 20,317 |
| Norway | 4,132 | 10,803 | Japan | 14,560 | 14,844 |
| VietNam | 2,444 | 8,029 | China | 3,126 | 8,501 |
| Thailand | 4,060 | 6,565 | Spain | 5,222 | 7,051 |
| United State of America | 3,851 | 6,144 | France | 4,176 | 6,670 |
| Chile | 2,501 | 5,854 | Germany | 2,805 | 6,205 |
| India | 1,409 | 5,604 | Italy | 3,904 | 6,166 |
| Denmark | 3,566 | 4,765 | Sweden | 1,301 | 4,783 |
| Netherland | 2,452 | 4,555 | United Kindom | 2,812 | 4,638 |
| Canada | 3,487 | 4,503 | Republic of Korea | 2,250 | 4,271 |
| Top ten subtotal | 34,539 | 77,802 | Top ten subtotal | 52,120 | 83,446 |
| Rest of world total | 37,330 | 70,346 | Rest of world total | 23,583 | 57,169 |

Aquaculture product is considered one of the main food sources of protein for human beings; according to FAO, fishery protein accounts for 16.7% of total animal protein consumed by humans [18]. In the United States alone, people consume 5,550,744 tons of seafood a year [18]. One of these major aquaculture organisms is rainbow trout. It is considered one of the main seafood sources in the United State and worldwide. Rainbow trout (*Oncorhynchus mykiss*), a member of *Salmonidae* family, is a native species of the Pacific coast of North America. In addition, rainbow trout is considered a model for other species. Several studies have been done on rainbow trout, including studying genetics, ecology [19], pathology [20], physiology [21], toxicology [22] and carcinogenesis [23]. Having a complete and well-annotated rainbow trout reference genome will provide genomic tools for scientists; these in turn will help in finding markers, single nucleotide polymorphisms (SNP), quantitative trait loci (QTL), and gene annotations and providing basic functional genomics information that will increase opportunities for genetic improvement that could be used to increase fish production efficiency and value-added products and could increase its usefulness as a biomedical research model.

## Genome Annotation

Genomics is the science of studying the structure and content of a genome for each species. With the help of supercomputing and the advancement of sequencing technology, the cost of sequencing is becoming affordable by almost any institute[24]. Data emerging from next generation sequencing offers unprecedented opportunity to study any genome.

In April 2014, Berthelot C, *et al.* publish the first draft reference genome for rainbow trout [25]; however, the reference genome is not complete yet. The estimated length of the rainbow trout genome is 2.4-3.0 GB [26, 27]. But the total length of the assembled genome reference is 2.1 GB, and only 1.023 GB (48%) of the total assembly is anchored to chromosomes. Additionally, there are 30,339,800 ambiguous nucleotides representing unknown gaps [25]. The current version of the rainbow trout genome is not well-annotated, with many genes misassembled, missing or fragmented, and lacks non-coding RNA. Similarly, majority of the previous work aimed at identification of trait-associated genetic markers were not robust due to limitation of genomic resource available at that time. Therefore, we are proposing to improve the genome reference annotation, we will use genomics and transcriptomic approach to identify missing genetic elements including long non-coding RNA and protein coding genes in the draft reference genome. And we will utilize these genomics resources to identify genes and genetic variations, especially SNPs, associated with growth and muscle quality traits in rainbow trout.

### Transcriptome *De novo* Assembly and Gene Annotation

One of the main objectives of any genome study is to identify genes and their characterizations. To understand how genes and gene products interact, we need to determine gene structure and function annotations [24, 28-31]. The gene identification process involves multiple stages [24, 28-31]. In this study, two different methods were used to assemble the rainbow trout transcriptome. The first method is reference-based using Tophat and cufflinks package and the second method uses a *de novo* approach using Trinity [32] software. After assembling short read sequences into contiguous sequences (contigs), protein coding sequences were detected  by searching for homology in a protein sequence database, using BLASTX [33] search against the NCBI non-redundant (nr) protein database which translates transcripts to all possible open reading frames (ORF) that can provide a functional annotation. Further analyses were performed to determine complete ORF. Gene annotation is completed with gene ontology by determining biological process, molecular function, and cellular component of a protein.  Providing basic functional genomics information in addition to classifying and annotating the coding nucleotide sequences will improve opportunities for genetic improvement of rainbow trout.

## Long non Coding RNA (lncRNA)

Different studies showed that almost 70% of a genome is transcribed, where only 2% of the genome transcribed into protein [34-36]. That raises an important question about the function of the remaining transcriptome (68%) of transcribed genome. These genes are transcribed into noncoding protein (ncRNA). There are different categories of ncRNAs, such as transfer RNAs (tRNAs), microRNA (miRNAs), small nuclear RNA (snRNAs), small nucleolar RNA (snoRNAs), small interfering RNA (siRNAs), signal recognition particle (SRP) RNAs, in addition to lncRNAs which constitute the majority of the transcribed RNAs. LncRNAs are greater than 200 nucleotides (nt), are not translated into protein, and most of lncRNAs have an open reading frame (ORF) less than 100 amino acid [37-39]. LncRNAs have fewer exons than coding genes, and on average, they are shorter than coding genes. LncRNAs can be classified into two categories: genic, which either partially or fully overlap with protein coding gene (sense, antisense, intronic, and exonic), and intergenic which exist close by protein coding genes, most of them within 10k nt. LncRNAs have different functions [40], such as lncRNAs can act as scaffolds [40], where lncRNAs link multiple protein factors together to create more complex cellular machines [40, 41]. Other lncRNAs work as decoys, where they interact with the promoters of genes under some signals such as stress or heat and prevent the genes from being transcribed [40, 42]. Some lncRNAs work as molecular guides by localizing particular ribonucleoprotein complexes to specific chromatin targets[40]. Other lncRNAs participate in signaling, where they combines with protein factors to activate pathways. LncRNAs are not generally evolutionarily conserved, which makes it hard to detect them [43]. To improve the genome annotation, one of the major objectives of this study was to identify the transcribed lncRNAs and use this information to annotate the rainbow trout genome reference.

We used these lncRNA data in a functional genomics study "*Differential expression of long non-coding RNAs in three genetic lines of rainbow trout in response to infection with Flavobacterium psychrophilim."* One of the major causes of mortality of salmonids is the Bacterial Cold Water Disease which is caused by *Flavobacterium psychrophilim* (*Fp*). Among the functions of lncRNAs is regulation of transcription and post-transcriptional events of protein-coding genes, which are including in cellular processes, such as disease immunity. In this study, lncRNAs that are associated with genetic resistance against *Fp,* a causative agent of Bacterial Cold Water Disease (BCWD) were identified. An RNA-Seq approach was used to quantify differentially expressed (DE) lncRNAs in response to *Fp* challenge in three genetic lines (Resistance, Susceptible, and Control). Strong expression correlation with their overlapped, neighboring, and distant immune related protein-coding genes involved in immunity was discovered.

## Single Nucleotide Polymorphism (SNP)

To complement the genome annotation, different tools have been used to identify Single Nucleotide Polymorphisms (SNPs) in the transcribed regions of the genome. SNPs related to specific growth and quality traits were targeted. If we take a close look at two genomes of the same species, we can identify three different types of variations. These are copy number variation (CNV), insertion/deletions (indels), and SNPs, which are the sites in the genome where single nucleotides vary from one genome to another. The majority of variation comes from SNPs which constitute about 90% of all variations. There are hundreds of thousands of SNPs located across the genome in different locations. SNP classification depends on where they are located. Some SNPs are located in the non-protein coding region, such as in the introns or at the 5’ and 3’ untranslated (UTR) ends of the gene. Other SNPs are located in the protein coding region, which are called cSNP. The cSNPs can be of two types, synonymous polymorphisms and non-synonymous polymorphisms. Synonymous polymorphisms cause a change in the codon but not in the amino acid. Non-synonymous polymorphisms change both the codon and the amino acid which could lead to change in the function of the protein. Other SNPs are classified as regulatory polymorphisms. Those functional SNPs can affect the transcriptional or translational regulation of a gene.

For the last 15 years, a plethora of studies have been done on SNPs and their role in several aspects in the medical and agricultural fields [44-46]. Different studies have been done for SNP discovery in rainbow trout fish, such as recent studies by Palti et al., [47, 48]. These studies targeted SNPs across the entire genome, coding and non-coding parts, which identified 57K SNPs that are spread across the 2.7 billion nt, or about 27 SNPs for each 1,000,000 nt. Those SNPs have been used to build a relatively low-density SNP chip that cannot capture most of the SNPs in the gene coding regions.  Typically, one SNP exists for every 1,000 nt [49].

In our study, we are extending the literature by generating SNP data for the coding region in fish with phenotypic variations in specific traits. These data will be used to build a 50K SNP chip specifically for the coding (transcribed) part of the genome, which represents 1-2% of the whole genome. In 2012, Salem *et al.* used RNA-Seq to identify SNP markers for growth traits in rainbow trout [50]. We have used the same approach, but on a larger scale, to identify up to 50K cSNPs from a large number of fish and fish families and identify cSNPs that could serve as genetic markers for multiple traits including whole body weight, muscle yield, muscle sheer force (softness), muscle fat content, and muscle whiteness. The 2012 study by Salem *et al*. was done without the use of a reference genome, which was published in 2012 [50]. Having a reference genome should help filtering out the false SNPs. SNPs generated from this study were used to build a cSNP-chip for rainbow trout that could be used by other researchers. The cSNP-chip may be used to develop SNP markers for genetic selection of improved fish production traits in the USDA aquaculture research center at Leetown, WV.

One of the challenging problems in working of non-model eukaryotic species is to determine splice variants without a reference genome. To overcome this challenge, we introduce a new approach to determine splice variants without the need of a reference genome.

## Specific Objectives

This work includes the following chapters:

1. Assembly of a reference transcriptome and protein-coding gene discovery and gene annotations for rainbow trout.
2. Identification and characterization of long non-coding RNAs (lncRNA) in rainbow trout genome to create a global gene expression atlas of lncRNAs in several vital tissues.

3. Identification of cSNPs with variations in production traits such as growth rate and muscle quality traits. This will lead to building a cSNP-chip for rainbow trout that could be used by other researchers.
4. Identification of lncRNAs that are associated with genetic resistance against *Fp* and to identify immune-relevant protein-coding genes that might be regulated by lncRNAs.
5. Introduction of a new method to detect splice variants without the need of a reference genome, using a new approach that depends on *de novo* assembly.

This study provides the following benefits for the science and research community:

a) Improved genome annotation (transcriptome reference) for rainbow trout;

b) A database of lncRNA for rainbow trout;

c) Data to construct a cSNP-chip and genetic markers for rainbow trout.

# CHAPTER II: TRANSCRIPTOME ASSEMBLY, GENE ANNOTATION AND TISSUE GENE EXPRESSION ATLAS OF THE RAINBOW TROUT [51]

## Introduction

Rainbow trout (*Oncorhynchus mykiss*), a member of *Salmonidae* family, is a native species of the Pacific coasts of North America and Russia [52]. They are extensively cultivated worldwide for food, and commercial rainbow trout production significantly contributes to the aquaculture industry in several countries including the USA. In addition, rainbow trout is one of the most extensively studied fish species as it is widely used as a model organism in biomedical research including immunology [53], carcinogenesis [54], physiology [21], nutrition [55], toxicology [22, 56], microbial pathogenesis [20], and ecology [19]. More than 9,686 biomedical articles and abstracts have been published on rainbow trout [57].

Over the past decade, international efforts have been made to increase the genomic data on rainbow trout resulting in a significant amount of information in public databases [25, 58-67]. *De novo* transcriptome sequencing has been successfully used for gene discovery, single nucleotide polymorphism (SNP) identification, molecular marker development, detection of expression quantitative trait loci (eQTL), and differential gene expression profiling [68-70]. The available rainbow trout transcriptomic resources include a transcriptome reference sequence that has been developed in our laboratory using a 19X coverage of Sanger and 454-pyrosequencing data [71]. In addition, another reference transcriptome was sequenced in our laboratory representing responses to several stressors affecting the aquaculture production environments [72]. Further, a transcriptome sequence of the anadromous steelhead (*Oncorhynchus mykiss*) was recently reported [73]. While the first study aimed at assembling a transcriptomic reference for gene discovery, the latter two studies complemented the existing transcriptomic resources and facilitated evaluating gene expression associated with adaptation to ecological and environmental factors in rainbow trout.

Identifying and annotating the coding nucleotide sequences and providing basic functional genomics information will enhance opportunities for genetic improvement of this fish for aquaculture production efficiency and product value and increase its usefulness as a biomedical research model. Recently, unannotated genomic scaffolds and contigs with ~70% coverage of the genome length were assembled from the Swanson River clonal line [74]. More successfully, a draft of the genome sequence has been assembled from a single homozygous doubled haploid YY male from the same clonal line [25]. A gene models approach based on both a genome and transcriptome sequences was used to annotate the genome sequence, predicting 69,676 transcripts. However, the genome sequence still is not complete, with a total length of 2.1 Gb and only 1.023 Gb (48%) of the total assembly anchored to chromosomes [25]. To improve annotation of the under developed trout genome sequence and estimate assembly coverage, a complete and well-annotated transcriptome reference sequence is still needed. Therefore, a *de novo* approach was used in this study to sequence and assemble the rainbow trout transcriptome using in-depth (4,333X) sequence coverage.

Next-generation sequencing is a rapid and cost-effective method for sequencing. However, short sequencing reads generated by most high-throughput sequencing techniques pose difficulties in *de novo* assembly resulting in short/fragmented assemblies of genes [75]. In addition, about 50% of the genes in salmonids are duplicated [76], which makes *de novo* assembly and annotation of the transcriptome difficult and complicates SNP/variant discovery [77-80]. To help overcome these bioinformatics challenges of the trout duplicated genome, we have sequenced the transcriptome of a single doubled haploid fish from a clonal line in an effort to remove sequence variation resulting from polymorphism [25]. This doubled haploid clonal line, which contains two identical copies of each chromosome, was previously established by chromosome set manipulation techniques [81, 82] and has been used in sequencing the rainbow trout genome and transcriptome [25, 71, 83]. Recently, dramatic improvements in genome assembly of *Takifugu rubripes* were achieved by using doubled-haploid individuals compared to the wild type [84].

Housekeeping genes were initially described as genes which are always expressed in the cell [85]. Later, this concept has been refined to refer to genes with constitutive expression that maintain normal cellular functions [86]. In contrast, tissue-specific genes are transcripts whose functions and expressions are favored in specific tissue/cell types [87]. Tissue-specific gene expression is crucial for maintaining specificity and determining complexity of multicellular organisms as they affect the development, function and maintenance of diverse cell types within an organism. Studying the ubiquitous versus the tissue-specific expression of genes enables greater understanding of organismal development, complexity and evolution at the systems level. Large scale gene expression profiling has been done on a small number of organisms [88-93]. In fish, gene expression atlases were characterized in only few model species [94, 95]. Identification of housekeeping versus tissue-specific genes provides important molecular information that is needed for genetic improvement of fish for food production and for biomedical research purposes.

Salmonids underwent an evolutionarily recent whole genome duplication event and are in the process of returning to a diploid state [96]. Therefore, some fundamental scientific questions can be explored by decoding the rainbow trout transcriptome including how many genes exist in the rainbow trout, which genes are ubiquitously expressed and which genes and splice variants are uniquely expressed in each tissue to provide tissue specificity. In addition to the fundamental knowledge, this information can be used for the genetic improvement of rainbow trout for aquaculture by eliminating the need to positionally clone genes, facilitating resequencing to identify genetic variants, and identifying candidate genes for traits of interest.

To address the questions above, this study sequenced and *de novo* assembled the rainbow trout transcriptome from 13 vital tissues. High throughput Illumina sequencing in conjunction with the Trinity assembly package were used to: (1) sequence the rainbow trout transcriptome to provide a reference sequence, (2) functionally annotate the transcripts, (3) characterize digital gene expression and alternative splicing in 13 vital tissues; and (4) identify full-length cDNAs in the rainbow trout genome. Illumina sequencing in conjunction with Trinity assembly provided an efficient approach for *de novo* assembly and characterization of the transcriptome with high depth and width of coverage. Results of the *de novo* approach, used in this study, were compared to results of the gene models approach that was previously used in annotating the genome sequence [25].

## Materials and Methods

### Ethics statement

The fish used for this study was reared and euthanized under protocol #02456 approved by the Washington State University Institutional Animal Care and Use Committee.

### Production of doubled haploid rainbow trout

The rainbow trout from the Swanson clonal line used in the study was produced at the Washington State University (WSU) trout hatchery using previously described techniques [81, 82, 97, 98]. First generation homozygous rainbow trout were produced by androgenesis using gamma irradiation of eggs prior to fertilization [81, 82] and by gynogenesis by blockage of first cleavage using hydrostatic pressure shock [81, 82, 98]. When fish reached sexual maturity, homozygous clones were produced by collecting sperm from homozygous males and doing another cycle of androgenesis, or by stripping the eggs from homozygous androgenetically or gynogenetically produced females and performing gynogenesis by retention of the second polar body [98].

### Tissue collection and RNA isolation

Thirteen different tissues were collected from a single immature (2-year old, 250 g) male homozygous rainbow trout of the Swanson clonal line. Tissues collected were brain, white muscle, red muscle, fat, gill, head kidney, kidney, intestine, skin, spleen, stomach, liver, and testis. Tissues were quick-frozen in liquid nitrogen and were shipped to WVU from

WSU in dry ice. Tissues were kept at -80°C until RNA isolation. Total RNA was isolated from each tissue using TRIzol™ (Invitrogen, Carlsbad, CA) according the manufacturer's procedure as previously described [71] .

### Illumina paired-end sequencing

Construction of RNA-Seq libraries and sequencing on an Illumina Genome Analyzer IIx was performed at Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign. RNA-Seq libraries were constructed with the mRNA Sequencing Sample Preparation Kit (Illumina, San Diego, CA). Briefly, polyA+ messenger RNA was selected from 1 µg of RNA with magnetic oligo (dT) beads, chemically fragmented and converted to cDNA with random hexamers. Double stranded cDNAs were end-repaired, and the 3'-ends were A-tailed followed by ligation of Illumina sequencing and amplification adapters randomly to the ends. The adaptor-ligated cDNAs were loaded onto 2% agarose E-gels® (Invitrogen, Carlsbad, CA) and the fraction containing 200-500 bp was excised. Size-selected cDNAs were amplified by PCR with primers that introduced unique barcodes to each library. The final libraries were quantitated with Qubit (Life Technologies, Grand Island, NY) and the average size was determined on an Agilent bioanalyzer DNA7500 DNA chip (Agilent Technologies, Wilmington, DE) and diluted to 10 nM. The 10 nM dilution was further quantitated by qPCR on an ABI 7700. Each library was loaded onto one lane of an 8-lane flowcell for cluster formation and sequenced on an Illumina Genome Analyzer IIx according to the manufacturer's protocols (Illumina, San Diego, CA). The libraries were sequenced from both ends of the molecules to a total read length of 100 nt from each end. The FASTQ files were generated with Casava version 1.6.

### Trinity assembly and annotation

All 13 lanes of Illumina paired-end data were used to run Trinity assembler with default parameters. The Trinity software package combines three assembly algorithms: Inchworm, Chrysalis and Butterfly [99]. Assembly algorithms were run in C++ (Inchworm and Chrysalis) and Java (Butterfly) scripts. FASTQ formatted sequencing reads were converted into FASTA format by Fastool software, and extraction and computation of k-mer abundance from the sequencing reads were done by Jellyfish software. During assembly of contigs by Inchworm, minimum k-mer threshold abundance was set to 1 (default). The program was run at default parameters to cluster the Inchworm contigs into components (min_glue <int> =2, min_iso_ratio <float>=0.05 and glue_factor <float>=0.05). Transcript reconstruction from a deBruijn graph by Butterfly was also performed at default parameters (max_number_of_paths_per_node <int>=10, group_pairs_distance <int>=500, path_reinforcement_distance <int>=75, lenient_path_extension=1). Trinity contigs that were more than 500 nucleotides long were BLASTed against NCBI non-redundant (NR) protein database. The longest transcript of each Trinity contig group that matched a given protein in the NR database was selected as a representative sequence for each contig group.

### ORF/full-length cDNA prediction and gene ontology analysis

All representative transcripts selected from contigs having hits to the NCBI NR protein database were analyzed by ESTScan [100] to search for an open reading frame (ORF), which distinguishes coding and non-coding sequences [100, 101]. Whenever an ORF began and ended within a contig, it was considered as full length. If an ORF began at the first base or ended at the last base, it was not considered as full length. In addition, TransDecoder [http://transdecoder.sf.net] was used to identify ORFs with complete coding sequences. Gene ontology analysis was performed by BLASTx search against the NCBI NR protein database using the Blast2GO suite [102]. Blast2GO analysis provides a controlled vocabulary to describe gene product characteristics in three independent ontologies: biological process, molecular function, and cellular component [103, 104].

### Identification of housekeeping and tissue-specific genes

Housekeeping and tissue-specific genes were identified using a CLC genomics workbench. A total of 44,990 transcripts selected as representative sequences for each contig group from all 13 tissues were used as a reference sequence. Reads from each tissue (two libraries from each tissue) were mapped against the reference. Transcripts with RPKM (Reads per kilo base per million) value ≥1 in all 13 tissues were defined as housekeeping genes. For the tissue-specific genes, expression level of a gene in a particular tissue was compared to its expression level in all remaining 12 tissues. For distinction of tissue-specific genes, the fold-change in expression level was set as ≥ 8 fold, i.e. genes with an expression level in one tissue that is equal to 8 fold or higher than the maximum value in any of the other 12 tissues. As explained above, a single doubled haploid individual was used in this study to overcome the assembly bioinformatics challenges of the trout duplicated genome. Therefore, inferences regarding the housekeeping and tissue-specific gene expression should be considered with caution because results may be limited to this fish and to the conditions and time period during which the tissues were collected.

### Complexity and composition of tissue specific transcriptome

Sequence reads from each tissue were mapped to the 44,990 transcripts used as a reference sequence in this study. After mapping, numbers of genes expressed in each tissue were reported at four different threshold RPKMs (5, 1, 0.5 and 0.1). Transcripts having an RPKM value above the threshold were counted to obtain the number of genes expressed in each tissue. The mRNA abundance of the tissue-specific genes were calculated by dividing the sum of RPKM values of the tissue-specific genes by the sum of RPKM values of all genes expressed in that particular tissue (at an RPKM threshold of 0.5). A similar method of comparing the composition and complexity of tissue-specific transcriptomes was employed by Jongeneel and coworkers [105]. A multivariate Principal Component Analysis (PCA) analysis was applied to cluster tissues types according to gene expression patterns using a CLC genomics workbench.

### Assessment of the assembled rainbow trout transcriptome

Reference proteome sets of seven model fish species with known reference genome (*Danio rerio*, *Oreochromis niloticus*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Gadus morhua*, *Gasterosteus aculeatus,* and *Oryzias latipes*) were downloaded from the Uniprot database. Rainbow trout protein coding sequences resulting from the Trinity assembly were searched against the reference proteome of each fish species by BLASTx with a cut off E value of 1.00E-10. To obtain the expected range of sequence conservation between model fish species, cDNA sequences of model fish species were downloaded from the NCBI database. The cDNA sequences of each fish species were searched against the reference proteome set of the other model fish species by BLASTx with a cut off E value of 1.00E-10.

### Genome read mapping, annotation and assessment of alternative transcription/splicing

Alternative transcription/splicing events were assessed using the Bowtie2, TopHat and Cufflinks software package [106, 107]. First, a rainbow trout draft genome assembly was downloaded from http://www.genoscope.cns.fr/trout-ggb/data/ [25]. Then, sequence reads from all 13 tissues were mapped to the genome reference using Bowtie2/TopHat. Cufflinks was used to generate a transcriptome assembly for each tissue using alignment files from TopHat. Assemblies were then merged together using the Cuffmerge utility. Reads and the merged assembly were then analyzed using Cuffdiff to identify alternative transcripts (produced by alternative splicing/start sites) from each genomic locus (gene).

To identify novel genes, gene loci predicted by Cufflinks were filtered against the trout genome annotated loci first by BLASTn against the mRNAs (E-value $10^{-5}$) then by comparing the genome annotation coordinates (gtf files) using in-house script. TargetIdentifier [108] and TransDecoder [http://transdecoder.sf.net] were used to determine novel genes with ORFs. In addition, an in-house software (available upon request) was used to determine novel genes with 80% and 100% match to the NR database at an E value 10-3.

BLAT [109] with default parameters was applied to map the Trinity transcripts to the reference genome. The pslReps programs in the BLAT suite was used to select the best alignments for each query sequence. BLAT hits were classified based on the percentage of sequence identity covering the reference coding sequence at 100%, 90% and 50% of the entire coding sequence.

## Results and Discussion

### Illumina sequencing and Trinity assembly

To improve assembly and annotation of the rainbow trout reference transcriptome, libraries were constructed from a single double-haploid individual of the Swanson homozygous clonal line that has been used in sequencing the rainbow trout genome [25, 83] and in our previous transcriptome assembly [71]. Total RNA was isolated and sequenced from 13 different tissues of vital importance to fish life. These tissues were brain, white muscle, red muscle, fat, gill, head kidney, kidney, intestine, skin, spleen, stomach, liver and testis.

To maximize transcript coverage, cDNA libraries were sequenced on 13 separate lanes of an Illumina's Genome Analyzer using a paired-end protocol, yielding a total of 1.167 billion paired-end reads (100 bp). The cDNA library and sequencing information is given in Table 3. To allow identification of housekeeping and tissue-specific gene expression, sequences were generated from non-normalized libraries from different tissues. To facilitate the assembly, sequence reads were preprocessed to remove artifacts including sequencing adapters, low complexity reads and near-identical reads to improve read quality and efficiency of assembly [110].

Table 3: cDNA library information and summary of the high-throughput sequencing yield.

| | **Tissue** | **Number of reads** |
|---|---|---|
| 1 | Red Muscle | 93,064,168 |
| 2 | Skin | 87,743,778 |
| 3 | Fat | 93,546,068 |
| 4 | Brain | 84,816,430 |
| 5 | Gill | 92,670,670 |
| 6 | Spleen | 93,532,200 |
| 7 | Head kidney | 92,168,818 |
| 8 | Liver | 85,281,910 |
| 9 | Stomach | 91,231,186 |
| 10 | Intestine | 91,613,688 |
| 11 | Testis | 85,389,746 |
| 12 | White Muscle | 86,643,770 |
| 13 | Kidney | 89,642,288 |

RNA-Seq data were *de novo* assembled using the Trinity assembly package which comprises combining sequence reads into larger contigs (by Inchworm), clustering contigs into a component (by Chrysalis), and producing the most plausible sets of transcripts from these groups (by Butterfly) [99]. An assembly of 1.167 billion paired-end reads gave 1,371,544 Inchworm contigs (contig length > 200bp, ave = 744 bp). Inchworm contigs longer than 500 nucleotides (474,524 contigs) were used for downstream analysis. Assembly statistics and length distribution of contigs are given in Table 4 and Figure 3. These Inchworm contigs were clustered into a set of connected components to construct deBruijn graphs for assembly components. Each component defines a collection of contigs that are derived from alternative splicing or closely related paralogs [99]. These contigs were categorized into 163,411 components. Of them, 57,467 components contained more than one contig, while the remaining 105,944 were single contig components. The Trinity assembly package was used based on previous studies done in model species that suggest better performance of Trinity over some other assemblers, its ability to construct full-length transcripts, and the quality of the constructed transcripts [99, 111].

Table 4: Assembly statistics of Illumina paired-end data.

| | **All contigs** | **Long contigs (≥ 500 nt)** |
|---|---|---|
| Number of bases | 1,020,368,806 | 753,301,781 |
| Number of contigs | 1,371,544 | 474,524 |
| N50 (nt) | 1,369 | 2,188 |
| Largest contig length (nt) | 54,460 | 54,460 |
| Smallest contig length (nt) | 201 | 500 |
| Average contig length (nt) | 744 | 1,587 |

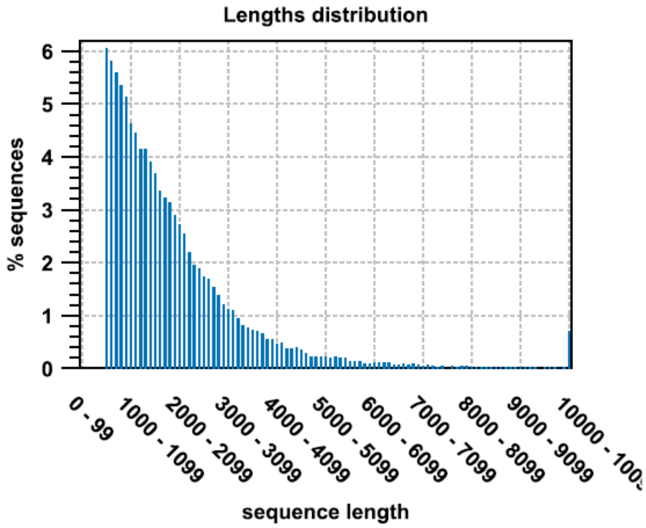


Figure 3: Distribution of contig (≥ 500 nt) length of a rainbow trout Illumina/Trinity transcriptome assembly.

All 474,524 Trinity contigs longer than 500 nucleotides were searched against the NCBI non-redundant (NR) protein database. A total of 287,593 (60.60%) contigs had hits to the database proteins. Importantly, 92.5% (266,188) of these contigs were part of the components with more than one contig, indicating the existence of a large number of transcript variants possibly due to alternative splicing, variable transcription start or termination points, or paralogous loci.

One of the remarkable findings of the project was the failure of a significant number of contigs (39.40% of 474,524 contigs) to have hits to the NR database, a finding similar to that observed previously in rainbow trout [112]. Similarly, in a catfish EST project Wang et al (2010) reported over 40,000 unique catfish sequences containing ORFs had no significant hits to the NCBI protein database [113]. Likewise, three transcriptomes from Antarctic notothenioid fish revealed 38-45% significant BLASTx hits in the NR protein database [114]. The unmatched contigs were used to identify a large number of non-coding RNAs [115]. In addition, the unmatched contigs may result from mistakes in assembly (contigs from reads with sequence errors) [99], lack of protein sequences of related fish in the database, or "trout-specific" diverged sequences due to the whole genome duplication [116, 117].

Previously, we utilized Sanger-based and 454-pyrosequencing approaches for transcriptomic analysis of the rainbow trout [71]. Figure 4 shows comparisons of the total number of sequenced bases, number of contigs, number of long contigs (≥500 bp), and average length of contigs obtained from Illumina, Sanger-based, and 454-pyrosequencing techniques. Compared to Sanger based and 454-pyrosequencing, Illumina allowed more effective assembly of the transcriptome with tremendous increases in the total number of contigs, total number of long contigs (>500 bp), and average length of contigs. However, the percentage of long contigs (>500 bp) was only 34.59% in the current Illumina/Trinity assembly compared to 56% in the 454-pyrosequencing assembly, which may be attributed to longer sequence reads with 454-pyrosequencing (Figure 4).

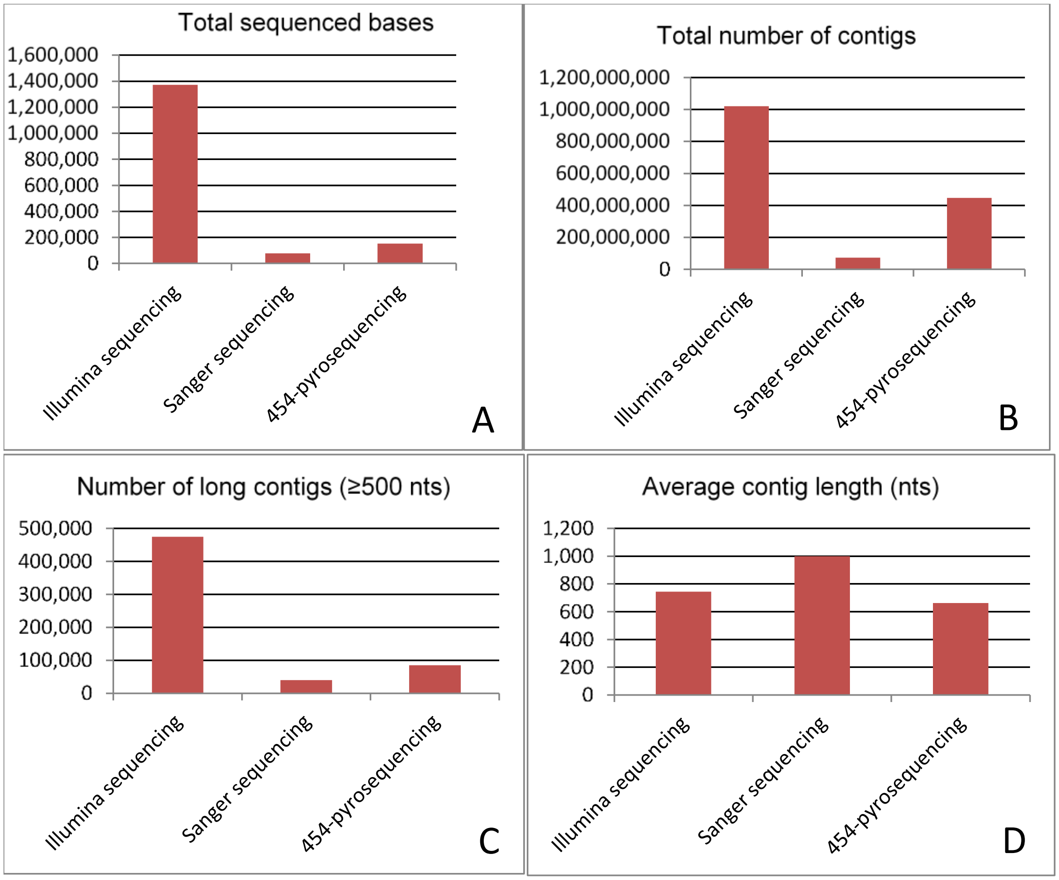


Figure 4: Comparison of total number of sequenced bases (A), total number of contigs (B), number of long contigs (≥500 bp) (C) and average length of contigs (D) obtained from Illumina, Sanger-based and 454-pyrosequencing techniques. Data on Sanger-based and 454-pyrosequencing techniques were obtained from Salem *et. al* [25].

### Gene identification and annotation

Transcript annotation was performed by BLASTx similarity search of the Trinity contigs against the NR protein public database. All contigs that had matches in the NR database were further analyzed to select a set of transcripts that could be used for functional genomics downstream analysis and ORF searching. For contigs that belonged to multiple contig components, the longest contig in a component was selected as a reference transcript of each component. For the single contig components, the longest contig was selected when more than one contig had aligned to any database protein with the same gene annotation. After removal of redundant transcripts, 44,990 were selected as a reference set of transcripts, including 34,260 contigs from multiple contig components and 10,730 contigs from single contig components. Of the total 44,990 representative contigs, ESTScan detected 43,824 (97.4%) sequences as having coding regions. The average length and number of the representative contigs is close to those predicted in the rainbow trout genome, 1.97 kb, versus 1.64kb and 44,990 versus 46,585 in the Trinity assembly and the rainbow trout genome, respectively [25]. In a catfish EST project, a 1.29 kb average length was observed and 98% of the unique sequences with significant hits to a protein database had ORFs [113]. About 2.6% of the contigs in this study (1,166) contained no coding regions (data not shown). These transcripts may represent pseudogenes or transcripts with intron-retaining cDNAs. Most of the contigs having hits to the NR database (97.49%) were identified within coding regions, which supports the credibility of the sequence assemblies.

So far, the international effort of sequencing the rainbow trout transcriptome has led to the discovery of 136,979 UniGenes (NCBI UniGene downloaded August, 2014), 1,610 genes and 13,166 proteins that are available in the public NCBI database [57]. Coding sequences were annotated in a recent assembly of the rainbow trout genome [25], however, UniGene sequence information is not yet updated at NCBI. The number and average length of the rainbow trout protein coding transcripts identified in this study (44,990 transcripts; 1.97 kb) are similar to the number and average length of UniGenes from model fish species (Figure 5). For example, zebra fish has 53,558 transcripts with a 1.04 kb average length. These data suggest that this sequencing project has captured the vast majority of the rainbow trout transcriptome. The protein coding Trinity transcripts are available at the USDA/NAGRP website http://www.animalgenome.org/repository/pub/MTSU2014.1218/

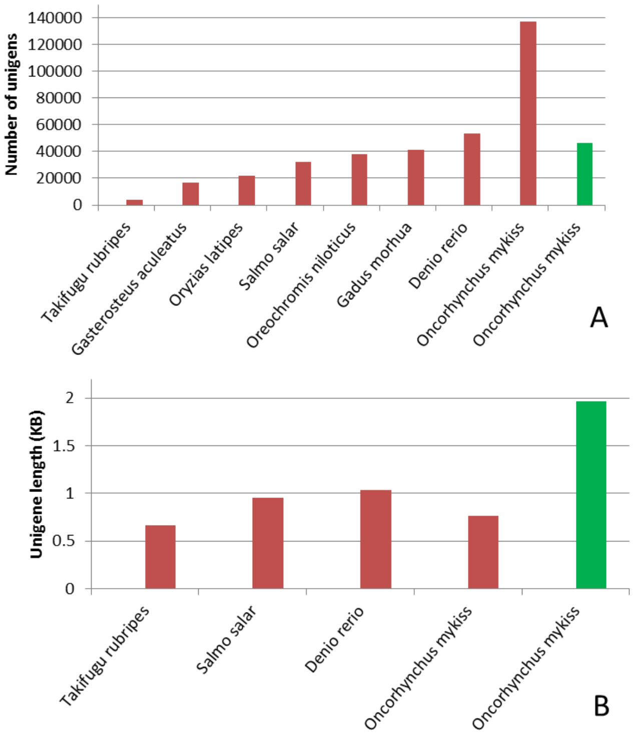


Figure 5: A: Number of UniGenes of model fish species and rainbow trout UniGenes that are available in the NCBI database (red bars) compared with number of rainbow trout protein coding transcripts obtained from Illumina sequencing (green bar). B: Average length of UniGenes of model fish species and rainbow trout UniGenes that are available in the NCBI database (red bars) compared with average length of rainbow trout protein coding

(Figure 5 cont.) transcripts obtained from Illumina sequencing (green bar). High number and short length of rainbow trout UniGenes suggest incomplete partial sequences. Illumina sequencing and Illumina/Trinity assemble resulted in 44,990 protein-coding transcripts with an average length of 1.97 kb, which is very close to number and average length of UniGenes in model fish species.

Grabherr *et. al.* found that Trinity was more sensitive than some other assemblers (Trans-ABySS, SOAP, Cufflinks and Scripture) in terms of percentage of full-length transcript reconstruction [99]. In another study comparing *de novo* assembly by various assemblers (SOAPdenovo, ABySS, Trans-ABySS, Oases and Trinity), Trinity assembly gave the highest (90%) RMBT value (Reads that can be mapped back to transcripts) and that the Trinity transcripts aligned better to the reference genome, indicating high quality of the transcripts [111]. One reason for the high quality of the transcripts constructed by Trinity may be its use of a fixed k-mer approach. In a previous study, Zhao *et. al.* found an increase in frequency of incorrect assemblies and artificially-fused transcripts by applying a multiple k-mer approach to the assemblers [111].

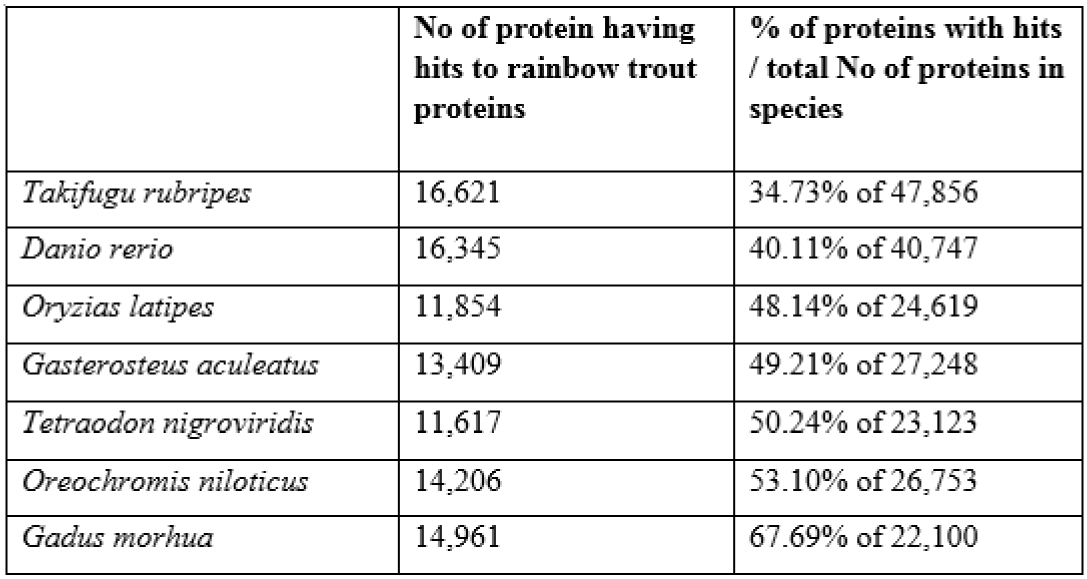
**Prediction of full-length cDNAs**

Illumina sequencing in conjunction with Trinity assembly provided a platform for identification and characterization of full-length cDNAs without the need for laborious cloning/primer walking approaches. Putative gene identification was done first by BLASTx against the NR protein database and then by identification of coding regions using ESTScan. ESTScan uses a Markov model to recognize the bias in hexanucleotide usage that exists in coding regions compared to non-coding regions [100]. In the context of this work, whenever an ORF began and ended inside a contig it was considered as full-length cDNA. This means if the ORF began at the first base and ended at the last base, it was not considered as full length. A total of 15,736 putative full-length cDNAs with an average length of about 2.4 kb were identified. In addition, TransDecoder [http://transdecoder.sf.net] identified 25,705 unique transcripts with complete coding sequences. Full-length transcripts identified by the ESTScan and TransDecoder were aligned to the reference genome using BLAT [109]. There were 9,000 (57.2%) and 14,213 (55.3%) unique transcripts mapped at 90% of their total length, respectively. The average lengths of the full-length cDNAs were more than that of Atlantic salmon obtained from ESTs using TargetIdentifier (17,399 cDNAs with average length 1.36 kb). The same study reported 10,453 full-length cDNAs from the 51,199 rainbow trout ESTs [118]. A well-characterized full-length cDNA set from rainbow trout will be necessary for the annotation of the rainbow trout genome sequences as well as for comparative, structural and functional genomics studies.

### Assessment of the sequenced rainbow trout transcriptome

In order to assess the level to which the rainbow trout transcriptome has been captured, the 44,990 reference transcripts were BLASTx searched against reference proteome sets of seven different model fish species with known reference genomes. Out of 44,990 reference transcripts, a total of 30,880 (68.3%) sequences matched to protein sequences of all seven fish species and 37,753 sequences (83.9%) matched to protein sequences of at least one fish species with a cut off E value of 1.00E-10. These findings suggested a high degree of sequence conservation and homology with these fish species. Variable numbers of significant hits were identified within each species; *Danio rerio* (40.11%), *Oreochromis niloticus* (53.10%), *Takifugu rubripes* (34.73%), *Tetraodon nigroviridis* (50.24%), *Gadus* *morhua* (67.69%), *Gasterosteus aculeatus* (49.21%) and *Oryzias latipes* (48.14%) with cut off E values of 1.00E-10 (Table 5). Similar levels of homology to model fish species were reported in a catfish EST project (54% to 57%) [113] and a common carp transcriptome study (47.7% to 54.2%) [119]. To allow a fair comparison of the rainbow trout protein coverage with that expected between fish species with complete known reference genomes, cDNA sequences from each fish species were searched against complete reference proteome sets of other fish species using BLASTx search with a cut off E value of 1.00E-10. *Gadus morhua* cDNA sequences had hits to 64.97% (15,022 out of 23,118) proteins of *Tetraodon, Takifugu rubripes* sequences had hits to 64.45% (17,775 out of 27,576) proteins of *Gasterosteus aculeatus* and *Danio rerio* sequences had hits to 66.43% (17,779 out of 26,763) proteins of *Oreochromis niloticus* (data not shown). Since rainbow trout protein coverage observed in this study is within the expected range, we anticipate that the project has captured the vast majority of the rainbow trout transcriptome.

Table 5: Summary of BLASTx search analysis of rainbow trout sequences against different model fish species with known reference genomes



| | No of protein having hits to rainbow trout proteins | % of proteins with hits / total No of proteins in species |
|---|---|---|
| *Takifugu rubripes* | 16,621 | 34.73% of 47,856 |
| *Danio rerio* | 16,345 | 40.11% of 40,747 |
| *Oryzias latipes* | 11,854 | 48.14% of 24,619 |
| *Gasterosteus aculeatus* | 13,409 | 49.21% of 27,248 |
| *Tetraodon nigroviridis* | 11,617 | 50.24% of 23,123 |
| *Oreochromis niloticus* | 14,206 | 53.10% of 26,753 |
| *Gadus morhua* | 14,961 | 67.69% of 22,100 |

**Functional annotation and gene ontology analyses**

Gene ontology provides organized terms to describe characteristics of gene products in three independent categories: biological processes, molecular function, and cellular components [103, 104]. Functional annotation of the Illumina/Trinity transcriptome contigs was performed by BLASTx search against the NCBI NR protein database using the Blast2GO suite [102]. The BLAST result findings were used to retrieve the associated gene names and Gene ontology (GO) terms in all three areas of ontologies. BLASTx results showed that biological processes constituted the majority of GO assignment of the transcripts (22,416 counts, 49%), followed by cellular components (12,793 counts, 28.1%), and molecular function (10,325 counts, 22.67%). The biological processes category showed that 18% of the rainbow trout genes were associated with cellular processes, 16 % with metabolic processes, and 14% with biological regulation (Figure 6). The molecular function category showed that 49% of the genes were associated with binding and 30% with catalytic activities. Of the cellular components, 46% of the rainbow trout genes were components of the cell and 27% were related to cellular organelles (Figure 6).

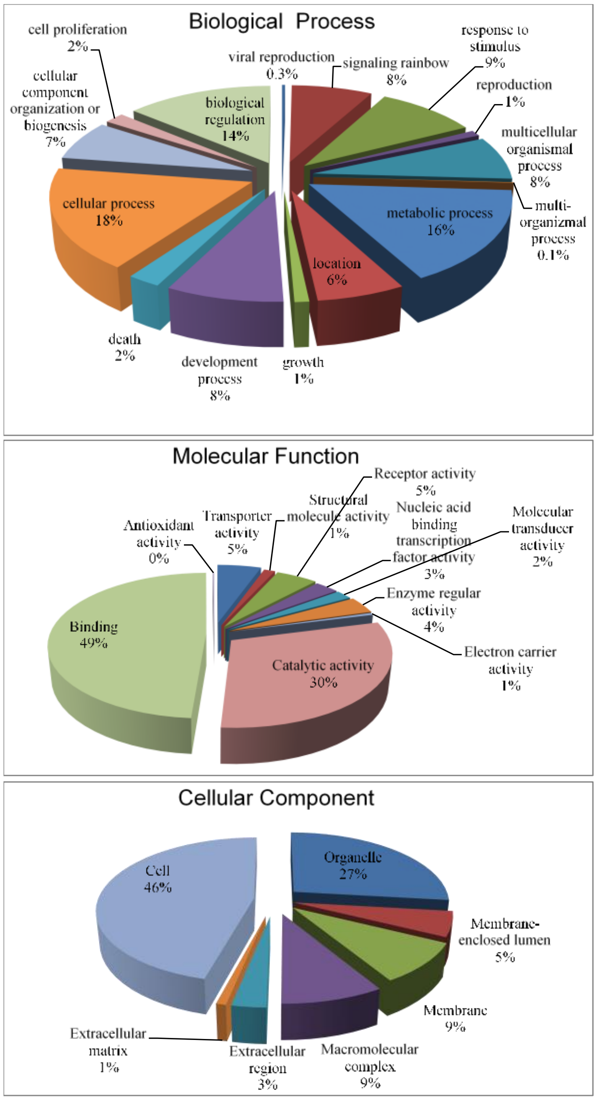


Figure 6: Gene Ontology (GO) assignment (2nd level GO terms) of the rainbow trout of 13 lanes of Illumina Trinity assembly.  Biological processes constitute majority of GO assignment of the transcripts (22,416 counts, 49%), followed by cellular components (12,793 counts, 28.1%) and molecular function (10,325 counts, 22.67%).

Previously, we performed functional annotation of rainbow trout transcripts sequenced using Sanger based and 454-pyrosequencing techniques [71]. Compared to the Illumina/Trinity assembly, there were some noticeable differences in distribution of genes in all three areas of ontologies (data not shown). The most noticeable difference was observed in distribution of genes in biological process. As an example of the previous assembly, in the biological process category the highest number of transcripts were associated with biological regulation and cellular processes (25% each) followed by metabolic processes (18%). Similarly, in the molecular function category, a larger number of transcripts was found to be associated with binding function (46%) than with catalytic activity (32%). In the cellular component category, transcripts associated with the cell and organelles were 59% and 24%, respectively. Possible reasons for these differences may include variations in nature of cDNA libraries (non-normalized in this assembly versus normalized in the previous assembly) and number of sequences used to retrieve GO terms (161,818 versus 44,990). In addition, Illumina data have higher coverage and are expected to be more representative of the transcriptome. These dissimilarities may have resulted in differences in the number and types of genes captured by the sequencing projects, which might have resulted in slightly different GO distribution profiles.

**Taxonomic analysis**

BLASTx top-hit species distribution of the gene annotations showed the highest number of matches to Nile tilapia (*Oreochromis niloticus*) followed by Zebrafish (*Danio rerio*) and Atlantic salmon (*Salmo salar*) (data not shown). Other fish species in the BLASTx top-hit list were Japanese puffer fish (*Takifugu rubripes*), puffer fish (*Tetraodon nigrovirdis*) and

European sea bass (*Dicentrarchus labrax*). Most of the species on the top hit list were fishes, suggesting high quality of the assembled genes and a high level of phylogenetic conservation of genes between rainbow trout and other fish species. As Nile tilapia showed high similarity to rainbow trout on the BLASTx top hit species distribution, the transcriptome of rainbow trout was compared to that of the Nile tilapia (Figure 7). Gene ontology for biological process and molecular function showed a homogeneous distribution of GO terms of transcripts between rainbow trout and Nile tilapia, suggesting that our transcriptome from Illumina/Trinity assembly represents all transcribed genes of rainbow trout. However, there were some slight differences in GO distribution of transcripts, especially in the =cellular component category (Figure 7). This variation in GO distribution may be attributed to differences in the sequencing approaches used for rainbow trout and Nile tilapia as well as their phylogenetic differences.

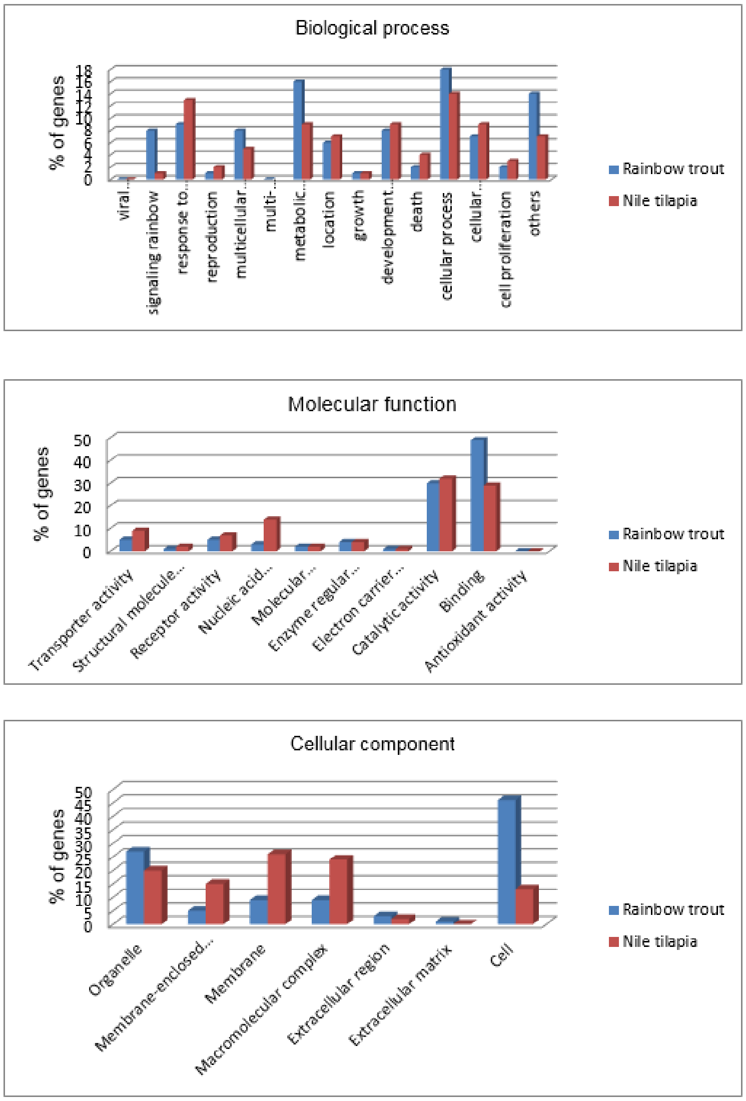


Figure 7: Gene Ontology (2nd level GO terms) comparison of rainbow trout and Nile tilapia. GO comparison shows a high resemblance of GO terms between rainbow trout and Nile tilapia (Oreochromis niloticus).

## Characterization of housekeeping and tissue-specific genes

An important outcome of this transcriptome sequencing project was identification of housekeeping and tissue-specific genes from 13 vital tissues. By mapping reads from each tissue to the Illumina/Trinity transcriptome reference, we identified a total of 7,678 (17.0%) housekeeping transcripts expressed in all 13 tissues with a minimum of 1 RPKM value in each tissue (Supplementary table S1) [120]. In comparison with mammals, a wide range of housekeeping gene percentages (1-38%) were reported in the mouse and human genomes using chip hybridization, MPSS (massive parallel signature sequencing) and next generation sequencing technologies [90, 105, 121]. Clearly, the differences are due to variations in technologies, number of tissues included, and nature of the duplicated rainbow trout genome.

Regarding the tissue-specific genes, a total of 4,021 transcripts with predominant expression in various tissues were identified in this dataset (Figure 8). The level of gene expression of each of these tissue-specific genes was at least 8-fold higher in one tissue relative to the rest of the tissues. Using these criteria, there was no tissue-specific gene that matches any housekeeping gene in the dataset. Testis expressed the highest number of tissue-specific genes followed by brain, gill, and then kidney. Conversely, liver expressed the lowest number of tissue-specific genes followed by spleen, skin, and then white muscle (Figure 8 and Supplementary table S2) [122]. A similar trend of tissue specificity was observed in the human and mouse genomes [121]. Examples of the highly expressed genes shown in Supplementary table S2 include two brain transcripts that had expression levels more than 30 fold higher than the rest of the tissues. Of them, metabotropic glutamate receptor-5 is involved in signal transduction for glutamatergic neurotransmission in the human brain [123, 124], and GABA (gamma-aminobutyric acid) receptor A is the principal inhibitory neurotransmitter in the mammalian central nervous system [125]. In skin, one of the three most highly expressed proteins is lily-type lectin which is a predominant protein in mucus of fish skin and provides important innate immunity [126, 127]. Similarly, myosins and troponins were among the most highly expressed tissue-specific transcripts predicted in muscle, both of which play important roles in muscle contraction. In red muscle, four transcripts characteristic of slow (red) muscle were identified (Slow myosin light chain, Troponin-I, Slow skeletal muscle, Slow troponin-T family-like, and Slow myosin heavy chain-1). The tissue-specific expression results warrant further work to reveal how expression patterns are regulated in different tissues and how the functions of genes are influenced by the cellular context.

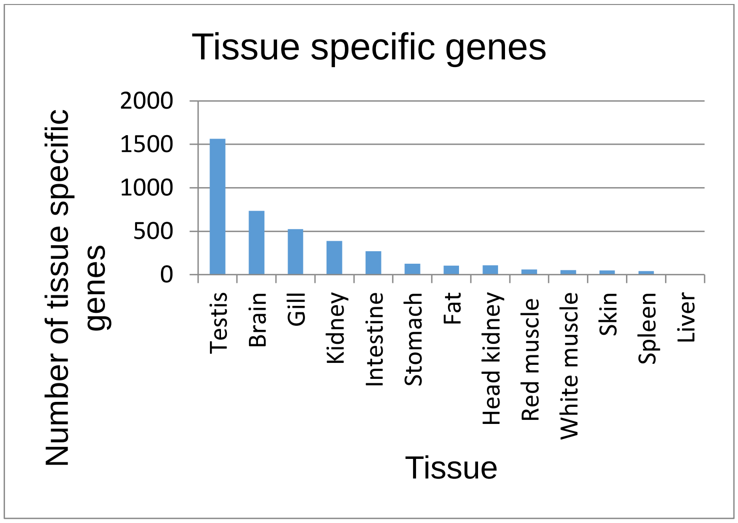


Figure 8: Number of tissue-specific genes predicted in different tissues. A transcript was classified as tissue-specific if it had expression level in one tissue that is ≥8 fold higher all other tissues.

Gene ontology comparison of housekeeping and tissue-specific genes showed differences in patterns of GO distribution. For example, in the molecular function category, the percentage of transcripts involved in the transport, receptor activities, and DNA binding were notably higher among tissue-specific genes than housekeeping genes (3.8%, 3.0%, 1.4% versus 1.2%, 0.7%, 0.7%; respectively). Conversely, the percentage of transcripts involved in protein binding was greater among housekeeping genes in comparison to tissue-specific genes (26.2% versus 11.2%; respectively). More than half of the DNA binding transcripts have tissue specific expression, similar to the proportion reported in humans [121]. Additionally, in the cellular component category relatively more tissue-specific transcripts were associated with plasma membrane than transcripts from housekeeping genes (1.1% versus 0.7%; respectively). Conversely, more genes connected with the nucleus, cytoplasm and mitochondrion were classified as housekeeping genes (3.3%, 2.6%, 2.2% versus 2.3%, 1.6%, 0.6%; respectively). Further, in the biological function category, there were more tissue-specific genes linked to signaling, developmental processes, and response to stimulus (2.6%, 6.6%, 0.7% versus 1.7%, 4.6%, 0.3%; respectively). Similar trends in gene ontology comparisons between tissue-specific and housekeeping genes have been reported in mammals [121].

Taken together, these data indicate the major biological role of the housekeeping genes in performing basic cellular functions needed to sustain life including metabolism, cellular processes, and biological regulation. However, tissue-specific genes were more involved in specialized functions such as signaling, responding to stimuli, development, organismal process, etc., suggesting diverse and specialized roles of tissue-specific genes in the cell.

## Complexity and composition of tissue-specific transcriptome

In an attempt to investigate the tissue complexity and composition of the rainbow trout transcriptome, the first question we asked was how many transcripts are expressed in a tissue? From 16,000-32,000 genes (at RPKM threshold of 0.5) were found to be expressed in the 13 studied tissues (Table 6). This range is slightly higher than what has been reported

(12,170) in various mammalian tissues using RNA-Seq data at the same RPKM threshold [121]. The difference may be attributed to the duplicated nature of the rainbow trout genome. Other studies utilizing non-RNA-Seq experimental techniques reported expression of about 10,000-30,000 genes in different mammalian tissues [128-130]. Our data suggested that expression of about 35-71% of total genes (at RPKM of 0.5) seems to account for all basic and specialized functions of the 13 studied tissues (Table 6). This expression level is marginally different from the level reported in humans (61%-84%) using MPSS, but at less stringent conditions (RPKM threshold of 0.3) [105].

Table 6: Number of genes expressed in 13 rainbow trout tissues at different RPKM threshold.

| Tissue | RPKM ≥5.0 | | RPKM≥ 1.0 | | RPKM≥ 0.5 | | RPKM ≥0.1 | |
|---|---|---|---|---|---|---|---|---|
| | Number of genes expressed | Fraction of total genes | Number of genes expressed | Fraction of total genes | Number of genes expressed | Fraction of total genes | Number of genes expressed | Fraction of total genes |
| White muscle | 2,949 | 0.06 | 10,798 | 0.24 | 15,970 | 0.35 | 27,593 | 0.61 |
| Red muscle | 6,425 | 0.14 | 18,991 | 0.42 | 24,136 | 0.54 | 33,079 | 0.74 |
| Head kidney | 7,461 | 0.17 | 19,699 | 0.44 | 24,368 | 0.54 | 32,022 | 0.71 |
| Skin | 6,646 | 0.15 | 20,951 | 0.47 | 27,796 | 0.62 | 38,669 | 0.86 |
| Spleen | 10,277 | 0.23 | 22,150 | 0.49 | 26,009 | 0.58 | 32,850 | 0.73 |
| Fat | 9,584 | 0.21 | 22,837 | 0.51 | 27,059 | 0.6 | 35,251 | 0.78 |
| Testis | 16,374 | 0.36 | 26,385 | 0.59 | 30,289 | 0.67 | 38,027 | 0.85 |
| Kidney | 12,253 | 0.27 | 25,856 | 0.57 | 29,964 | 0.67 | 36,783 | 0.82 |
| Gill | 13,804 | 0.31 | 26,149 | 0.58 | 29,757 | 0.66 | 36,440 | 0.81 |
| Brain | 11,464 | 0.25 | 27,151 | 0.6 | 32,053 | 0.71 | 39,697 | 0.88 |
| Intestine | 13,655 | 0.3 | 27,018 | 0.6 | 31,168 | 0.69 | 38,186 | 0.85 |
| Liver | 5,181 | 0.12 | 16,293 | 0.36 | 21,236 | 0.47 | 29,698 | 0.66 |
| Stomach | 6,982 | 0.16 | 19,462 | 0.43 | 24,460 | 0.54 | 33,807 | 0.75 |

The second question we asked is how various tissues differ in composition and complexity of their transcriptomes? Brain, testis and intestine had complex transcriptomes in that they expressed larger percentages of the genes in the genome (Table 6) with a small fraction of the mRNA pool contributed by the most highly expressed genes (Figure 9). On the other hand, white muscle and stomach had less complex transcriptomes, expressing fewer genes in the genome with a large fraction of the transcriptome contributed by the most highly expressed genes. As an example, the top hundred most highly expressed genes contributed 80% of the mRNA population in white muscle, while contributing only ~16% of mRNA pool in testis (Figure 9). Similar trends in transcriptome complexity were reported from previous studies in mammals [105, 121] suggesting conservation of the tissue-specific expression patterns. Conserved expression of more than a third of the core tissue-specific gene expression was reported across major vertebrate lineages [131].

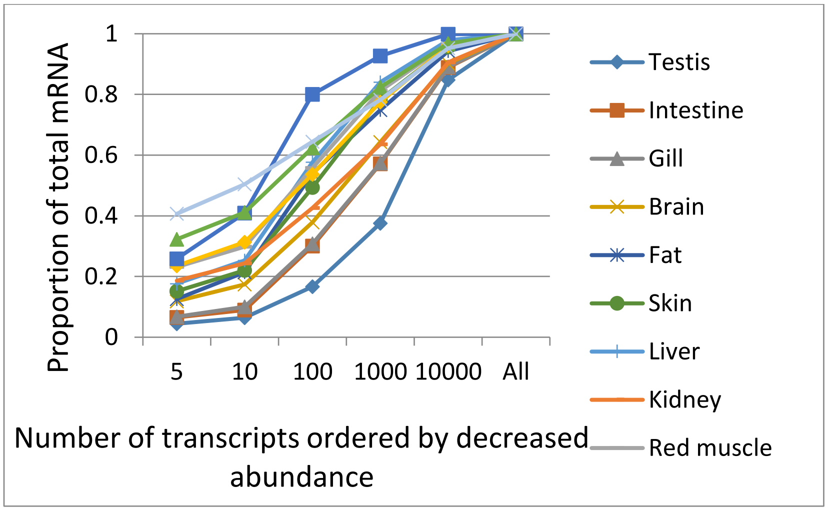


Figure 9: Distribution of gene abundance in various tissues. Proportion of the transcriptome contributed by the most abundant genes is plotted in various tissues. In

(Figure 9 cont.) testis, intestine, gill and brain, there was little contribution of the most highly expressed genes to the mRNA pool. Conversely, in white muscle, spleen, and stomach, large fraction of the transcriptome was contributed largely by the most highly expressed genes.

The third question we asked is what is the contribution of the tissue-specific genes to the transcription pool in different tissues? Stomach, white muscle and fat had high abundances of tissue-specific transcripts; and skin, liver, spleen, brain, kidney and intestine had low abundances of tissue-specific transcripts (Figure 10). Although stomach, white muscle, and fat expressed relatively fewer tissue-specific genes (51-127 genes), these transcripts significantly contributed to the total cellular mRNA pool (31-39% of total mRNA) (Figure 10 and Supplementary table S2). Conversely, in brain, kidney, and intestine, which expressed a large number of tissue-specific genes (734, 390 and 271 genes, respectively), these genes contributed only 2-3% of total cellular mRNA. These results indicate wide variation in the number of genes and regulation of gene expression that determine tissue specificity.

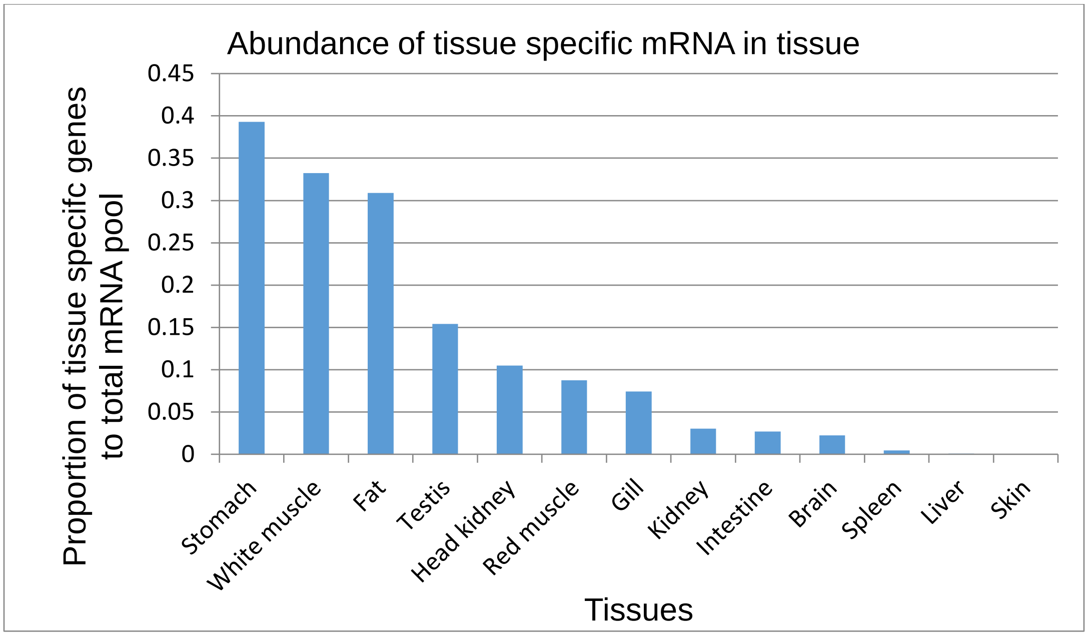


Figure 10: Transcript abundance of tissue-specific genes in various tissues. White muscle, stomach and fat showed high abundances of tissue specific transcripts; while, skin and liver exhibited low abundance of tissue-specific transcripts.

This complexity in the expression pattern of genes may be explained in terms of not only the degree of specialization but also the types of cells in each tissue. For example, brain has a variety of cells specialized for equally important but different functions. As different cell types express different cell-specific genes, tissue as a whole has a large collection of equally important tissue-specific genes expressed at comparable rates (Figure 10). In contrast, in fat, a majority of gene expression is directed to the manufacture of necessary enzymes to carry out basic fat metabolic pathways. Therefore, there is an abundance of a relatively small number of fat metabolic transcripts. The other possibility is that most of the cells in fat tissues are alike and the genes taking part in some important function may be expressed highly in all cells so that their mRNA population may be dominated in non-normalized libraries.

A multivariate Principal Component Analysis (PCA) analysis was applied to cluster tissues types according to gene expression patterns. Two dimensional covariance matrix of the different tissue samples revealed distinct expression of both the spleen and the kidney (Supplementary figure 1) [132]. Recently, we reported a detailed expression in the spleen transcriptome in rainbow trout [133]. The distribution of rest of the tissues were clearly classified into 2 clusters (head kidney, red muscle and stomach) and (testis, gill, fat, skin, intestine, brain, white muscle and liver).

### Comparison of the Trinity assembly to the reference genome annotation

Berthelot et al used a gene models approach based on both a genome and a transcriptome sequences to predict 46,585 annotated protein-coding genes [25]. To assess the *de novo* transcriptome assembly approach used in this study against the gene models approach used by Bethelot et al, we first ran a reciprocal BLAST search between the two datasets. A total of 4,146 contigs of the Trinity assembly (9.2%) including, 710 full-length sequences, did not match any mRNA sequences identified in the genome reference (BLASTn, E value > 1.00E-10). These contigs may represent unannotated, incomplete, or absent loci in the trout genome. On the other hand, 2,641 mRNAs sequences in the genome reference did not match any of the Trinity contigs. All teleost protein sequences were used, at least partially, to annotate the trout genome [25]. Therefore, some of these 2,641 missing transcripts may represent predicted gene models that are not expressed in rainbow trout, at least in the single individual used in this study.

In addition, we ran BLASTx of the two datasets against the zebrafish proteome (with a cut off E value of 1.00E-3, downloaded from Ensembl 11/17/2014). A total of 19,390 (44.9%%) of the zebrafish proteins had hits by at least one of the Trinity contigs, compared to 21,119 (48.9%) proteins in case of the trout genome mRNA sequences. There were 16,046 (39.6%) zebrafish protein hits shared between the two datasets. A total of 4,378 and 1,077 transcripts of the Trinity and the genome reference mRNAs had no hits to the zebrafish proteome, respectively. When the two datasets were compared by BLAST with proteome sequences of seven model fish species (with known genomes), there were 3,297 and 195 transcripts of the Trinity and the trout genome reference mRNAs with no hits, respectively. TransDecoder recognized 25,705 (57.1%) and 38,313 (82.2%) transcripts with complete ORFs in the Trinity and the trout genome mRNAs, respectively. Taken together, the comparison of *de novo* transcriptome assembly approach (used in this study) and the gene models approach used by Bethelot et al, indicate some differences in the transcripts/annotations identified by each method. It is worth mentioning that, in this study, the transcriptome was sequenced from the Swanson clonal line which is the same source used for the rainbow trout genome sequencing. However, a large proportion of the transcriptomic data used by Berthelot and coworkers to annotate the genome came from a different clonal line [25].

To assess the percentage of the mappable Trinity transcripts to the genome reference, Trinity transcripts were aligned to the reference genome using BLAT and then the best hits were selected using the pslReps program of the BLAT suite [109]. BLAT hits were classified according to the percentage of Trinity sequence identity covering the reference coding sequence of the genome. There were 1,434 (3.2%); 25,860 (57.5%) and 38,367

(85.3%), unique Trinity transcripts mapped at 100%, 90% and 50% of coverage, respectively. These results, at least partially, validate the Trinity assembly. However, the current version of the genome sequence is still not complete which prohibits a complete assessment of the Trinity assembly based on the BLAT results.

In an effort to find novel loci (not annotated) in the genome, sequence reads were mapped to the genome reference using TopHat and Cufflinks software packages [106]. A total of 223,751 gene loci were predicted with 286,561 potential transcripts (average of 1.28 transcripts/gene). These gene loci were filtered against the trout genome annotated loci first by BLASTn against the mRNAs (E-value $10^{-5}$) and then by comparing the genome annotation coordinates (gtf files) using an in-house script (available upon request). Using this approach a total of 78,592 novel loci were identified. Further investigation used TargetIdentifier [108] and TransDecoder [http://transdecoder.sf.net] to determine novel genes with ORFs. TargetIdentifier recognized 10,195 full ORFs and TransDecoder identified 12,652 ORFs with 3,420 complete ORFs. There were 1,432 transcripts, with complete ORF common between the TargetIdentifier and TransDecoder datasets. Using an in-house script based on a BLASTx to the NR database with and e value 10-3, there were 128 genes with 100% matches and 832 genes with 80% matches to the NR database not annotated in the reference genome. After redundant removal, 11,843 transcripts were recognized as new transcription loci. To provide a comprehensive list of all new transcripts that were identified in this study (not annotated in the trout genome), those 11,843 were screened to remove redundancy with the 4,146 contigs of the Trinity contigs that had no match with any mRNA sequences in the genome reference. A total of 14,827 (11,843+2,984) were counted as new transcripts. FASTA and annotation (gtf) files of those new transcripts are provided (Supplementary files S3 and S4) [134, 135] and available for download http://www.animalgenome.org/repository/pub/MTSU2014.1218/

### Comparison of the Trinity assembly to the marine rainbow trout transcriptome

The anadromous steelhead (*Oncorhynchus mykiss*) transcriptome was recently sequenced [73]. To assess gene expression associated with adaptation to ecological and environmental factors in the marine versus the freshwater rainbow trout, we ran a reciprocal BLASTn search. A total of 8,312 contigs of the Trinity assembly (18.4%) did not match any sequences in the marine rainbow trout (BLASTn, E value > 1.00E-3). On the other hand, 12,207 (9.3%) marine rainbow trout transcripts did not match any of the Trinity contigs. These results should be considered with caution because of the unbalanced amount of data (~1.167 billion paired-end reads [100bp] in the freshwater trout, compared to 41 million 76-mer reads in in the marine trout). Gene ontology comparison of the marine versus freshwater unmatched transcripts did not show significant gene enrichment for salinity adaptation (data not shown).

### Assessment of alternative transcription/splicing

Trinity assembler is capable of predicting alternative splicing events. There were a total of 287,593 Trinity contigs longer than 500 nucleotides that had hits to the NR protein database. A total of 92.5% (266,188) of these contigs were part of the components with more than one contig, indicating the contigs had alternative transcription/splicing. However, these contigs may also be separately expressed from paralogous genes.

Therefore, the TopHat and Cufflinks read mapping to the genome, described above, were used to assess the percentage of alternative transcription/splicing events. Out of 223,751 predicted genes, 27,471 (12.8.) genes had at least two transcripts from alternative transcription/splicing; 4,663 (2.08%) genes had five and more transcripts and 634 genes had 10 or more transcripts. A total of 1,064,892 exons were detected yielding an average of 4.75 exons/locus.

The low percentage of genes with alternative splicing is unexpected because alternative splicing is one of the important components adding functional complexity to vertebrates; in humans about half of the genes have at least one splice variant [136]. However, because of the whole genome duplication event in teleost fish, many genes have paralogous duplicates [137-139]. Indeed, gene duplication can lead to loss of alternative splicing of genes [140, 141] and many of the splice variants present in an ancestor are found to be expressed separately from duplicated genes in teleost fish [142]. The rate of alternative splicing was lowest (17%) in the highly duplicated genome of zebrafish compared to the compact genome of the pufferfish (43%) [143]. Availability of a complete and annotated sequence of the rainbow trout genome is needed to fully characterize transcripts representing splice variants and separately expressed sequences of paralogous genes.

## Conclusion

High throughput Illumina sequencing of non-normalized cDNA libraries from 13 tissues was used together with the Trinity assembler to generate a high-quality draft of the rainbow trout transcriptome. A single doubled haploid rainbow trout fish, from the same source used for the rainbow trout genome sequence, was used to address problems associated with the nature of the rainbow trout duplicated genome. Results of the *de novo* approach, used in this study, were compared to results of the gene models approach that was used in annotating the genome sequence. A total of 14,827 sequences were identified as new transcripts (not annotated in the trout genome). A digital gene expression atlas revealed 7,678 housekeeping and 4,021 tissue-specific genes. In addition, expression of 16,000-32,000 genes (35%-71% of the transcriptome) was revealed in various tissues. White muscle and stomach showed the least complex transcriptomes, with high fractions of their total mRNA expressed by a small number of genes.  In contrast, Brain, testis and intestine had complex transcriptomes with large numbers of genes involved in their gene expression.

# CHAPTER III: GENOME-WIDE DISCOVERY OF LONG NON-CODING RNAS IN RAINBOW TROUT [115]

## Introduction

Global gene expression data in different mammalian species have demonstrated that protein-coding sequences occupy less than 2% of the genome, and the vast majority of the genome is transcribed into non-coding RNAs [25, 34-36]. These non-coding RNA molecules include small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), microRNA (miRNA), small interfering RNA (siRNA), piwi RNA (piRNA), signal recognition particle (SRP) RNA and lncRNA. LncRNAs are defined as non-protein-coding RNAs greater than 200 nucleotides in length, distinguishing them from small non-coding RNAs [37, 39]. Based on their proximity to the protein-coding genes in a genome, lncRNAs are subdivided as genic (intronic or exonic with sense, antisense, and bidirectional orientation) or intergenic [144, 145]. Unlike small non-coding RNAs, lncRNA sequences are less conserved and are expressed at relatively low levels, and these characteristics make their computational identification and annotation difficult [43].

Like protein-coding genes, lncRNAs are often transcribed by RNA polymerase II and can be post transcriptionally modified by splicing, capping and polyadenylation [146-149]. In contrast to protein-coding  genes, a majority of lncRNA transcripts tend to have fewer exons [43] and a shorter transcript size (average of 800 nucleotides) [150]. LncRNAs usually exhibit highly cell- and tissue-specific expression patterns and sometimes they are uniquely localized to a specific cellular compartment [42, 151-153].

Even though a small number of lncRNAs have experimentally validated molecular functions, a substantial number of lncRNAs have been functionally annotated. LncRNAs are considered important gene regulators due to, at least, three important molecular roles; these RNAs serve as decoys, scaffolds or guides. Many lncRNAs serve as decoys that preclude access to DNA by regulatory proteins; this role affects transcription of protein-coding genes [154, 155]. Some lncRNAs regulate genes by acting as scaffolds to bring two or more proteins into a discrete complex [41, 156-158]. Other lncRNAs regulate different developmental and cellular processes by guiding a specific protein complexes to a specific promoter in response to certain molecular signals [159-161]. LncRNA mediated guidance of chromatin modifying proteins affects expression of neighboring genes (*cis*) or distant genes (*trans*) and there is evidence that even cis acting lncRNAs have ability to act in *trans* [162-164]. Beside transcriptional control, lncRNAs regulate many molecular processes including alternative splicing [165, 166], other post transcriptional processes [167], and mRNA transport [168].

Aquaculture of rainbow trout supplies a significant portion of aquatic food in the USA and worldwide. In addition to its importance as a food species, rainbow trout is one of the most widely used fish species as a model in biomedical research [19-22, 53, 54, 56, 169]. In order to improve aquaculture production and efficiency and facilitate biomedical research of involving rainbow trout, a great deal of genetic information has been accumulated for this species that includes a recently published initial draft of the genome [25] and several assemblies of the transcriptome [51, 71, 73]. However, a complete understanding of the trout's genome biology is still lacking. Recent studies in mammalian and non-mammalian species have resolved some long-standing mysteries in biology by functionally characterizing lncRNAs as important regulators of protein-coding genes [158, 170-174]. With growing interest in lncRNAs-mediated gene regulation, these RNAs have been characterized, genome-wide, in limited animal and plant species in recent years [38, 151]. And, our knowledge of lncRNAs in fish is still very limited [175]. Therefore, the objective of this study was to identify and characterize lncRNAs in rainbow trout genome and create a global gene expression atlas of lncRNAs in several vital tissues.

## Materials and Methods

### Data source

To facilitate lncRNA discovery in rainbow trout, four high-throughput sequence datasets were used in this study. 1) About 1.16 billion Illumina sequence reads as we previously described [51]. Briefly, 13 tissues including brain, white muscle, red muscle, fat, gill, head kidney, kidney, intestine, skin, spleen, stomach, liver and testis were sequenced from a single male-doubled haploid rainbow trout. Sequencing libraries were constructed using poly-A selection technique and cDNA libraries were sequenced using Illumina's paired-end protocol. Data were generated from a single doubled haploid individual to overcome the assembly bioinformatics challenges of the trout duplicated genome. 2) Similarly, about 0.75 billion Illumina single reads, used in annotating the rainbow trout genome and sequenced from a doubled haploid female rainbow trout, as previously described by Berthelot et al. [25]. Briefly, 13 vital tissues including (liver, brain, heart, skin, ovary, white and red muscle, anterior and posterior kidney, pituitary gland, stomach, gills) were sequenced. Sequencing libraries were constructed using poly-A selection technique and cDNA libraries were sequenced using Illumina's 101 base-lengths single read protocol. 3)

About 0.25 billion reads used in assembling the anadromous steelhead (Oncorhynchus mykiss) transcriptome by Fox et al. [73]. 4) About 89 million reads data set from redband trout (Oncorhynchus mykiss) by Narum *et al.* [176]. Data from Narum *et al.* were chosen because Ribo-Zero RNA-Seq libraries were sequenced to capture both the polyadenylated and the non- polyadenylated RNAs with information about transcript strand orientation.

**Computational prediction pipeline**

Sequencing reads were mapped to the genome reference [25] using the TopHat and Cufflinks software packages [106]. An in house Perl script was written to filter the transcripts shorter than 200 nt. Several stages of filtration were performed to remove protein-coding transcripts and small non-coding RNAs. First, transcripts were searched against NCBI nr protein database (updated on 10/01/2014). All the transcripts which had an open reading frame more than 100 amino acids were removed. Next, protein-coding calculator (CPC) was used to remove any remaining potential protein-coding transcripts (Index value <-0.5) [177]. To remove other classes of RNAs (tRNA, rRNA, snoRNA, miRNA, siRNA and other small non-coding RNAs) transcripts were searched against multiple RNA databases including genomic tRNA database, mirBase, LSU (large subunit ribosomal RNA) and SSU (Small subunit ribosomal RNA) databases [178-181]. Any transcripts which showed sequence similarity with any of these classes of RNAs with cut-off E value of ≤ 0.0001 were removed. After these filtration steps, putative lncRNA transcripts were searched against several noncoding-RNA databases to explore sequence similarity of putative rainbow trout lncRNAs transcripts to previously characterized lncRNAs in other species [175, 182-186]. All prediction steps were applied independently to the four transcriptome datasets. All putative lncRNAs from all four datasets were blasted against each other. LncRNA which were identified in at least 2 of the 4 datasets were chosen for further analysis. Data set from Narum et al., is the only one with information about strand orientation [176]. To ensure correct sense and antisense orientations of lncRNAs from the other three sources, their strand orientation was assigned by matching to counterparts from Narum and coworkers (based on sequence similarity match of more than 95% and same genomic location coordinates). A total of 54,503 non-redundant lncRNAs were identified in this dataset.

For the extra filtration steps, more stringently selected lncRNAs, any putative lncRNA containing ORF covering more 35% of its length or more than 83 amino acid were filtered out [187]. In addition, the cut-off value for the CPC [177] was decreased from -0.5 to -1.0. Further, if any lncRNA overlapped with more than 100 nt with another lncRNA from a different dataset, we filtered out the shortest lncRNA. Furthermore, any lncRNA that overlapped with a protein-coding gene in the sense orientation was removed. Lastly, any single-exon lncRNA that was adjacent to a protein-coding gene within 500nt was removed.

### Identification of tissue expression

For lncRNA tissue distribution, sequencing reads from 13 tissues were independently mapped to all putative lncRNAs and gene expression level were measured in terms of RPKM. House-keeping and tissue-specific genes were determined as we previously described [51].

### Gene clustering

Sequencing reads from each tissue were mapped to combined reference consisting of the lncRNAs and mRNAs from the rainbow trout genome reference [25]. Expression of lncRNAs and protein-coding genes was determined in terms of RPKM. Expression value of each transcripts in each tissue was normalized using a scaling method in CLC genomics workbench with mean as the normalization value. Normalized expression values of transcripts in each of the 13 studied tissues were used to cluster protein-coding genes and lncRNAs using a clustering feature in Multi-experiment Viewer (MeV) program [188, 189]. The minimum correlation threshold to generate clusters was 0.97.

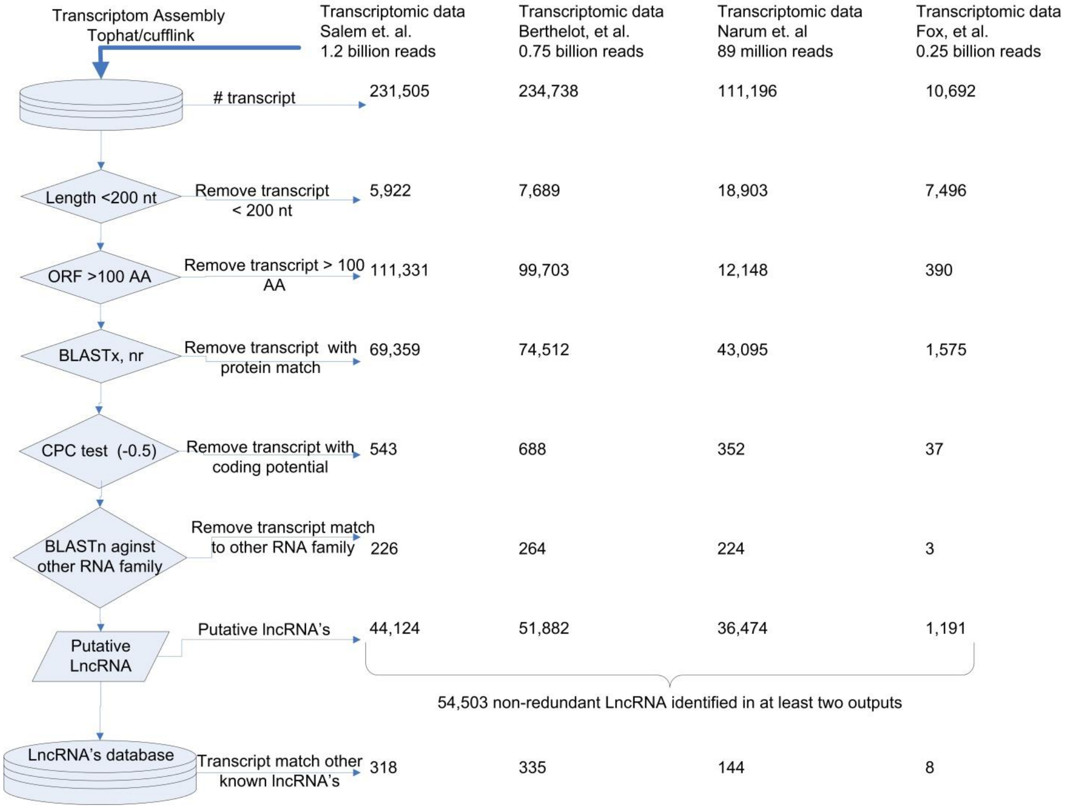
### Identification of genomic location of lncRNAs relative to neighboring protein-coding genes

LncRNAs were classified based on their intersection or relative location to protein-coding genes using in-house Perl scripts using the rainbow trout genome data (downloaded from **http://www.genoscope.cns.fr/trout/data/**).

## Results and Discussion

### Identification of putative lncRNAs in rainbow trout

The main objective of this study was to identify a comprehensive list of putative lncRNA genes in the rainbow trout genome. To accomplish this, we sequenced poly-A selected cDNA libraries using total RNA isolated from 13 tissues. Recently, we used the same sequencing data to identify protein-coding  transcripts in the trout genome [51].  In this study, sequence data for about 1.167 billion, paired-end reads (100 nt) were mapped against a reference rainbow trout genome using the Cufflink and TopHat software [106, 107], resulting in 231,505 putative transcripts. Several filtration steps were used to distinguish lncRNAs in the transcript list by removing the protein-coding transcripts, pseudogenes and other classes of non-coding RNAs including rRNA, miRNA, tRNA, snRNA, snoRNA (Fig 11). First, all transcripts shorter than 200 nt were removed, and then transcripts with an open reading frame (ORF) longer than 100 amino acids were filtered out. Next, remaining transcripts were BLASTx searched against the NCBI non-redundant protein database to eliminate transcripts with sequence similarity to known proteins at a cut off E-value of ≤ 0.0001. To further filter remaining protein-coding transcripts, we used the Coding Potential Calculator (CPC) software that assesses quality and completeness of query ORF to proteins in the NCBI database using six biologically meaningful sequence features [177]. These filtration steps left 44,350 transcripts from this data set that had very little or no evidence of protein-coding ability. Because most of the small non-coding RNAs like miRNA and tRNA are shorter than 200 nt, the first filtration step should be enough to remove most of the small non-coding RNAs. To confirm removal of any remaining small non-coding RNAs (tRNA, rRNA, snoRNA, miRNA, siRNA and other small non-coding RNAs), transcripts were searched against multiple RNA databases including genomic tRNA database, mirBase, and LSU (large subunit ribosomal RNA) and SSU (Small subunit ribosomal RNA) databases [178-181]. After application of the above filtration steps, we found 44,124 putative lncRNAs from our sequence data set (Salem et al., [51]). These lncRNAs exhibited little or no evidence of coding potential or belonging to other non-coding classes of RNA.



| | Transcriptomic data<br>Salem et. al.<br>1.2 billion reads | Transcriptomic data<br>Berthelot, et al.<br>0.75 billion reads | Transcriptomic data<br>Narum et. al<br>89 million reads | Transcriptomic data<br>Fox, et al.<br>0.25 billion reads |
|---|---|---|---|---|
| # transcript | 231,505 | 234,738 | 111,196 | 10,692 |
| Remove transcript < 200 nt | 5,922 | 7,689 | 18,903 | 7,496 |
| Remove transcript > 100 AA | 111,331 | 99,703 | 12,148 | 390 |
| Remove transcript with protein match | 69,359 | 74,512 | 43,095 | 1,575 |
| Remove transcript with coding potential | 543 | 688 | 352 | 37 |
| Remove transcript match to other RNA family | 226 | 264 | 224 | 3 |
| Putative lncRNA's | 44,124 | 51,882 | 36,474 | 1,191 |
| Transcript match other known lncRNA's | 318 | 335 | 144 | 8 |

Figure 11: Bioinformatics pipeline used in prediction of rainbow trout lncRNAs. LncRNAs were predicted from four different transcriptomic datasets, then all putative lncRNAs from all data were blasted against each other. A total of 54,503 non-redundant lncRNAs identified in at least 2 of the 4 data sets were chosen for further analyses in order to increase the confidence of lncRNA prediction. Vertical arrows are pointing toward the subsequent prediction and filtration steps of the workflow. First horizontal arrow pointing toward the right is referring to the number of initial transcripts predicted from the four datasets. Middle six horizontal arrows indicate the number of transcripts filtered at each step and the final horizontal arrow points to the number of putative lncRNAs with significant hits to noncoding-RNA databases from each dataset.

Because some of the lncRNAs are thought to be due to expression noise [190], we conceptualized that prediction of lncRNAs from different reliable data sources would be an important step in removing false lncRNAs. To achieve this goal, the same lncRNAs prediction pipeline was applied to discover putative lncRNAs from three other rainbow trout transcriptomic datasets that are available on NCBI (Fig 11). Those three sources were sequence data used by Berthelot et al. [25] in annotating the rainbow trout genome, a data set used by Fox et al. [73] in assembling the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome and a data set from redband trout (*Oncorhynchus mykiss*) that was reported by Narum et al. [176]. Data from Narum et al. were particularly useful because Ribo-Zero RNA-Seq protocols were used which allow sequencing both the polyadenylated and the non- polyadenylated RNAs. In addition, the strand orientation sequence information was preserved. From these three sequence data sources, a total of 0.75B reads, 89M reads, and 0.25B reads were used in the prediction pipeline that yielded 51,882; 1,191; and 36,474 putative lncRNAs in the three datasets, respectively. LncRNAs predicted in at least 2 of the 4 data sets were considered for the subsequent analyses. After removal of redundant transcripts, we had a total of 54,503 putative lncRNAs. Fig 11 illustrates the bioinformatics pipeline used in prediction of lncRNAs in all four datasets, and Table 7 and S1 table report the number of putative lncRNAs predicted in each dataset. FASTA and gtf annotation files are available at http://www.animalgenome.org/repository/pub/MTSU2015.1014/.

Table 7: Number of lncRNA predicted in at least 2 of the 4 datasets and final numbers.

| Source | LncRNAs common between two data sources | | | | Putative non-redundant lncRNA from each sources after combining all four | |
|---|---|---|---|---|---|---|
| | Salem et. al. | Berthelot et. al. | Narum et. al. | Fox et. al. | Source | Number |
| Salem et. al. | x | 35,307 | 13,557 | 268 | Salem et. al. | 21,617 |
| Berthelot et. al. | 35,307 | x | 13,993 | 291 | Berthelot et al. | 22,568 |
| Narum et. al | 13,557 | 13,993 | x | 401 | Narum et. al | 10,097 |
| Fox et. al. | 268 | 291 | 401 | x | Fox et al. | 221 |
| | | | | | total | 54,503 |

To look for evolutionarily conserved lncRNAs in rainbow trout, all putative lncRNA transcripts (54,503) were searched against several noncoding-RNA databases ($E \leq 0.0001$) [175, 182-186]. Of those 54,503 lncRNAs, only 421 had sequence homology to lncRNAs from other species (S1 table). This low evolutionary conservation of lncRNAs is in agreement with previous reports [43].

### Characterization of lncRNAs

Studies on mouse, zebra fish and maize have suggested that lncRNAs are shorter than protein-coding genes, have relatively fewer exons, and are expressed at a lower level [38, 175, 191]. Consistent with previous reports, our study indicates that trout lncRNAs were shorter (0.821 kb) than protein-coding genes (1.636 kb) (Fig 12). In addition, the average number of exons in lncRNAs was 1.14 compared to 4.75 in protein-coding genes. Unlike the trout protein-coding genes, ~90% of the trout lncRNAs had one exon. Fig 12 and Table

8 show distribution and number of exons in lncRNAs compared to protein-coding genes. Data regarding exon numbers in lncRNAs from different species are inconsistent. Similar to our findings, some plant and animal studies reported one-exon bias for lncRNAs [38, 192]. Conversely, some human studies showed a remarkable two-exon prevalence in the majority of lncRNAs [43]. Several reasons may explain these discrepancies including tissue variation, developmental stages, sequencing techniques and biases due to variations in number and length of genes in different species.

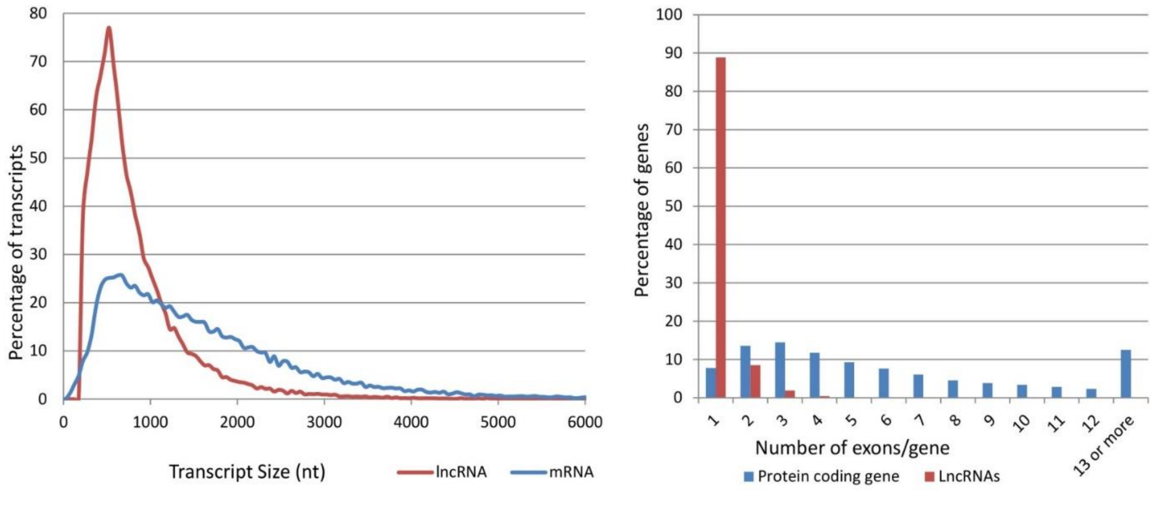
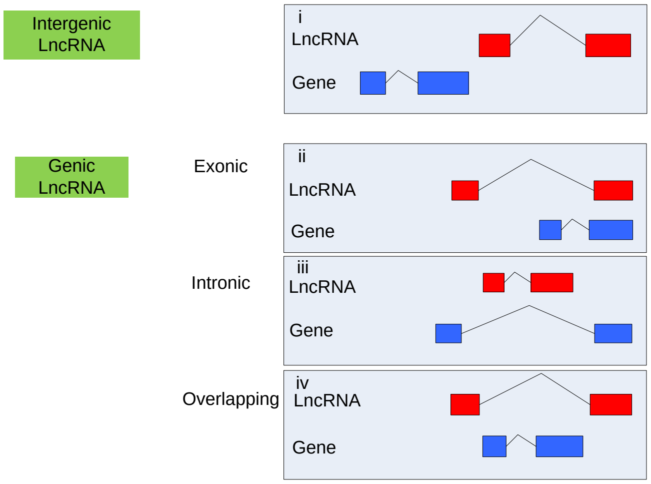


Figure 12: Distribution of sequence length of LncRNAs compared to protein-coding transcripts in rainbow trout. LncRNAs were shorter than protein-coding genes with (0.821 kb) and (1.636 kb) average length in each, respectively (Left). Distribution of number of exons (Right).

Table 8: Number of exons and average length of lncRNAs in different data sets

| | Salem et al. | | Berthelot et al. | | Narum et al. | | Fox et al. | | Common | |
|---|---|---|---|---|---|---|---|---|---|---|
| **# of exon** | LncRNA % | Average length | LncRNA % | Average length | LncRNA % | Average length | LncRNA % | Average length | LncRNA % | Average length |
| **1** | 86.14 | 790 | 88.52 | 682 | 96.62 | 453 | 98.24 | 353 | 88.84 | 796 |
| **2** | 10.63 | 888 | 8.71 | 846 | 2.79 | 462 | 1.34 | 377 | 8.49 | 1007 |
| **3** | 2.37 | 973 | 2.07 | 893 | 0.43 | 480 | 0.42 | 359 | 1.91 | 1044 |
| **4** | 0.51 | 1090 | 0.47 | 1030 | 0.10 | 475 | 0.00 | 0.0 | 0.46 | 1225 |
| **5** | 0.15 | 1284 | 0.11 | 1217 | 0.02 | 792 | 0.00 | 0.0 | 0.13 | 1390 |
| **6** | 0.08 | 1289 | 0.04 | 1157 | 0.02 | 514 | 0.00 | 0.0 | 0.07 | 1206 |
| **7** | 0.05 | 1379 | 0.03 | 1076 | 0.01 | 477 | 0.00 | 0.0 | 0.03 | 1183 |
| **8** | 0.03 | 1322 | 0.01 | 1227 | 0.00 | 631 | 0.00 | 0.0 | 0.02 | 1364 |
| **9** | 0.01 | 1217 | 0.01 | 1394 | 0.01 | 620 | 0.00 | 0.0 | 0.01 | 1302 |
| **10** | 0.02 | 1167 | 0.01 | 1199 | 0.00 | 0.0 | 0.00 | 0.0 | 0.01 | 1181 |

LncRNAs are classified, based on their intersection with protein-coding genes, as genic and intergenic [43]. Some of the lncRNAs are located in transcriptionally-active regions and influence expression of neighboring genes [145, 193]. Therefore, the genomic position of lncRNAs relative to protein-coding genes can possibly provide important clues about lncRNA-mediated regulation of protein-coding genes [194]. Our data indicate that 7,847 (14.4%) of the lncRNAs intersected with protein-coding gene and thus are called genic (Fig 13). Of these lncRNAs 4,697 (8.6%), were intronic lncRNAs, existing in introns of protein-coding genes but do not intersect with any exons, and 3,091 (5.6%) exonic, sharing at least part of a protein-coding exon. Among those lncRNAs, 248 were sense and 1,488 were antisense; and 6,052 lncRNAs had an unknown orientation. In addition, there were 59 lncRNAs that completely overlapped with a protein-coding gene by containing this protein-coding gene within its intron.  Fig 13 and S1 table show classification and number of lncRNAs based on their intersection with protein-coding genes. There were 46,656 (85.6%) intergenic lncRNAs in the trout genome that did not intersect but were within 15 kb of the nearest protein-coding gene.   Those intergenic lncRNAs were further divided into 3,588 convergent (same sense) and 3,428 divergent (opposite sense). Consistent with our study, previous reports in humans indicate that the majority of lncRNA transcripts do not intersect with protein-coding genes [43].



| Interginc | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Salem et. al.(39,613) | | | Berthelot et. al.(44,312) | | | Narum et. al.(17,071) | | | Fox et. al.(535) | | | Common(46,656) | | |
| 1,001C | | 908D | 1,003C | | 881D | 8,558C | | 8,513D | 83C | | 69D | 3,588C | | 3,428D |
| 33,733U | | | 42,428U | | | 0U | | | 383U | | | 39,640U | | |
| Genic | | | | | | | | | | | | | | |
| Salem et. al.(4,511) | | | Berthelot et. al.(7,570) | | | Narum et. al.(10,505) | | | Fox et. al.(656) | | | Common(7,847) | | |
| Exonic | Intronic | Overlapping | Exonic | Intronic | Overlapping | Exonic | Intronic | Overlapping | Exonic | Intronic | Overlapping | Exonic | Intronic | Overlapping |
| 1,915 | 2,555 | 41 | 3,344 | 4,148 | 78 | 3,043 | 7,459 | 3 | 156 | 500 | 0 | 3,091 | 4,697 | 59 |
| S30<br>A146<br>U4,294 | | S18<br>AS23<br>U0 | S35<br>AS225<br>U7,232 | | S0<br>AS6<br>U72 | S1,300<br>AS9,202<br>U0 | | S2<br>AS1<br>U0 | S14<br>AS167<br>U475 | | | S248<br>AS1,488<br>U6,052 | | S2<br>AS5<br>U52 |

Figure 13: Classification of lncRNAs based on their intersection with protein-coding genes and number of lncRNAs in each class. Diagram on the top is a visual illustration of each class of lncRNAs relative to nearest protein-coding gene(s) based on genomic position and direction of transcripts. Bottom figure in tabular format presents number of different classes of lncRNAs from each class. Numbers inside brackets following data source references indicate total number of that particular class of lncRNAs. Letters C, D, S, AS and U indicate number of convergent, divergent, sense, anti-sense and transcripts with unknown directionality, respectively.

### Expression of lncRNA in different tissues

A comparison of lncRNA expression to protein-coding genes showed that transcript abundance of lncRNAs is lower than that of protein coding genes. Average RPKM (Reads Per Million per Kilo-base) of the most abundant 40,000 transcripts was 3.49 and 15.69 in LncRNAs and protein-coding genes, respectively (Fig 14). Similar trends, showing lower lncRNAs expression in all human tissues compared to mRNAs, were reported [43].

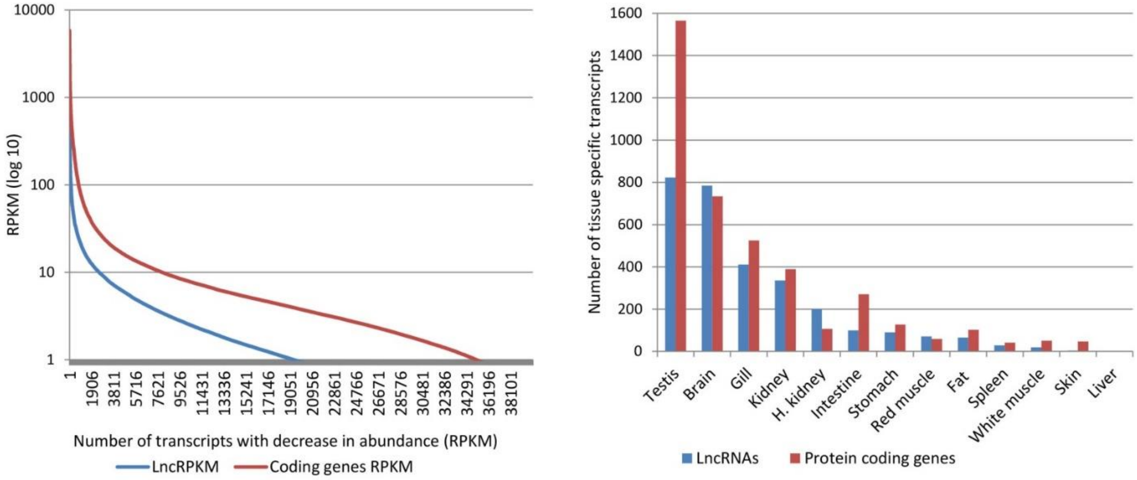


Figure 14: RPKM comparison of protein-coding genes and lncRNAs. Transcript abundance of lncRNAs is lower than that of protein-coding genes. Average RPKM of the most abundant 40,000 genes is 15.69 and 3.49 for protein coding genes and LncRNAs, respectively (Left). Number of tissue-specific lncRNAs and protein-coding genes in various tissues. Expression of lncRNAs and protein-coding genes showed similar patterns among different tissues (Right).

Evidence is clear that lncRNAs exhibit strict cell/tissue specificity and play a significant role in development and differentiation of tissues in plants and animals [38, 151]. Nonetheless, their tissue specificity and potential role in tissue development are not well studied in fish. Lack of sequence conservation of lncRNAs across diverse species demands study of their expression in vital tissues as a method to identify lncRNAs with tissue-specific roles in rainbow trout. In this study, lncRNA expression was studied in 13 vital tissues of rainbow trout. Out of 54,503 putative lncRNAs, 3,269 (~5.9%) exhibited expression across all tissues with a minimum RPKM value of 1.0 (S2 table). On the other hand, 2,935 tissue-specific lncRNAs (5.4%) were identified from 13 tissues (S3 table). In this report, transcripts were described as 'tissue specific' if their expression in one tissue was 8-fold or higher compared to the maximum value for any of the other 12 tissues with a minimum RPKM of 0.5 [51] (Fig 14). Previously, we reported 17.1% and 8.9%, respectively, for housekeeping and tissue-specific protein-coding genes [51]. To gain insight into the expression and tissue specific differences between lncRNAs and protein-coding genes, the number of each was examined in 13 different tissues (Fig 14). Testis expressed the highest number of tissue-specific lncRNAs followed by brain, gill, and kidney. Conversely, liver expressed the lowest number of tissue-specific lncRNAs followed by skin, white muscle then spleen, in increasing order. We previously reported that the number of tissue-specific protein-coding transcripts follows similar patterns in various tissues [51]. Similar to the protein-coding genes, expression patterns of tissue-specific lncRNAs can be explained in terms of tissue complexity [51].

Previously, we showed that tissues are different in terms of the protein-coding transcriptome composition and complexity. Brain and testis possess the most complex transcriptomes. These tissues express large numbers of the genes; however, only a small part of the mRNA pool is expressed by the most abundant genes [51]. On the other hand, white muscle and stomach revealed simpler transcriptomes. These tissues express fewer genes and a greater proportion of the transcriptome comes from the most highly expressed genes. Similarly and in this study, complex tissues like brain and testis, expressed a larger number of lncRNAs with equal dominance of many transcripts (Fig 15). Conversely, white muscle, fat and liver showed less complex transcriptomes; a vast majority of the transcriptome included a few dominant lncRNAs. Similar expression patterns between protein-coding genes and lncRNAs may suggest common mechanisms of gene expression regulation and important role of lncRNAs in regulating protein-coding RNAs. Regardless, these data suggest that lncRNAs may be significant in determining tissue complexity.

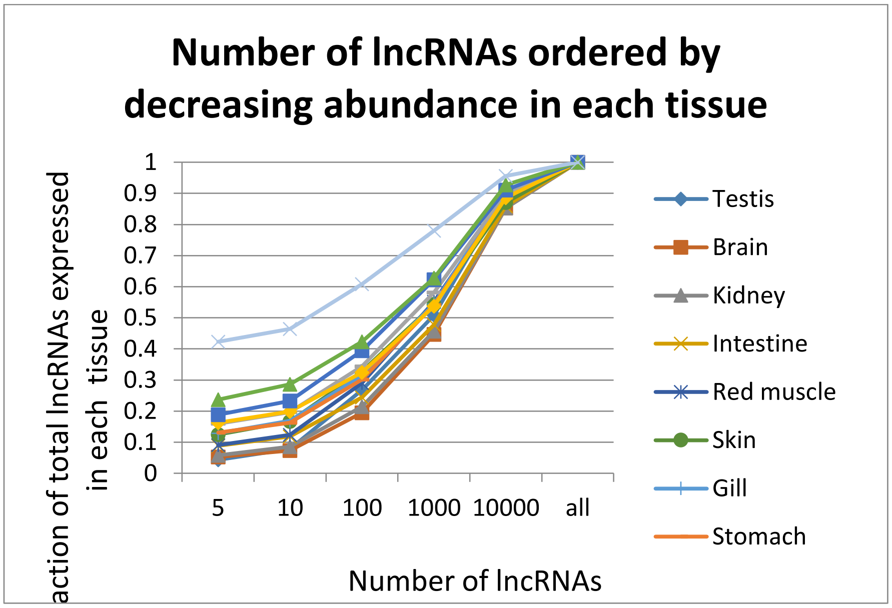


Figure 15: Distribution of lncRNA expression in various tissues. Proportion of the transcriptome that is contributed by the most abundant lncRNAs is plotted in various tissues. In complex tissues like brain and testis, larger number of lncRNAs were expressed with fairly equal dominance of many transcripts. On the contrary, less complex tissues like white muscle, fat and liver showed that majority of transcriptome is contributed by few dominant lncRNAs.

### Correlation in expression patterns of lncRNA and protein-coding genes across tissues

Very low sequence conservation of lncRNAs hinders their molecular annotation. In order to look for possible functional significance of lncRNAs in regulating protein-coding genes, we constructed an expression-based relevance network between protein-coding genes and lncRNAs using a clustering algorithm in Multi-experiment Viewer software package (MeV) [188, 189]. In this study, biological correlation in expression patterns were compared across 13 tissues representing vastly different cellular and functional complexities. After clustering, genes of each cluster were ranked based on their entropies, and the top 20% of genes with the highest entropy were retained to construct networks. This approach identified 15 clusters containing protein-coding and lncRNA genes with strong correlation in their expression patterns ($R^2$ >0.97) (S4 table). Examples of functionally important clusters include lncRNA Omy100084431 that was highly, positively correlated with splicing factor 3B (GSONMT00018324001) and transcription elongation factor SPT5 isoform X1 (GSONMT00067984001). In addition, expression of lncRNAs Omy200064145 and Omy100138726 was positively correlated with NF-kappa B inhibitor-like protein (GSONMT00082784001). Furthermore, a strong positive correlations in expression pattern between lncRNAs Omy300110093 and mitogen activated protein kinase1-like (GSONMT00053903001); Omy300072481 and thyroid hormone receptor alpha-like (GSONMT00066016001); Omy200106644 and histone deacetylase 3-like (GSONMT00058062001); and Omy300066671 and double-stranded RNA-specific adenosine deaminase (GSONMT00000999001) were observed. Proteins listed in these clusters have important functional roles in the cell including protein quality control (derlin-2) [195], RNA editing (adenosine deaminase) [196], transcriptional control (histone deacetylase 3) [197], splicing, and development. These findings nicely correlate with previously characterized molecular functions of lncRNAs in different species [157, 165, 166]. In order to explore additional underlying biological relationships between lncRNAs and protein-coding genes, more samples from different individuals and developmental stages should be studied as lncRNAs may be specific to developmental stages.

### More stringently selected lncRNAs

The aforementioned 54,503 putative lncRNAs were identified using filtration steps with traditional cutoff values [175, 191]. To provide an optional more stringently selected list of lncRNAs, we performed extra filtration as follows. First, we calculated the average amino acid length for the shortest 10% of the rainbow trout protein-coding genes [25]; this calculation yielded 83 amino acids. Using 83 amino acids as the cut-off value of the lncRNA, 5,836 lncRNAs were filtered out of 54,503. In addition, lncRNA containing ORF covering more 35% of its length were filtered out [187]. Second, we decreased the cut-off value for the CPC [177] from -0.5 to -1.0, which filtered out an extra 4,978 leaving 43,689 putative lncRNA. The next filtration step was performed based on location of the lncRNAs in the genome predicted from a comparison of different datasets. If any lncRNA overlapped fully or partially by more than 100 nt with another lncRNA from a different dataset, we filtered out the shortest lncRNA; this step eliminated 5,945 putative lncRNAs. In addition, we filtered out any lncRNAs that overlapped with a protein-coding gene in the sense orientation and this filtration eliminated an additional 354 lncRNAs. The last filtration step removed any single-exonic lncRNA that was within 500 nt of a protein-coding gene; as a result, 1,538 putative lncRNAs were removed. The final number of putative lncRNAs was 31,195 (S1 table). FASTA and gtf annotation files are available at http://www.animalgenome.org/repository/pub/MTSU2015.1014/. Because the criteria for distinguishing lncRNAs are still loosely defined [198], filters applied in this study (with traditional or stringent cutoff values) should be considered arbitrary, hence, the identified lncRNAs may or may not reflect biological functions. For example, some of the well characterized lncRNAs in mammals contain more than 100 AA ORF. In this study, two sets of lncRNAs were obtained with traditional or stringent cut off values. All above mentioned analyses were done using lncRNAs from the traditional filtrations.

# CHAPTER IV: IDENTIFICATION OF SNPS ASSOCIATED WITH MUSCLE YIELD AND QUALITY TRAITS USING ALLELIC-IMBALANCE ANALYSIS IN POOLED RNA-SEQ SAMPLES IN RAINBOW TROUT [199]

## Introduction

Fish growth rate, muscle yield and fillet quality are major traits affecting profitability of aquatic food animal production. As feed cost is a major factor influencing the profitability, efficiency of growth is important and related to growth rate and muscle yield and composition. Skeletal muscle constitutes about 50-60% of the fish weight [200]. Given that growth efficiency and fillet firmness and appearance are critical for profitability and production of premium products [2], optimizing fish growth, muscle yield and fillet quality traits is a key objective in aquaculture breeding programs. Traditional phenotype-based selection is typically used to select for fast growth; however, muscle yield and quality traits are difficult to improve by conventional selection because measurement of these traits requires sacrificing the animal [201].

Genomic selection tools have been created to improve economically important traits in plants and livestock. Genetic maps, which characterize the linkage or co-inheritance patterns of genetic markers, have been developed for a wide range of species, including fish, with the aim of discovering allelic variation affecting traits; and ultimately identify DNA sequences underlying phenotypes [202, 203]. Markers have been identified by various molecular techniques, including numerous and genome-wide single nucleotide polymorphisms (SNPs). In addition, recent technological developments have enabled high throughput genotyping of these SNPs rendering them useful for genome-wide association studies [47, 204-206]. Functional SNPs are generally defined as SNPs from genome sequences with a functional effect. These sequences include coding SNPs (e.g. non-synonymous, splicing), promoter and noncoding SNPs, as well as functional elements identified from studying of genome conservation [207]. Functional/coding SNPs are especially important because they have the potential to change the function of a protein [50, 203, 208]. In addition, functional/coding SNP markers are unlikely to become unlinked from their associated genes due to genetic recombination. Therefore, functional/coding SNPs can be useful genetic markers for detecting significant associations with phenotypes. Understanding molecular mechanisms of muscle growth and quality can help in making better selection decisions. In terrestrial livestock, several genes, genetic markers and QTLs associated with production traits, including growth, have been characterized using molecular techniques [209, 210]. In addition, marker-assisted selection has been used to enhance genetic improvement in livestock breeding programs by direct selection on genes affecting economic traits [211] and to optimize selection for quantitative traits [209, 210]. However, the genetic basis of muscle growth and quality traits is not well studied in fish [212].

Rainbow trout is the most cultivated cool and cold freshwater fish in the U.S. [213], and it is considered a model species for studies in several fields of biology, including ecology [19], pathology [20], physiology [214], toxicology [22] and carcinogenesis [23]. Several studies used RNA sequencing to identify markers in human [215, 216] and non-model species [50, 217, 218]. However, most SNP detection algorithms were developed for DNA-

Seq analyses and are not optimized/tested for RNA-Seq, especially in pooled samples. The objective of this study was using RNA-Seq analyses of pooled samples to identify functional/coding SNP markers and develop a resource for studies of marker association with production traits in rainbow trout. First, transcriptome-wide SNP allele frequencies were correlated to phenotypic variations in fish whole body weight (WBW) and muscle yield, fat content, shear force and whiteness. Second, SNPs with allelic imbalance scores (ratios between the allelic frequencies of the high-end families and that of the low-end families) were identified. Then, a subset of the putative SNPs was validated for allelic polymorphism and tested for trait association. Finally, genes harboring SNPs with allelic imbalances were annotated to obtain insight into the potential functional effects of the SNPs.

## Methods

### Ethical statement

Institutional Animal Care and Use Committee of the United States Department of Agriculture, National Center for Cool and Cold Water Aquaculture (Leetown, WV) specifically reviewed and approved all husbandry practices used in this study (IACUC approval #056).

### Fish population and sequencing

Phenotypic data and muscle samples were collected from ~500 fish representing 98 families (5 fish/family) from the growth-selected line at NCCCWA (year class 2010) as previously described [50]. Families were produced and reared until ~13 months post-hatch as described in reference [219]. At about ~13-months old and in each of five consecutive weeks, approximately 100 fish (i.e., 1 fish per full-sib family per week) were anesthetized in approximately 100 mg/L of tricaine methane sulfonate (Tricaine-S, Western Chemical, Ferndale, WA) weighed, slaughtered, and eviscerated. A muscle sample was excised from the left dorsal musculature and frozen in liquid nitrogen. Head-on gutted carcasses were packed in ice, transported to the West Virginia University Meats Processing Laboratory (Morgantown, WV), and stored overnight. The next day, carcasses were hand-processed into trimmed, skinless fillets by a trained faculty member and weighed. Fresh fillet surface color was measured with a Chroma meter (Minolta, Model CR-300; Minolta Camera Co., Osaka, Japan) calibrated using a standard white plate No. 21333180 (CIE Y 93.1; x 0.3161; y 0.3326). L* (lightness), a* (redness), and b* (yellowness) values were recorded at three locations above the lateral line along the long axis of the right fillet, and these values were used to calculate a fillet whiteness index according to the following equation: Whiteness = $100 – [(100 – L)^2 + a^2 + b^2 ]^{1/2}$ [81]. The left-side fillet was frozen for subsequent proximate analysis, and a 4 × 8 cm fillet section was cut from the left side for subsequent cooked texture analysis [220].

For RNA-Seq study, eight different families (5 fish each) showing opposite phenotypes for each of the 5 traits were analyzed (4 high ranked families versus 4 low ranked families on average for each trait). Each family represented a full-sib family from the above-described growth-selected line. Muscle tissues were collected from each fish and flash frozen in liquid nitrogen then stored at −80°C until RNA isolation. Total RNA was isolated from each sample using TRIzol™ (Invitrogen, Carlsbad, CA). Equal masses of total RNA from 5 samples of each family were pooled and used for RNA-Seq sequencing. cDNA libraries were prepared and sequenced on an Illumina HiSeq2000 (single-end, 100bp read length) using multiplexing standard protocols.

### SNP detections using

#### SAMtools/Popoolation2

For each trait (WBW, muscle yield, muscle fat content, shear force, and whiteness), sequence reads from each family were aligned to the rainbow trout genome using STAR [221]. After read alignment, the SAMtools view/sort and mpileup functions were used within the Popoolation2 package (version 1.201) to determine the genotype for each variant and calculate allele frequencies [222, 223]. Initial SNPs were considered at minimum reads > 10 and minor allele count > 4 and MAF > 0.05. Putative SNPs associated with each trait were determined using an in-house Perl script at allelic imbalance scores (the ratio between the allelic frequencies of the high-end families and that of the low-end families) >2.0 as an amplification and <0.5 as loss of heterozygosity.

#### SNP detection using GATK tools

For the GATK pipeline [224], reads from each sample were aligned to the rainbow trout genome using STAR [221] as recommended by the GATK practice. Picard tools were used to sort the SAM files and to mark duplicates, a step used by GATK to reduce a false positive due to error in duplicate that could be falsely detected as a SNP. The following steps were performed according to GATK pipeline for RNA-Seq (Split and trim to reassign mapping quality, Indel realignment, local realignment around Indel in order to clean up any mapping artifacts and Base Quality Score Recalibration). After data preparation, variants were called using HaplotypeCaller followed by hard-filtering using the following parameters: Qual By Depth (QD) 2.0, FisherStrand (FS) 60.0: RMS Mapping Quality (MQ) 40.0, MAF > 0.05. Since GATK was not optimized to calculate allelic imbalances in RNA-Seq data, putative SNPs identified in each family were analyzed using an in-house Perl script to determine the allelic imbalances applying the criteria that we used in the SAMtools/Popoolation2 method.

### SNP validation

Flanking sequences (up to 250 bp on each side) of putative SNPs were extracted from the reference genome [25]. Some SNPs were removed from SNP assay design because either a sequence gap was located less than 60 bp from the SNP site or a non-target SNP was located less than 30 bp away from the target SNP. A total of 92 SNP assays were developed and evaluated with 282 DNA or cDNA samples. These included 85 DNA samples derived from 19 full-sib families used for RNA-Seq and their parents (38 DNA samples), DNA samples of 2 full-sib mapping families (2 parents and 19 offspring per family), 64 DNA samples from two commercial populations (Troutlodge Inc. and Clear Springs Foods Inc.) and 35 cDNA samples derived from the RNA samples used for RNA-Seq high versus low muscle yield. The SNP genotyping was performed following the instructions of the Fluidigm genotyping user guide. Briefly, DNA and cDNA samples were pre-amplified, diluted and used for genotyping with 96.96 Dynamic Array IFCs (Integrated Fluidic Circuits). The arrays were read using EP1 system, and genotypes were called automatically using Fluidigm SNP genotyping analysis software 4.1 with a confidence threshold of 85. The genotype clusters were examined for each assay and any wrong calls or no calls were corrected manually. The program Pedcheck [225] was used to identify genotypes inconsistent with Mendelian inheritance between parents and offspring. Chi-square goodness of fit tests were performed to identify SNPs with significant segregation distortion ($P < 0.01$) in the two mapping families. Those SNPs were reported as assay-failed SNPs.

For the Sanger sequencing validation of the SNPs showing potential mon-allelic gene expression, flanking sequences (up to 250 bp on each side) of each SNP were PCR amplified from DNAs and cDNA from the same 35 fish samples that were used for RNA-Seq high versus low muscle analyses. PCR amplicons were Sanger sequenced and manually inspected for consistency between DNA and cDNA genotypes or mono-allele specific gene expression as explained in the results section.

### Functional annotation of SNPs

SNP annotation by functional class (genic/intergenic etc.) for different SNP sets and their genome distributions were conducted using in-house Perl scripts. The gff file of the rainbow trout genome reference was used to determine if a SNP is located within an mRNA start and end positions (genic), within a CDS, 5’UTR or 3’UTR. Upstream/ downstream SNPs were determined if located within 5kb of a protein-coding gene. SNPs were called intergenic if located more than 5Kb of protein-coding genes. SNPs within lncRNAs were determined using gtf file of our previously reported lncRNA reference [115]. Functional annotation of the SNP-harboring genes was performed using the Blast2GO suite [30] and KEGG pathway mapping [226].

## Results and Discussion

### Phenotypes

SNPs were identified in fish families with divergent phenotypes in WBW, muscle yield, fat content, shear force (texture) and whiteness of the fillet. These rainbow trout were from a growth-selected line developed by the NCCCWA breeding program [219]. Briefly, this line was created through artificial selection, starting in 2004, from 7 founder strains with documented diversity and domestication history. Over five generations, the population responded to selection by 9.8-12.7% increase in WBW per generation, and rate of inbreeding averaged 0.86% per generation [219]. In this study population, which was sampled after three generations of selection (hatch year of 2010), WBW was positively correlated with muscle yield and muscle fat content ($R^2$= 0.56 and 0.50 respectively, data not shown). Our previous reports showed that fast growth may be genetically associated with improved muscle yield, paler fillets (affected by intramuscular fat content) and firmer texture [227]. The trait heritability estimates for muscle yield, muscle weight, WBW10, WBW13, carcass weight, fat percentage, shear force and fillet color were moderate to high (0.31–0.81) [205, 227]. Those moderate to high heritability estimates imply that substantial additive genetic variation exist in the study population for growth and carcass traits.

For RNA sequencing, muscle samples were collected from 7-9 different full-sib families showing divergent phenotypes per trait (i.e. 3-5 high ranked families versus 3-5 low ranked families per trait). Five fish were sampled from each family. Divergent phenotypic attributes (Figure 16) were different ($P$<0.01): WBW (1221.6g ± 84.25 vs. 502.1±28.0g), muscle yield (50.9% ± 1.6 vs. 43.3% ± 2.3), muscle crude-fat (9.24% ± 1.2 vs. 4.77% ± 1.3), shear force (grams force/grams of sample; 539.64± 12.3 vs. 310.01± 49.2), and fillet whiteness index (44.7 ± 0.8 vs. 41.23 ± 0.4) for high- vs. low-ranking groups, respectively. Means and standard deviations of these traits were calculated from the family averages.

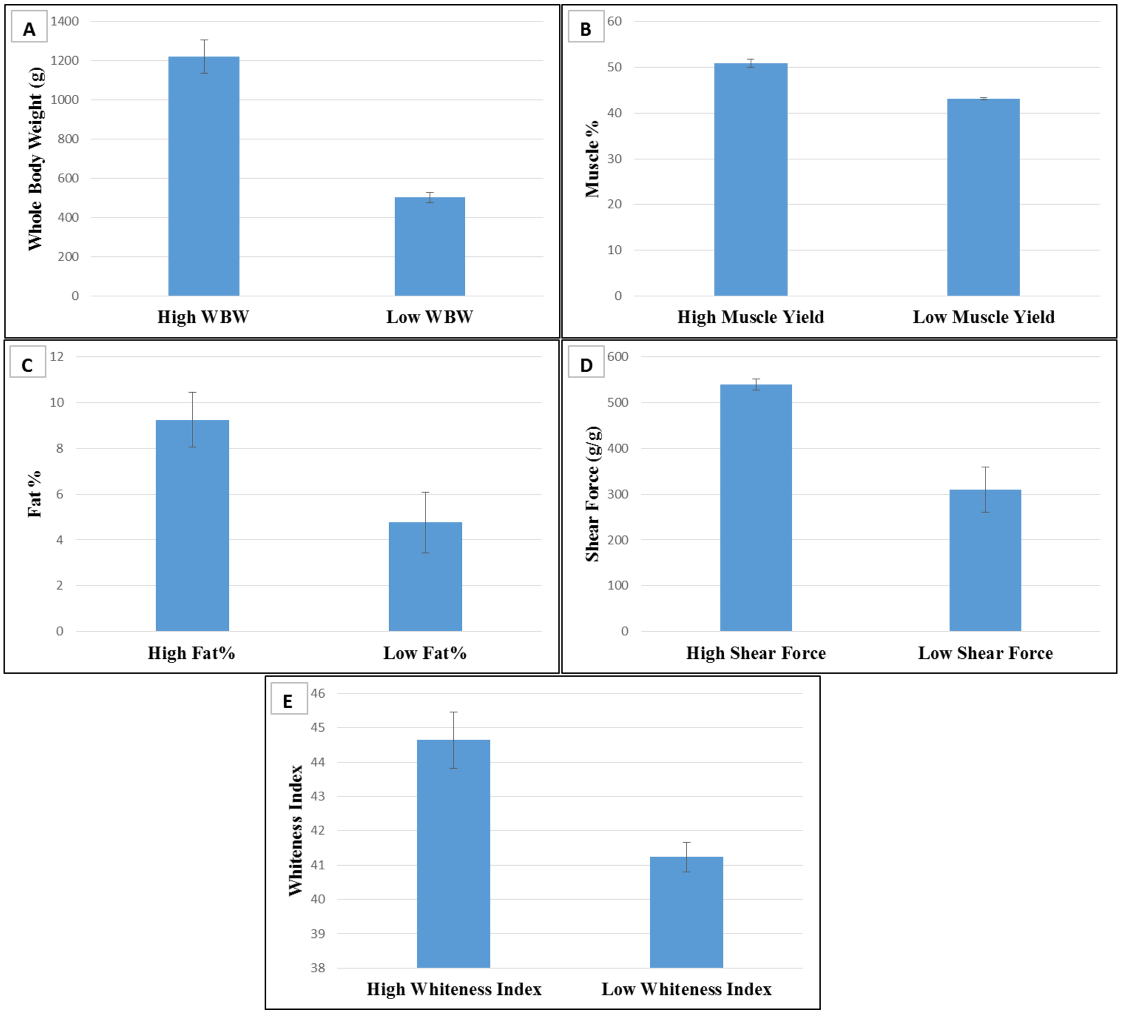


Figure 16: Phenotypic variations in fish families with contrasting phenotypes for five different traits; whole-body weight, muscle yield, fat content, shear force and fillet whiteness index. All differences were statistically significant (p< 0.01).

### Identification of putative SNPs

RNA pools from muscle tissues of 5 fish per family were used for RNA-Seq analyses. A total of 259,634,620 reads (100 bp single-end) were generated from 22 families at an average of 11,801,573 reads per family. Reads were aligned against the rainbow trout genome [25] using the STAR [221] alignment tool. Percentage of reads mapped to the genome ranged from 80% to 82% per family.

A total of 204,604 putative SNPs were detected for the five traits using Haplotypecaller tool of Genome Analysis Toolkit v3.3.0 (GATK) [228], with an average of 40,920 SNPs per trait. Using the SAMtools/Popoolation software package [229, 230], a total of 304,805 putative SNPs were predicted, with an average of 60,961 SNPs per trait (Table 9). After removing redundant SNPs among all traits, we had 59,112 SNPs from GATK and 87,066 from SAMtools/Popoolation2 with 50,885 shared between the two bioinformatics pipelines (Table 9).

After identifying putative SNPs, an in-house Perl script was used to estimate allelic imbalances of the SNPs in each trait. A total of 6,275 SNPs with allelic imbalances were identified from the GATK dataset at cutoff values of >2.0 as an amplification and <0.5 as loss of heterozygosity. In addition, 969 SNPs explicitly existed in only the high or low phenotypic group. After removing redundant SNPs between traits at the two cutoff values, there were 4,798 unique SNPs (Table 9). Similarly, SAMtools/Popoolation2 identified 5,070 SNPs with allelic imbalances at cutoff values of >2.0 as an amplification and <0.5 as loss of heterozygosity. In addition, 1,450 SNPs existed in families at one of the two ends of each trait variation scale but not in the other (Table 9). There were 4,962 non-redundant

SNPs among the five traits that were identified with SAMtools/Popoolation2 at the two cutoff values. There were only 1,829 non-redundant SNPs shared between GATK and SAMtools/Popoolation2. Differences in variant calling and filtering steps might have caused the observed differences in number of SNPs between GATK and SAMtools/Popoolation2.

For subsequent analyses, we combined SNPs from GATK and SAMtools/Popoolation2 into three different groups: 1) Non-redundant SNPs with allelic imbalances from both methods (7,930 SNPs); 2) Common putative SNPs from both methods (50,885 SNPs); 3) Putative non-redundant SNPs from both methods (95,234 SNPs) (Table 9). All SNPs data are provided in Additional file 1.

Table 9: Summary of putative SNPs and SNPs showing allelic imbalances identified by SAMtools and GATK for each trait. Allelic imbalances were calculated at >2 for amplification and <0.5 for loss of heterozygosity. SNPs explicitly existing in only the high or low phenotypic group are indicated in the table by the 0.0/1.0 ratio. * 59 SNPs were multi-allelic, showing different alleles in association with different phenotypes. ** 1 SNP was multi-allelic showing different alleles predicted by different pipelines.

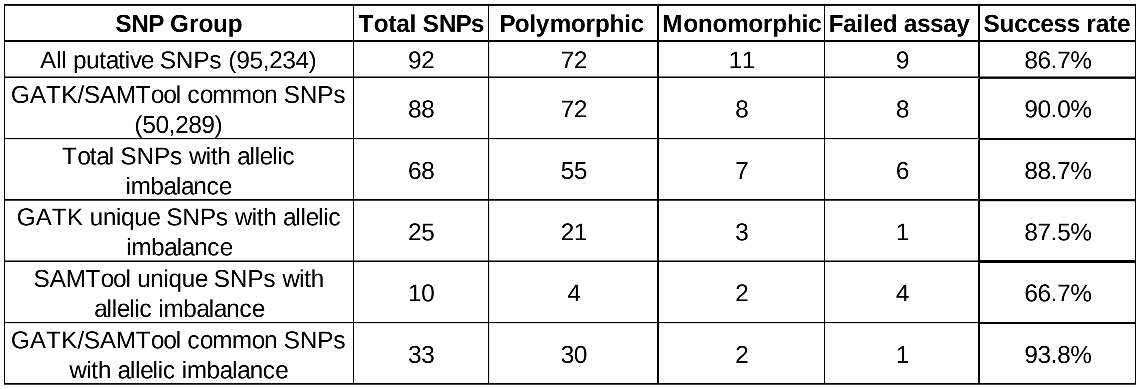
| Trait | No. of putative SNPs | | No. of SNPs with Allelic imbalance | | | |
|---|---|---|---|---|---|---|
| | SAMtools/ Popoolation2 | GATK | SAMtools/ Popoolation2 | | GATK | |
| | | | 0.5/2.0 | 0.0/1.0 | 0.5/2.0 | 0.0/1.0 |
| **Fat%** | 59,032 | 38,808 | 662 | 406 | 877 | 270 |
| **Shear** | 60,309 | 38,960 | 910 | 488 | 1,152 | 261 |
| **Muscle%** | 61,117 | 42,383 | 1,321 | 116 | 1,507 | 76 |
| **Whiteness** | 64,636 | 44,460 | 1,011 | 347 | 1,283 | 298 |
| **WBW** | 59,711 | 39,993 | 1,166 | 93 | 1,456 | 64 |
| **Total # SNPs** | 304,805 | 204,604 | 5,070 | 1,450 | 6,275 | 969 |
| **Total # SNPs non- redundant** | 87,066 | 59,112 | 4,962 | | 4,798 | |
| **Total Common SNPs** | 50,885 | | 1,829 | | | |
| | **All putative SNPs(MAF>0.05)** | | **Total No. of SNPs with allelic imbalance = 7,930 **** | | | |

### SNP validation

A total of 92 putative SNPs including 88 SNPs from the GATK/SAMtools common pool (50,885 SNPs) were selected for SNP validation. Among the 92 putative SNPs, 68 SNPs showed allelic imbalances (Table 10), including 25 SNPs identified by GATK pipeline, 10 SNPs identified by SAMtools pipeline, and 33 SNPs identified by both pipelines (Table10). Among the 92 tested SNPs, 72 (78.2%) SNPs were polymorphic, 11 (11.9%) SNPs were monomorphic and 9 failed the assay (Table10). Failure of the Fluidigm assay can be caused by unsuccessful or non-specific primer binding to the target genomic DNA. Therefore, we cannot assume that a failed assay indicates failure of our bioinformatics pipeline to detect a SNP in the RNA sequence data, and can remove the failed SNP assays from the calculation of SNP validation rate. As 72 out of the 83 working Fluidigm SNP assays were polymorphic we can claim 86.7% validation rate in detecting polymorphic SNPs in the overall putative SNP pool and 90% validation rate in the GATK/SAMtools shared SNPs pool. This success rate is much higher than what we previously achieved in rainbow trout using RNA-Seq (70%) and genomic reduced representation libraries (48%) [48, 50]. The improved success rate in this study is perhaps due to use of a reference genome instead of *de novo* assembled references used in the previous studies. In addition, a transcriptome sequence coverage of ~7.4X per fish was used compared to only ~0.97X in our previous RNA-Seq study [50]. The 90% successful SNP validation rate is comparable to that reported in diploid fish or using genomic RADs and doubled haploid fish in rainbow trout [47, 231]. In addition, a recent rainbow trout genome re-sequencing study with at least 10x genome coverage per fish had 86% successful validation rate [47]. Relatively lower success rates in SNP detection were reported from RNA-Seq studies in rainbow trout due to genome duplication and assembly errors in the genome/transcriptome references [50, 232, 233]. Noteworthy and in a separate study, we found variation in gene expression in only 75 genes distributed between all 5 traits (data will be published elsewhere). Therefore, differential gene expression effects on estimating allelic imbalances were negligible as only 75 genes distributed between all five traits were differentially expressed between the high and low families. Minor effects of variation in gene expression on allele frequency estimation accuracy were previously reported [234]. The SNP validation data, albeit small, indicated that the GATK method was more successful in calling polymorphic SNPs with allelic imbalances than the SAMtools pipeline; 87.5% versus 66.7%, respectively. However, combined GATK and SAMtools data had a 93.8% success rate. Success rates between SNPs with and without allelic imbalances were 88.7% and 86.7%, respectively. Importantly and out of 72 validated SNPs, 61 (84.7%) and 58 SNPs (80.5%) were polymorphic in fish from two different commercially important rainbow trout populations in the US, Troutlodge Inc. and Clear Springs Foods Inc., respectively. These results suggest that the SNPs identified in this study are also useful for other commercial rainbow trout populations.

To evaluate ability of the pipeline in calculating allelic imbalances, DNA and cDNA of the 35 fish used for RNA-Seq analyses of high versus low muscle yield were also genotyped. For all 72 validated SNPs, all DNA and cDNA genotypes were consistent except for 4.64% that indicated mono-allele specific gene expression as explained below.

Table 10: Number of putative and validated SNPs from each dataset.



| SNP Group | Total SNPs | Polymorphic | Monomorphic | Failed assay | Success rate |
|---|---|---|---|---|---|
| All putative SNPs (95,234) | 92 | 72 | 11 | 9 | 86.7% |
| GATK/SAMTool common SNPs (50,289) | 88 | 72 | 8 | 8 | 90.0% |
| Total SNPs with allelic imbalance | 68 | 55 | 7 | 6 | 88.7% |
| GATK unique SNPs with allelic imbalance | 25 | 21 | 3 | 1 | 87.5% |
| SAMTool unique SNPs with allelic imbalance | 10 | 4 | 2 | 4 | 66.7% |
| GATK/SAMTool common SNPs with allelic imbalance | 33 | 30 | 2 | 1 | 93.8% |

## Assessment of Mono-allelic Gene Expression

Out of the 72 validated polymorphic SNPs (Table 10), there were 46 SNPs that showed potential mono-allelic expression in cDNA in at least one fish. In other words, the genomic DNA is heterozygous for the SNP while cDNA is monomorphic. Thirty-three of the 35 fish showed mono-allelic expression in at least one SNP. Out of the aforementioned 46 SNPs, 5 SNPs were randomly selected for validation using Sanger sequencing. All SNPs were heterozygous at the DNA level. However, manual investigation of the cDNA sequence chromatograms exhibited existence of substantial allelic imbalances ranging from existence of two alleles with >2.0 X peak height ratios between the 2 alleles at the SNP base to a complete mono-allelic expression (a single peak). Overall, approximately 4.64% random mono-allelic/allelic imbalances existed in gene expression of rainbow trout. These data are consistent with a recent study in human stem cells showing that most allelic imbalances did not represent ‘on/off’ events, but instead revealed biased expression from each allele [235]. None of the 8 tested families in our study showed mono-allelic expression in all individuals specific to a given family, indicating no parental origin effect through genomic imprinting. Likewise, the human stem cell study suggested that most of the allele-biased gene expression is not due to genomic imprinting [235]. Compared to our estimated 4.64% mono-allelic expression, recent studies showed 12-24% random mono-allelic expression in mammals and 7-9% in interspecies catfish [203, 236-238]. Our mono-allelic expression assessment is based on only 72 SNPs, and hence a genome-wide assessment of mono-allelic expression in rainbow trout warrants further investigation.

### SNP Genomic/Functional Classification

Three sets of SNPs were considered for genomic/functional classifications. For the 7,930 SNPs with allelic imbalances, 2,898 (37.69%) were intergenic. Of them, 635 (8.01%) and 721 (9.09%) SNPs were located within 5Kb upstream or downstream of protein-coding genes, respectively. The rest of the intergenic SNPs, 1,633 (20.59%) were located more than 5Kb distant to protein-coding genes.

On the other hand, 4,941 (62.31%) SNPs were genic, including 214 (2.70%) that were located within the 5’ untranslated region (5’UTR) and 1,677 (21.15%) that were located in the 3’ untranslated region (3’UTR) of protein coding genes. In addition, 2,548 (32.13%) SNPs were located within coding DNA sequences (CDS) and 502 (6.33%) SNPs were located within introns. Of the CDS SNPs, 504 (6.36%) were non-synonymous; 4 of these caused early stop codon, and 500 caused amino acid substitution (Table 11). There were 684 (8.63%) SNPs located within 295 lncRNAs (Table 11).

Regarding the GATK/SAMtools shared SNPs (50,885 SNPs), there were 20,356 (40.00%) intergenic SNPs. Of these shared SNPs, 4,594 (9.03%) were located within 5Kb upstream, and 5,208 (10.23%) downstream of protein-coding genes. In addition, 10,554 (20.74%) were intergenic, more than 5Kb distant to protein-coding genes. In contrast, 30,529 (60.00%) SNPs were genic. And, 1,389 (2.73%) of these SNPs were in the 5'UTR; 10,259 (20.16%) were in the 3'UTR, 15,178 (29.83%) were within CDS; and 3,703 (7.28%) were within introns. Out of those within CDS SNPs, 3,919 (7.70%) were non-synonymous SNPs. Fifty of these CDS SNPs were nonsense (causing premature stop codon), and 3,869 (7.60%) were missense SNPs (Table 11).

Concerning all the putative SNPs, there were 46,901 (49.25%) intergenic SNPs. Of these, 9,005 (9.46%) were located within 5Kb upstream; and 10,245 (10.76%) were downstream of protein-coding genes. In addition, 27,651 (29.03%) were more than 5Kb distant from protein-coding genes. Alternatively, 48,333 (50.75%) SNPs were genic, and of these genic SNPs, 2,247 (2.36%) were in the 5'UTR; 16,420 (17.24%) were in the 3'UTR; 22,616 (23.75%) were within CDS; and 7,050 (7.40%) were within introns. Of the CDS SNPs, 5,853 (6.15%) were non-synonymous with 79 SNPs causing early stop codons and 5,774 (6.06%) causing amino acid changes (Table 11).

In these three SNP datasets, there were large percentages of intergenic and upstream/downstream SNPs (37-49%). Approximately 10% intergenic in addition to 30% non-coding SNPs were reported in humans from RNA-Seq data [239]. Our high percentages of intergenic SNPs may be partially explained by the incomplete annotation of protein coding genes and exons in the current version of the rainbow trout reference genome sequence [25].

Table 11: Summary of SNPs classification for different SNP sets.

| **Functional Class** | **SNPs with allelic imbalance 7.9 K** | **%** | **GATK/SAMtools Common SNPs 50.8 K** | **%** | **All putative SNPs 95.2K** | **%** |
|---|---|---|---|---|---|---|
| **Intergenic** | 2,989 | **37.69%** | 20,356 | **40.00%** | 46,901 | **49.25%** |
| Intergenic(>5K) | 1,633 | 20.59% | 10,554 | 20.74% | 27,651 | 29.03% |
| Upstream (<5K) | 635 | 8.01% | 4,594 | 9.03% | 9,005 | 9.46% |
| Downstream (<5K) | 721 | 9.09% | 5,208 | 10.23% | 10,245 | 10.76% |
| | | | | | | |
| **Genic** | 4,941 | **62.31%** | 30,529 | **60.00%** | 48,333 | **50.75%** |
| 5’UTR | 214 | 2.70% | 1,389 | 2.73% | 2,247 | 2.36% |
| 3’UTR | 1,677 | 21.15% | 10,259 | 20.16% | 16,420 | 17.24% |
| CDS | 2,548 | 32.13% | 15,178 | 29.83% | 22,616 | 23.75% |
| Intronic | 502 | 6.33% | 3,703 | 7.28% | 7,050 | 7.40% |
| | | | | | | |
| **Non-synonymous** | 504 | **6.36%** | 3,919 | **7.70%** | 5,853 | **6.15%** |
| Stop gain | 4 | 0.05% | 50 | 0.10% | 79 | 0.08% |
| Missense | 500 | 6.31% | 3,869 | 7.60% | 5,774 | 6.06% |
| **LncRNA** | 684 | **8.63%** | 4,386 | **8.62%** | 10,465 | **10.99%** |
| **Total number/percentage** | 7,930 | 100.00% | 50,885 | 100.00% | 95,234 | 100.00% |

## Distribution and Density of SNPs in the Genome

Chromosome density distribution of the SNPs with allelic imbalances exhibited high density for all five traits in several chromosomes with the three highest peaks in chromosomes 9, 20 and 28 (Figure 17A). All five traits revealed very similar pattern of distribution with a single exception; shear force exhibited a relative higher density than the other traits on chromosome 9. The similarity in density distribution between traits may be explained at least in part by the positive correlation that we observed between the phenotypes in this population. WBW and thermal growth coefficient were used as selection criterion in this population [50, 219], and we found that WBW as an independent variable has significant effects on muscle yield and fat percentage (multivariable regression analysis [P<0.01], data not shown). However, despite the similarity in SNP density distributions, most of the identified SNPs were unique to each trait. From the 7,930 SNPs with allelic imbalances, only 27 were shared by all five traits, 161 were shared by four traits, 680 were shared by three traits and 1,783 were shared by two traits. In agreement with our results, a recent GWAS study identified two windows with effect on fillet yield located on chromosome 9 and explaining 1.0–1.5% of genetic variance in the same fish population [205].

As can be expected, the number of SNPs with allelic imbalances per chromosome was strongly correlated with chromosome length (Figure 17 B). In general, numbered unknown chromosomes, which are longer in the current reference genome [25], had more SNPs compared to the known chromosomes (Figure 17 B). Chromosome “Unknown” (1.1 Gb of scaffolds not assigned to chromosomes) had 4,086 (49.05%) SNPs (not shown in Figure 17 B). Previous genetic mapping reports showed that the growth-related SNPs/QTL are distributed over ~20 chromosomes [50, 240, 241]. Together with our data, these reports confirm the polygenetic nature of growth/muscle related traits in rainbow trout.

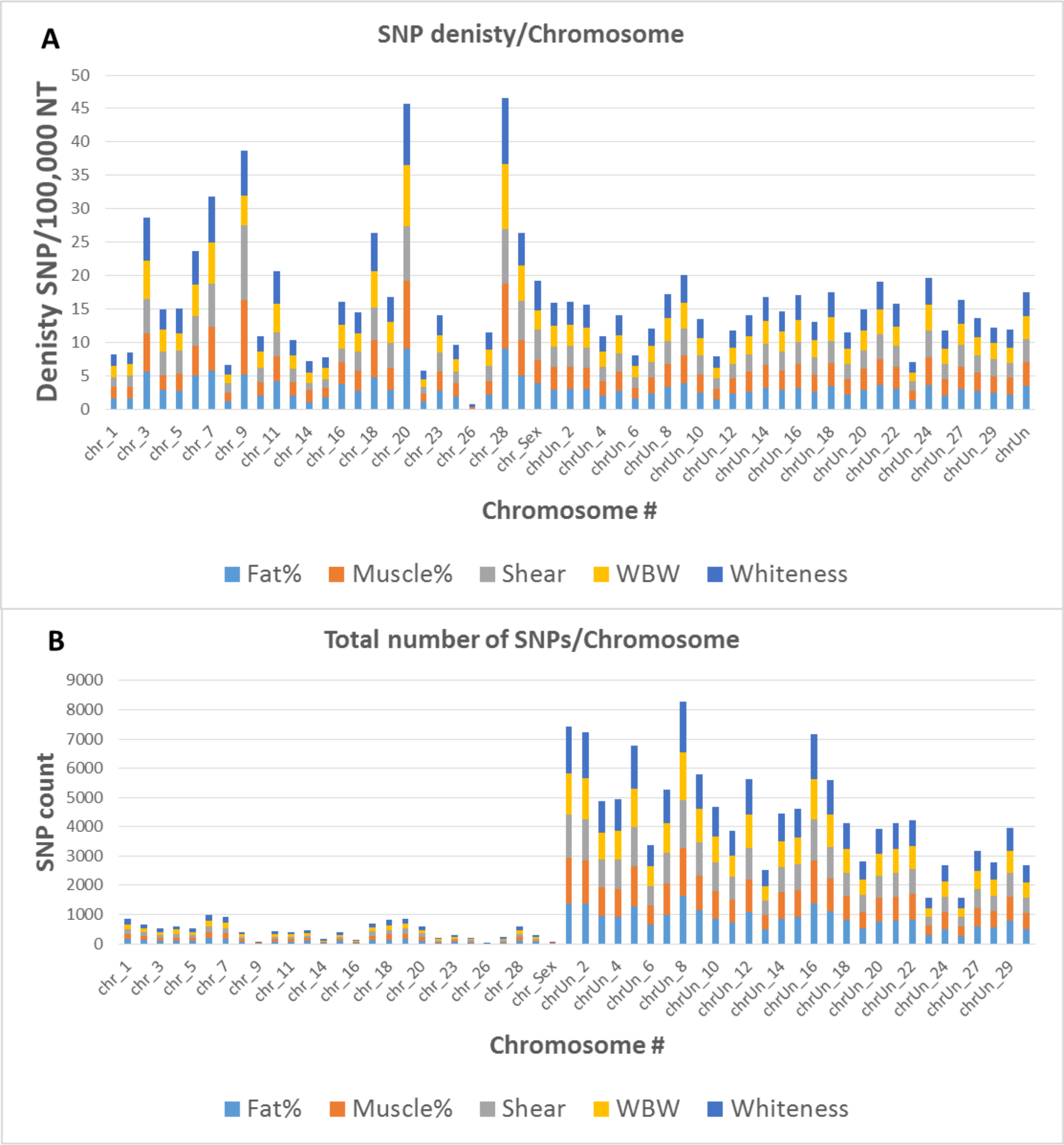


Figure 17: Genome distribution of the SNPs with allelic imbalances for all five traits. SNP density (SNPs per 100,000 NT) (A) and total number of SNPs (B) are shown for each chromosome. Chromosome "Unknown" (1.1 Gb scaffolds not assigned to chromosomes) had 4,086 (49.05%) SNPs is not shown in the lower panel.

### SNP Functional Annotation

Functional annotation of genes harboring SNPs with allelic imbalances were performed using the Blast2GO suite [102]. The SNP-flanking sequences were searched against the NCBI nr-protein database using BLASTx; then, associated genes and Gene Ontology (GO) terms were acquired. In the biological processes category, SNP-harboring genes were associated with various cellular processes mainly involved in growth-related mechanisms, including regulation of metabolic and oxidation-reduction processes and protein translation (Figure 18). In the molecular function category, SNP-containing genes were associated with binding metal ions, ATP, nucleic acid, and actin. In addition, a significant number of the genes were associated with transferase, motor, oxidoreductase, and structural molecule activities (Figure 18). In the cellular component category, many of the genes exhibited association with the cytoplasmic compartment, membranes, myosin complex, and extracellular region compartment (Figure 18). Genes with similar GO associated terms were previously reported to be involved in rainbow trout muscle growth and quality [50, 214, 240, 242-244].

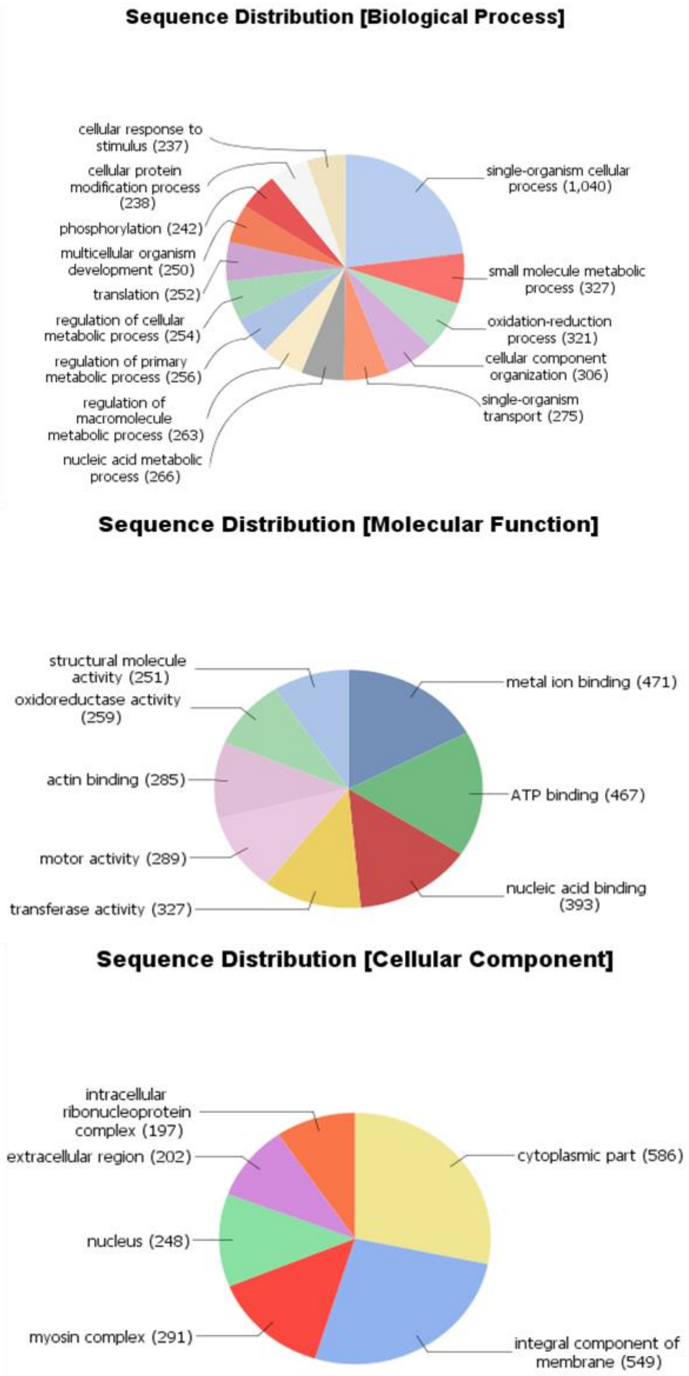


Figure 18: Gene Ontology (GO) assignment of the genes harboring SNPs with allelic imbalances in families with contrasting growth and muscle phenotypes.

Additionally, KEGG pathway mapping was used to assign enzyme function to the SNP-containing transcripts [226]. Searching transcripts against the KEGG database yielded 1,043 transcripts (13.15%) with significant KEGG hits to 632 KEGG Orthologies (KOs) belonging to different pathways (Table 12). Most of the transcripts were assigned to growth-related metabolic pathways. There were 275 transcripts (182 KOs) related to metabolism. Under this category, sequences matching energy metabolism (88 transcripts, 57 KOs) appeared on the top of the list, with 52 transcripts (37 KOs) assigned to oxidative phosphorylation. Sequences matching carbohydrate metabolism occupied the second place (77 transcripts, 43 KOs) and were further classified into glycolysis/gluconeogenesis (39 transcripts, 18 KOs), citrate cycle (19 transcripts, 14 KOs) and pyruvate metabolism (16 transcripts, 10 enzymes). The next metabolic subcategories in the metabolic list were amino acid metabolism (56 transcripts, 41 KOs), lipid metabolism (27 transcripts, 22 KOs), and cofactors and vitamins metabolism (14 transcripts, 11 KOs). These preliminary SNP functional annotations are in agreement with previous reports that showed strong association between 1) mutations and altered expression of glycolytic and oxidative phosphorylation enzymes and 2) rainbow trout growth and muscle degeneration [50, 214, 240, 242, 243].

In addition, 176 KEGG annotated sequences were assigned to the genetic information processing category (112 KOs) that included translation (105 sequences, 69 KOs), folding, sorting and degradation (62 sequences, 38 KOs), and transcription (9 sequences, 5 KOs) (Table 12). A significant number of the SNP-harboring genes matched ribosomal (68 sequences, 48 KOs) and RNA-transport proteins (22 sequences, 12 KOs). Previously, we showed that the atrophying muscle and muscle from fast versus slow growing rainbow trout had differentially expressed genes involved in RNA processing, protein synthesis, posttranslational modification, and intracellular protein trafficking [214, 240, 242].

Moreover, 166 sequences (99 KOs) were classified by KEGG mapping into the environmental information processing category; these sequences were further assigned to signal transduction (147 sequences, 87 KOs) and signaling and interaction molecules (19 sequences, 12 KOs) (Table 12). The PI3K-Akt signaling, Calcium signaling, MAPK signaling, and cGMP-PKG signaling pathways had the largest numbers of hits: 21, 18, 18, and 16 KOs, respectively. Previous studies indicated involvement of MAPK and Calcium signaling in fish/muscle growth [242, 245].

Furthermore, the cellular processes category contained 152 KEGG-annotated sequences matching 85 KOs, which were further classified into cellular community (54 transcripts, 27 KOs), transport and catabolism (42 transcripts, 24 KOs), and cell growth and death (36 transcripts, 22 KOs) (Table 12). In the organismal systems category, the most significant subcategories were endocrine (105 transcripts, 53 KOs), circulatory (49 transcripts, 30 KOs), immune (44 transcripts, 28 KOs), and digestive systems (32 transcripts, 16 KOs). Recently, a GWAS study using the same fish population identified a small number of genes involved in muscle development explaining ~1.0% of the total genetic variance of the muscle yield and growth rate [205].

Distributions of KEGG matches were generally similar among all five traits. Albeit, we noticed an increased number of hits related to fillet whiteness compared to other traits, for carbohydrate metabolism (47 transcripts, 28 KOs) and amino acid metabolism (32 transcripts, 26 KOs) (Table 12). Similarly, there was a noticeable increase in numbers of hits in whiteness for PI3K-Akt signaling, focal adhesion, gap junction and regulation of actin cytoskeleton (Table 12). Regulation of focal adhesion and actin cytoskeleton were associated with development of pale, soft, and exudative (PSE) meat in turkey [246]. In addition, the muscle yield trait exhibited an increased number of transcripts for energy metabolism, with 28 transcripts/18 KOs belonging to oxidative phosphorylation.  Shear force exhibited an increased number of transcripts belonging to lipid metabolism (16 transcripts, 14 KOs) (Table 12).

Our KEGG pathway mapping results have linked many of the genes harboring SNPs with allelic imbalances to potential regulation of growth and metabolic pathways, which may support pathway-based GWAS analyses in rainbow trout, similar to what has been recently applied to detect genetic pathways explaining live weight and muscle growth variation in cattle genotypes [247].

Table 12: KEGG biochemical mapping of the genes harboring SNPs with allelic imbalances in fish families showing contrasting growth and muscle phenotypes.

| | **Total (all traits)** | | **WBW** | | **Muscle %** | | **Fat%** | | **Shear** | | **Whiteness** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **KEGG categories** | **No. of sequences(%)** | **No. of KOs** | **No. of sequences(%)** | **No. of KOs** | **No. of sequences(%)** | **No. of KOs** | **No. of sequences(%)** | **No. of KOs** | **No. of sequences(%)** | **No. of KOs** | **No. of sequences(%)** | **No. of KOs** |
| **Metabolism** | **275** | **182** | **133** | **99** | **130** | **96** | **102** | **77** | **125** | **92** | **142** | **105** |
| **Energy Metabolism** | 88 (32.00) | 57 | 42 (31.58) | 32 | 45 (34.62) | 29 | 37 (36.27) | 26 | 30 (24.00) | 21 | 41 (28.87) | 32 |
| *Oxidative phosphorylation* | 52 | 37 | 20 | 17 | 28 | 18 | 18 | 15 | 13 | 11 | 19 | 16 |
| **Carbohydrate Metabolism** | 77 (28.00) | 43 | 42 (31.58) | 28 | 39 (30.00) | 27 | 26 (25.49) | 19 | 39 (31.20) | 25 | 47 (33.10) | 28 |
| *Glycolysis / Gluconeogenesis* | 39 | 18 | 19 | 10 | 19 | 14 | 13 | 7 | 21 | 11 | 20 | 13 |
| *Citrate cycle (TCA cycle)* | 19 | 14 | 8 | 8 | 9 | 7 | 7 | 6 | 10 | 9 | 9 | 7 |
| *Pyruvate metabolism* | 16 | 10 | 6 | 4 | 7 | 7 | 5 | 4 | 10 | 8 | 6 | 5 |
| *Pentose phosphate pathway* | 13 | 5 | 8 | 5 | 7 | 5 | 7 | 3 | 3 | 2 | 3 | 3 |
| **Amino Acid Metabolism** | 56 (20.36) | 41 | 26 (19.55) | 21 | 24 (18.46) | 20 | 21 (20.59) | 16 | 28 (22.40) | 22 | 32 (22.54) | 26 |
| **Lipid Metabolism** | 27 (9.82) | 22 | 14 (10.53) | 11 | 11 (8.46) | 11 | 7 (6.86) | 6 | 16 (12.80) | 14 | 13 (9.15) | 11 |
| *Fatty acid degradation* | 15 | 13 | 8 | 6 | 9 | 9 | 4 | 3 | 11 | 10 | 8 | 7 |
| **Metabolism of Cofactors and Vitamins** | 14 (5.09) | 11 | 4 (3.01) | 4 | 6 (4.62) | 4 | 5 (4.90) | 5 | 5 (4.00) | 5 | 4 (2.82) | 4 |
| **Nucleotide Metabolism** | 13 (4.73) | 8 | 5 (3.76) | 3 | 5 (3.85) | 5 | 6 (5.88) | 5 | 7 (5.60) | 5 | 5 (3.52) | 4 |
| **Genetic Information Processing** | **176** | **112** | **69** | **50** | **79** | **59** | **50** | **40** | **83** | **59** | **74** | **55** |
| **Translation** | 105 (59.66) | 69 | 36 (52.17) | 31 | 48 (60.76) | 39 | 30 (60.00) | 27 | 45 (54.22) | 35 | 47 (63.51) | 40 |
| *Ribosome* | 68 | 48 | 25 | 23 | 32 | 26 | 21 | 19 | 33 | 27 | 32 | 28 |
| *RNA transport* | 22 | 12 | 9 | 6 | 11 | 9 | 7 | 6 | 8 | 4 | 7 | 6 |
| **Folding, Sorting and Degradation** | 62 (35.23) | 38 | 28 (40.58) | 17 | 24 (30.38) | 16 | 19 (38.00) | 12 | 30 (36.14) | 19 | 23 (31.08) | 13 |
| *Protein processing in endoplasmic reticulum* | 23 | 14 | 11 | 7 | 7 | 4 | 7 | 5 | 14 | 9 | 7 | 4 |
| *RNA degradation* | 16 | 5 | 11 | 4 | 10 | 5 | 8 | 4 | 9 | 3 | 11 | 5 |
| *Proteasome* | 12 | 11 | 4 | 4 | 5 | 5 | 1 | 1 | 4 | 4 | 4 | 3 |
| *Ubiquitin mediated proteolysis* | 9 | 7 | 2 | 2 | 2 | 2 | 1 | 1 | 3 | 3 | 1 | 1 |
| **Transcription** | 9 (5.11) | 5 | 5 (7.25) | 2 | 7 (8.86) | 4 | 1 (2.00) | 1 | 8 (9.64) | 5 | 4 (5.41) | 2 |
| *Spliceosome* | 9 | 5 | 5 | 2 | 7 | 4 | 1 | 1 | 8 | 5 | 4 | 2 |
| **Environmental Information Processing** | **166** | **99** | **70** | **45** | **88** | **61** | **62** | **45** | **76** | **55** | **87** | **58** |
| **Signal Transduction** | 147 (88.55) | 87 | 62 (88.57) | 39 | 79 (89.77) | 53 | 56 (90.32) | 41 | 69 (90.79) | 50 | 74 (85.06) | 50 |
| *PI3K-Akt signaling pathway* | 35 | 21 | 12 | 8 | 13 | 10 | 12 | 8 | 13 | 10 | 24 | 16 |
| *Calcium signaling pathway* | 36 | 18 | 16 | 9 | 13 | 8 | 14 | 10 | 16 | 12 | 16 | 11 |
| *MAPK signaling pathway* | 26 | 18 | 10 | 6 | 17 | 12 | 7 | 6 | 12 | 9 | 10 | 8 |
| *cGMP - PKG signaling pathway* | 26 | 16 | 10 | 7 | 8 | 8 | 7 | 6 | 10 | 8 | 12 | 10 |
| *AMPK signaling pathway* | 21 | 12 | 10 | 5 | 14 | 8 | 9 | 3 | 10 | 6 | 12 | 7 |
| *cAMP signaling pathway* | 18 | 12 | 6 | 5 | 4 | 4 | 7 | 6 | 8 | 7 | 7 | 6 |
| *HIF-1 signaling pathway* | 11 | 9 | 4 | 2 | 7 | 4 | 3 | 3 | 9 | 6 | 8 | 4 |
| *Hippo signaling pathway* | 13 | 7 | 2 | 2 | 7 | 5 | 6 | 6 | 5 | 5 | 5 | 4 |
| *FoxO signaling pathway* | 7 | 6 | 3 | 3 | 2 | 2 | 3 | 3 | 1 | 1 | 4 | 3 |
| *mTOR signaling pathway* | 5 | 5 | 1 | 1 | 2 | 2 | 0 | 0 | 1 | 1 | 3 | 3 |
| Signaling Molecules and Interaction | 19 (11.45) | 12 | 8 (11.43) | 6 | 9 (10.23) | 8 | 6 (9.68) | 4 | 7 (9.21) | 5 | 13 (14.94) | 8 |
| *ECM-receptor interaction* | 17 | 10 | 7 | 5 | 7 | 6 | 6 | 4 | 7 | 5 | 13 | 8 |
| *Cell adhesion molecules* | 3 | 2 | 1 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 2 | 1 |
| **Cellular Processes** | **152** | **85** | **68** | **41** | **70** | **42** | **54** | **42** | **64** | **44** | **83** | **56** |
| **Cellular commiunity** | 54 (35.53) | 27 | 27 (39.71) | 13 | 29 (41.43) | 15 | 26 (48.15) | 16 | 27 (42.19) | 16 | 36 (43.37) | 21 |
| *Focal adhesion* | 35 | 21 | 13 | 10 | 17 | 11 | 14 | 11 | 15 | 11 | 23 | 17 |
| *Tight junction* | 19 | 10 | 18 | 6 | 17 | 6 | 15 | 8 | 16 | 7 | 10 | 6 |
| *Gap junction* | 8 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 7 | 2 |
| *Adherens junction* | 5 | 3 | 4 | 3 | 5 | 3 | 3 | 3 | 5 | 3 | 3 | 3 |
| **Transport and Catabolism** | 42 (27.63) | 24 | 17 (25.00) | 11 | 20 (28.57) | 14 | 8 (14.81) | 8 | 16 (25.00) | 11 | 18 (21.69) | 11 |
| **Cell Growth and Death** | 36 (23.68) | 22 | 16 (23.53) | 11 | 12 (17.14) | 8 | 13 (24.07) | 12 | 14 (21.88) | 13 | 19 (22.89) | 16 |
| *Apoptosis* | 19 | 13 | 13 | 7 | 5 | 4 | 6 | 6 | 7 | 7 | 11 | 9 |
| *p53 signaling pathway* | 7 | 5 | 6 | 4 | 3 | 1 | 1 | 1 | 3 | 3 | 4 | 4 |
| **Cell Motility** | 20 (13.16) | 12 | 8 (11.76) | 6 | 9 (12.86) | 5 | 7 (12.96) | 6 | 7 (10.94) | 4 | 10 (12.05) | 8 |
| *Regulation of actin cytoskeleton* | 20 | 12 | 8 | 6 | 9 | 5 | 7 | 6 | 7 | 4 | 10 | 8 |
| **Organismal Systems** | **274** | **154** | **108** | **66** | **124** | **84** | **109** | **81** | **122** | **82** | **129** | **89** |
| **Endocrine System** | 105 (38.32) | 53 | 44 (40.74) | 25 | 49 (39.52) | 32 | 36 (33.03) | 24 | 53 (43.44) | 33 | 56 (43.41) | 33 |
| *Glucagon signaling pathway* | 36 | 12 | 19 | 8 | 14 | 9 | 12 | 7 | 22 | 11 | 23 | 10 |
| *Insulin signaling pathway* | 32 | 12 | 14 | 7 | 11 | 6 | 7 | 4 | 12 | 6 | 22 | 10 |
| *Thyroid hormone signaling pathway* | 11 | 7 | 4 | 4 | 6 | 5 | 3 | 3 | 4 | 4 | 6 | 6 |
| *Thyroid hormone synthesis* | 6 | 4 | 3 | 2 | 1 | 1 | 3 | 3 | 2 | 2 | 1 | 1 |
| **Circulatory System** | 49 (17.88) | 30 | 18 (16.67) | 12 | 22 (17.74) | 15 | 19 (17.43) | 16 | 16 (13.11) | 13 | 15 (11.63) | 12 |
| **Immune System** | 44 (16.06) | 28 | 16 (14.81) | 10 | 21 (16.94) | 14 | 22 (20.18) | 17 | 24 (19.67) | 16 | 23 (17.83) | 17 |
| **Digestive System** | 32 (11.68) | 16 | 13 (12.04) | 9 | 9 (7.26) | 7 | 18 (16.51) | 11 | 13 (10.66) | 9 | 20 (15.50) | 14 |
| *Protein digestion and absorption* | 12 | 5 | 6 | 4 | 7 | 5 | 8 | 5 | 5 | 3 | 8 | 4 |
| *Mineral absorption* | 4 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 |
| **Nervous System** | 27 (9.85) | 17 | 7 (6.48) | 5 | 12 (9.68) | 9 | 11 (10.09) | 10 | 10 (8.20) | 9 | 9 (6.98) | 8 |
| **Aging** | 17 (6.20) | 10 | 10 (9.26) | 5 | 11 (8.87) | 7 | 3 (2.75) | 3 | 6 (4.92) | 2 | 6 (4.65) | 5 |
| **Total** | 1,043 | 632 | 448 | 301 | 491 | 342 | 377 | 285 | 470 | 332 | 515 | 363 |

## Acknowledgements

This study was supported by a competitive grant No. 2014-67015-21602 from the United States Department of Agriculture, National Institute of Food and Agriculture.

# CHAPTER V: DIFFERENTIAL EXPRESSION OF LONG NON-CODING RNAS IN THREE GENETIC LINES OF RAINBOW TROUT IN RESPONSE TO INFECTION WITH FLAVOBACTERIUM PSYCHROPHILUM [248]

## Introduction

World aquaculture industries suffer considerable economic losses annually because of infectious diseases [249]. *Flavobacterium psychrophilum (Fp),* a causative agent of Bacterial Cold Water Disease (BCWD), saddleback disease, fry mortality syndrome, or rainbow trout fry syndrome causes significant loss of trout and salmon each year and is a threat to many other salmonids (see review [250]). Infection of rainbow trout with *Fp* results in mortality of up to 30% and several complications in the survivors [251]. Originally, the pathogen was considered to be endemic to North America but in recent years it has been reported from almost every continent [252]. Multiple routes of transmission [253], wide geographical distribution, the ability of pathogen to cope with harsh survival condition [253], limited chemotherapeutic agents, and lack of a commercial vaccine make control measures inefficient. Live-attenuated *Fp* vaccines can provide protection against BCWD but environmental safety is a concern (see review [254]).

Harnessing the host’s immune system by selective breeding is a strategy being pursued to improve farmed fish health [255]. In order to improve resistance of rainbow trout against *Fp*, the National Center for Cool and Cold Water Aquaculture (NCCCWA) started a family-based selective breeding program in 2005. A closed genetic line, designated ARS-

Fp-R, has undergone multiple generations of selection for increased survival following standardized challenge. This line has improved disease resistance against *Fp* infection in both laboratory and field settings compared to a susceptible (ARS-Fp-S) and randomly bred control (ARS-Fp-C) lines [256]. Previously, we performed global expression analysis of protein-coding genes in these genetic lines upon *Fp* challenge [257]. The study identified a large number of DE protein-coding genes among genetic lines, a significant proportion of which were genes with described roles in the immune response, especially the innate immune system. We demonstrated transcriptome differences between lines in the absence of infection. However, altered transcriptome abundance of lncRNAs among genetic lines after mock and *Fp* infection was not addressed.

LncRNAs have appeared as critical regulators of transcription and post-transcriptional events of protein-coding genes [151]. LncRNAs regulate diverse cellular processes, including disease, immunity, development and cell proliferation [258]. In mammals, lncRNAs regulate various immune responses including the interferon response, inflammatory processes, and other aspects of innate and adaptive immune responses [170, 259-263]. TLR signaling and inflammatory responses increase the expression of lncRNA-Cox2 that regulates both activation and repression of innate response genes [259]. LncRNA NeST controls susceptibility to Theiler's virus and Salmonella infection through epigenetic regulation of the interferon-γ locus [263, 264]. A distinct differential expression profile of lncRNAs in response to microbial infection has been reported in mammals and salmonids, suggesting involvement of a set of lncRNAs in host defense against microbes [258, 265]. To date, most of the studies in the field of lncRNA influence on immune processes are limited to mammalian species, especially human and mouse. To the best of our knowledge, there are no studies exploring the expression of lncRNAs during host defense against bacterial infection in aquaculture finfish.  Such studies are difficult as low evolutionary conservation of lncRNAs across species prevents utilization of the information from mammalian species into aquaculture animals.

The overall objective of this study was to identify lncRNAs that are associated with genetic resistance against *Fp* and to identify immune-relevant protein-coding genes that might be regulated by lncRNAs.  To study the expression of lncRNA, we utilized a reference dataset that we recently identified (31,195 lncRNA) in rainbow trout [115].  Using the abovementioned three genetic lines of rainbow trout, we were able to characterize the transcriptome profile of lncRNAs associated with the early response to *Fp* infection.  We have identified DE lncRNAs between genetic lines of naive animals and in response to infection, identified their genomic co-localization relative to immune-relevant protein-coding genes, and explored their co-expression relationships to suggest possible regulation of immune-relevant protein-coding genes by lncRNAs.

## Materials and Methods

### Ethics statement

Fish were maintained at the NCCCWA and all experimental protocols and animal procedures were approved and carried out in accordance with the guidelines of NCCCWA Institutional Animal Care and Use Committee Protocols #053 and #076.

### Experimental animals and RNA-Seq experimental design

Three rainbow trout genetic lines ARS-Fp-R, ARS-Fp-C, and ARS-Fp-S used in this study were developed at National Center for Cool and Cold Water Aquaculture (NCCCWA) rainbow trout breeding program. These genetic lines differ significantly to their susceptibility to *Fp* infection as a result of genetic selection [256] and we have previously reported the challenge experiment utilized in this study [257]. Briefly, fifty randomly selected fish from each genetic line were assigned to four challenge tanks (total 12 tanks for three genetic lines). At the time of challenge, average body weight was 1.1g and fish age was 49 days post-hatch. For each genetic line, fish in two tanks were injected with *Fp* (experimental group) and fish in the other two tanks were injected with PBS (control group). Fish were injection challenged with either $4.2 \times 10^6$ CFU *Fp* suspended in 10 µl of chilled PBS or PBS alone, and survival was monitored daily for 21 days [257]. For RNA extraction, five individuals were sampled from each tank on days 1 and 5 post infections. Survival at 21 days post-challenge injection was monitored during the experiment. Post-challenged bacterial load in the body was measured in a subset of fish by qPCR and was expressed in terms of *Fp* genome equivalents (GE).

### RNA extraction, library preparation, and sequencing

Tissue sampling, RNA extraction, library preparation and sequencing were done as described previously [257]. Briefly, total RNA was extracted and equal amounts of RNA from five fish were pooled from each of the 12 tanks at each of the two time points (total of 24 pools, n = 120 fish total). cDNA libraries were prepared using Illumina's TruSeq Stranded mRNA Sample Prep kit following the manufacturer's instructions. The 24 indexed and barcoded libraries were randomly divided into three groups (eight libraries per group) and sequenced in three lanes of an Illumina HiSeq 2000 (single-end, 100 bp read length). RNA-Seq reads are available at the NCBI Short Read Archive (BioProject ID PRJNA259860, accession number SRP047070).

### Differential gene expression analysis of lncRNAs

Complete description of lncRNA reference dataset with their discovery pipeline has been recently described [115]. From this discovery datasets, a stringently selected set of lncRNAs (31,195) were used as a reference for gene expression analysis. For differential gene expression analysis, sequencing reads from each library were mapped to the lncRNA reference using a CLC genomics workbench. Mapping conditions were, mismatch cost=2, insertion/deletion cost=3, minimum length fraction=0.9 and similarity fraction =0.9. The expression value of lncRNAs was calculated in terms of RPKM (reads per kilobase per million). EDGE (extraction and analysis of differential gene expression) tests were performed to identify DE genes between various groups, e.g. infected vs. non-infected, day 1 vs. day 5, and one genetic line vs. other with or without *Fp* injection [51]. To control false discovery due to multiple testing, p-values were FDR-corrected. LncRNA was considered significant at a fold-change cutoff value of ±2 and a corrected p-value of less than 0.05.

### Validation of RNA-Seq data by qPCR

From DE lncRNAs in the RNA-Seq study, 7 were randomly selected from the DE day 5 susceptible line for experimental validation using individual (unpooled) samples. RNA isolation, cDNA synthesis and primer design were completed using the same technique as described previously [257]. Briefly, RNAs were treated with Optimize™ DNAase I (Fisher Bio Reagents, Hudson, NH) to eliminate genomic DNA. One microgram of the purified RNA was converted to cDNA using the Verso cDNA Synthesis Kit (Thermo Scientific, Hudson, NH) according to the manufacturer protocol. Reverse transcription was performed using My Cycler™ Thermal Cycler (Bio Rad, Hercules, CA) at 42°C for 30 min (one cycle amplification) followed by 95°C for 2 min (inactivation). Blend of random hexamer and oligo (dT) primer (3:1 V/V), at a final concentration of 25 ng/μL, was used to prime the reverse transcription reaction.

The Bio-Rad CFX96™ Real Time System (Bio-Rad, Hercules, CA) in conjunction with SsoAdvanced™ Universal SYBR® Green Supermix (Bio-Rad, Hercules, CA, USA) was used to quantify the amount of the expressed gene of interest in PBS and Fp injected whole-body fish homogenates. Each primer was used at a concentration of 0.1 nM/μL and cDNA template was used at a concentration of 0.006 μg/μL. Cycling temperatures were set up according to the manufacturer's protocol and different annealing temperatures were used depending on primers. Fold change in gene expression was calculated as described previously [257]. Briefly, β-actin (Accession: AJ438158) was used as endogenous reference to normalize each target lncRNA. qPCR data were quantified using delta delta Ct (ΔΔCt) methods [266]. Ct-values of β-actin were subtracted from Ct-values of the target gene to calculate the normalized value (ΔCt) of the target lncRNA in both the calibrator samples (PBS-injected) and test samples (Fp-injected). The ΔCt value of the calibrator sample was subtracted from the ΔCt value of the test sample to get the ΔΔCt value. Fold change in gene expression in the test sample relative to the calibrator sample was calculated by the formula 2−ΔΔCt and the normalized target Ct values in each infected and non-infected group was averaged. Correlation between gene expression fold-change measured by qPCR and RNA-Seq was performed by Pearson correlation. All statistics were performed with a significance of $P < 0.05$.

### Gene clustering and gene expression correlation

Sequencing reads from all 24 libraries (samples) were mapped to a combined reference sequence consisting of all lncRNAs, that we previously identified [115], and mRNAs that were identified in the rainbow trout genome [25]. Expression of lncRNAs and protein-coding genes was measured in terms of RPKM. The expression value of each transcript in each sample was normalized using the scaling method [267]. Mean was chosen as normalization value and median mean was chosen as reference. Five percent of the data on both sides of the tail were trimmed. Normalized expression values of transcripts in each sample were used to cluster protein-coding genes and lncRNAs using algorithms in Multi-experiment Viewer (MeV). Clusters were generated with a minimum correlation coefficient of 0.92. During clustering, 30% of the sequences with flat expression values over samples were excluded from cluster generation to prevent uninteresting cluster generation. Correlation in expression of lncRNAs and neighboring/overlapped protein-coding genes was performed in Excel using regression analysis using normalized expressions values of the transcripts.

### Discovery of novel lncRNAs in resistant and susceptible genetic lines

Novel lncRNA were identified according to Al-Tobasei et al., 2016 [115]. Briefly, sequencing reads from each genetic lines (resistance, control and susceptible) were aligned to a rainbow trout reference genome using TopHat [25]. Cufflinks, Cufflinks compare and

Cufflinks Merge were used to predict transcripts in each genetic line. Transcripts shorter than 200 nt were filtered out using in house Perl script. Transcripts which had open reading frame (ORF) longer than 100 amino acids were removed. In addition, if ORF of the transcript is longer than 35% of the transcript length, the transcript was filtered out even if the ORF is shorter than 100 amino acids. Subsequently, transcripts were searched against NR protein database (updated on May 2016) using BLASTx, and any transcripts with sequence homology to existing proteins were removed. To remove any remaining protein coding transcripts, coding potential calculator (CPC) was applied to the transcripts (Index value <-1.0). Other classes of non-coding RNAs (e.g. rRNA, tRNA, snoRNA, miRNA, siRNA and others) in the dataset were removed by blasting (BLASTn) the transcripts against multiple RNA databases including genomic tRNA database. Finally, any single exon transcripts within 500 nts of protein coding gene was removed. After these filtration steps, remaining transcripts were considered as putative lncRNAs. To identify lncRNAs specific to a particular genetic line, lncRNAs from one genetic line were compared with lncRNAs from other two genetic lines. Resistant and susceptible specific lncRNA were reported.

## Results and Discussion

### Global expression of lncRNA across dataset

Previously, we analyzed mRNA expression in three genetic lines of fish sampled at 1 and 5 days post-*Fp* challenge [257]. In our prior analyses, slightly more than half (51.77%) of the RNA-Seq reads aligned to the 46,585 predicted coding mRNAs and thus considerable sequence information remained unaligned and thus enigmatic. In present study, on average, 8.2% of the total RNA-Seq reads aligned to the 31,195 lncRNAs reference (Supplementary Dataset 1A) [268]. 94.5% of the reads were uniquely mapped to the reference. On average, each dataset expressed 87.2% of the putative reference lncRNA's at RPKM cut off ≥ 0.5. Out of 31,195 reference lncRNAs, only 933 were not expressed in any dataset (RPKM ≥ 0.5). One possible explanation of the low percentages of aligned read to lncRNA reference compared to protein coding mRNAs might be due to the lower lncRNAs expression compared to mRNAs. Recently, we reported that the average RPKM of the most abundant 40,000 transcripts was 3.49 and 15.69 in LncRNAs and protein-coding genes, respectively [115]. In this study, RNA was sequenced from a whole-body extract, which may be another reason for the low percentage of mappable reads because reference lncRNA dataset was sequenced from about 13 specific tissues. Out of the 933lncRNAs, only 109 were tissue specific indicating that most of the 933 are very lowly expressed on all tissues.

We utilized pairwise comparisons between different genetic lines and days of infection to identify a sum of 937 DE lncRNA from all comparisons (FDR-P-value <0.05) (Table 13). Of these, 556 were unique lncRNA showing differential expression in at least one comparison (Supplementary Dataset 2 tab "ALL DEF, non-redundant") [269]. In our previous study using the same genetic lines, ~2,600 DE immune-related and other protein-coding genes were identified in response to *Fp* infection [257]. We quantified the number of DE lncRNAs between different genetic lines and infection statuses (total 24 comparisons) and compared the number with that of DE protein-coding genes. Numbers of DE protein-coding genes and lncRNAs showed moderate positive correlation ($R^2$=0.40, p=0.0011) (Table 13). In general, within each pair-wise comparison, fewer differentially regulated lncRNA were identified as compared to DE protein coding transcripts (Table 13). This may, in part, be, due to the overall lower expression level of lncRNA as compared to protein-coding genes [43]. Numbers of the DE protein-coding genes as well as lncRNAs positively correlated with bacterial load in the body. The susceptible line showed more DE lncRNAs as well as protein-coding genes compared to the resistant and control genetic lines (Table 13). Similarly, more transcripts were expressed on day 5 of infection than on day 1. Correlation between body bacterial load and the number of DE lncRNAs on the 5th day of infection in control, susceptible and resistant genetic lines was strongly positively correlated ($R^2$>0.99); however, correlation of body bacterial load with the number of DE protein coding-genes was moderately positive ($R^2$=0.34). This finding suggests that, like protein-coding genes, lncRNAs may play a role in the host defense against *Fp*. Expression trends of seven randomly chosen regulated lncRNAs was verified by real time PCR. A consistent trend ($R^2$=0.84) between RNA-Seq and qPCR was observed, albeit with a somewhat lower relative expression measured by qPCR for 6 of the 7 measured lncRNA's (Supplementary Dataset 1B). Information about primers and the real time PCR cycling program is provided in Supplementary Dataset 1C.

Table 13: Comparison of differentially expressed lncRNA and protein coding genes in response to *Fp* infection. Four different comparisons were made to quantify the differentially expressed transcripts: infected vs. non-infected, one genetic line vs. another without infection, one genetic line vs. another post infection, and day 1 vs. day 5 of infection. Differential expression was considered at fold change ±2 and FDR-corrected $p<0.05$. Number of differentially expressed protein coding genes and lncRNAs showed positive correlation ($R^2=0.40$, $P=0.0011$).

| **Comparison** | **Day, genetic line and infection status** | **No. differentially expressed protein-coding genes[1]** | **No. differentially expressed lncRNAs** |
|---|---|---|---|
| Infected vs PBS | Day 1 R-line (Fp) vs. R-line (PBS) | 515 | 57 |
| | Day 5 R-line (Fp) vs. R-line (PBS) | 428 | 36 |
| | Day 1 C-line (Fp) vs. C-line (PBS) | 20 | 0 |
| | Day 5 C-line (Fp) vs. C-line (PBS) | 2201 | 54 |
| | Day 1 S-line (Fp) vs. S-line (PBS) | 1663 | 125 |
| | Day 5 S-line (Fp) vs. S-line (PBS) | 2225 | 196 |
| Genetic lines (PBS) | Day 1 R-line (PBS) vs. S-line (PBS) | 76 | 24 |
| | Day 1 R-line (PBS) vs. C-line (PBS) | 3 | 2 |
| | Day 1 S-line (PBS) vs. C-line (PBS) | 28 | 6 |
| | Day 5 R-line (PBS) vs. S-line (PBS) | 45 | 22 |
| | Day 5 R-line (PBS) vs. C-line (PBS) | 246 | 28 |
| | Day 5 S-line (PBS) vs. C-line (PBS) | 61 | 25 |
| Genetic lines (Fp) | Day 1 R-line (Fp) vs. S-line (Fp) | 150 | 15 |
| | Day 5 R-line (Fp) vs. S-line (Fp) | 1016 | 83 |
| | Day 1 R-line (Fp) vs. C-line (Fp) | 28 | 12 |
| | Day 5 R-line (Fp) vs. C-line (Fp) | 159 | 21 |
| | Day 1 S-line (Fp) vs. C-line (Fp) | 37 | 13 |
| | Day 5 S-line (Fp) vs. C-line (Fp) | 1758 | 5 |
| Time points | Day 5 vs. Day 1 R-line (PBS) | 1286 | 26 |
| | Day 5 vs. Day 1 C-line (PBS) | 294 | 36 |
| | Day 5 vs. Day 1 S-line (PBS) | 376 | 14 |
| | Day 5 vs. Day 1 R-line (Fp) | 334 | 22 |
| | Day 5 vs. Day 1 C-line (Fp) | 2469 | 70 |
| | Day 5 vs. Day 1 S-line (Fp) | 2434 | 45 |

[1] Data are from [257]

Recently, we reported tissue specificity of lncRNAs in rainbow trout [115]. A total of 35 DE lncRNAs were selectively expressed in specific tissues, 10 of them were gill-specific.

Out of 13 vital tissues, liver, spleen and head kidney did not have any DE lncRNA. Spleen and head kidney lymphoid organ are mainly involved in generation of antibody response and other humoral components of immune system, but in early phase of BCWD, the first line of defense includes skin, alimentary tract lining, and gill [270].

### Differential expression of lncRNAs between Fp infected and PBS injected fish

LncRNAs are involved in the host immune response by regulating various immune-related genes [170, 259-263]. In this study, we initially investigated DE lncRNAs associated with *Fp* injection at days 1 and 5 post-challenge. Pairwise comparison between challenged and time- and line-matched PBS-injected animals identified 327 unique lncRNAs with altered expression (fold change ±2 and FDR-corrected p value <0.05) (Supplementary Dataset 2 tabs 1-7, and tab “All Fp vs PBS, non-redundant”).

In order to identify lncRNAs that are broadly involved in the response to infection with *Fp*, we quantified the DE lncRNAs (and their correlated protein-coding genes) that were differentially regulated in all three genetic lines upon infection. On the 5th day of infection, 12 lncRNAs were significantly upregulated (>2-fold) in all three genetic lines (FDR-corrected *p*- value <0.05) (Table 14, top panel). These lncRNAs were most highly upregulated in the susceptible line followed by the control and resistant lines. These finding may indicate that these lncRNAs were either upregulated in response to bacterial load or extent of tissue damage caused by bacterial infection. Surprisingly, none of the lncRNAs was downregulated in all three genetic lines.

Table 14: LncRNAs upregulated in all three genetic lines (>2 fold) on 5th day post Fp challenge and their expression correlation with protein coding genes (top). LncRNAs showing highest fold change (>100-fold) upon Fp infection in at least one genetic line relative to the two other genetic lines and their associated protein coding gene in genome (bottom). Fold change was considered significant if FDR-corrected p value was < 0.05.

| LncRNAs upregulated in all three genetic lines (>2 fold) upon infection and their expression correlation with protein coding genes | | | | | | | |
|---|---|---|---|---|---|---|---|
| LncRNA feature ID | Resistant line | | Control line | | Susceptible line | | |
| | EDGE test: Fold change | FDR p-value correction | EDGE test: Fold change | FDR p-value correction | EDGE test: Fold change | FDR p-value correction | Correlation with ($R^2$) |
| Omy200117486 | 24.36 | 0 | 41.6 | 0 | 91.18 | 0 | Interferon-induced guanylate-binding protein 1 (0.82) |
| Omy100128008 | 14.95 | 0 | 22.23 | 0 | 46.4 | 0 | Complement protein component C7-1 (c7-1) |
| Omy200138656 | 24.3 | 0.000001 | 11.63 | 0.011489 | 28.75 | 0 | Complement C5 (0.66) |
| Omy100149048 | 7.93 | 0.000004 | 5.95 | 0.048727 | 14.15 | 0 | Unknown |
| Omy200107378 | 6.38 | 0.000062 | 11.22 | 0.004248 | 11.22 | 0 | Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (0.92) |
| Omy200165911 | 3.68 | 0.049287 | 4.5 | 0.040404 | 9.98 | 0 | Unknown |
| Omy100052789 | 5.51 | 0.000564 | 5.13 | 0.047151 | 8.65 | 0 | Unknown |
| Omy200107535 | 3.71 | 0.000147 | 6.16 | 0.012707 | 8.12 | 0 | Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (0.92) |
| Omy300025398 | 4.37 | 0.006514 | 5.15 | 0.002475 | 4.86 | 0 | Unknown |
| Omy300085997 | 3.68 | 0.000345 | 3.16 | 0.043867 | 4.02 | 0.001759 | Unknown |
| Omy200206941 | 3.16 | 0.000344 | 2.33 | 0.015415 | 3.44 | 0.000002 | Lysozyme C II precursor (0.83) |
| Omy300043066 | 3 | 0.000121 | 3.74 | 0.006748 | 3.03 | 0.000014 | Properdin (0.82) and complement factor b-like (0.89) |

| LncRNAs showing drastic (>100) fold change upon infection in one particular genetic line and | | | | | |
|---|---|---|---|---|---|
| Feature ID | EDGE test: Fold change | FDR p-value correction | Comparison | Classification of LncRNA | Associated Gene(s) ($R^2$) |
| Upregulated upon infection | | | | | |
| Omy200018785 | 136.06 | 0.001233 | D1_S_FP vs D1_S_PBS | Intergenic | |
| Omy200132807 | 121.83 | 0.000167 | D5_S_FP vs D5_S_PBS | Intergenic | |
| Omy100037031 | 105.28 | 0.01282 | D5_C_FP vs D5_C_PBS | Intergenic | |
| Downregulated upon infection | | | | | |
| Omy200194608 | -168.77 | 0.000001 | D1_S_FP vs D1_S_PBS | Genic, antisense | GSONMG00062425001 si:ch73- protein (0.27) |
| Omy200226560 | -121.9 | 0.001972 | D1_R_Fp vs D1_R_PBS | Genic, antisense | GSONMG00065518001 (fatty-acyl reductase-1) |
| Omy100064313 | -108.56 | 0.001716 | D1_R_Fp vs D1_R_PBS | Intergenic | |

Among DE lncRNAs, 6 lncRNAs showed fold changes >100 fold following *Fp* challenge (Table 14, bottom panel). Five out of six lncRNAs, all three upregulated (Omy200018785, Omy200132807 and Omy100037031) and two downregulated (Omy200226560 and

Omy100064313), exhibited fold change only in one particular ‘genetic line-by-day of infection’ comparison.

### Relationship between differentially expressed lncRNAs and immune-related protein-coding genes

LncRNAs can be classified as genic or intergenic based on their physical location in genome relative to protein coding gene [115]. Classification of all 556 DE lncRNA is given in Supplementary Dataset 1D. Lack of lncRNAs sequence conservation across species [43] makes their annotation difficult. In addition, currently there are no enough literature or database resources for rainbow trout and other salmonids to study lncRNAs’ involvement with the host immune system. Therefore, in an effort to implicate association between DE lncRNAs, identified in this study, and the fish defense system, we followed the following criteria based on our prior knowledge of lncRNAs classification and the genetic lines that we used in this study:

#### Differentially expressed lncRNAs that overlap in position and correlate with expression of immune-related protein-coding genes

Several lncRNAs have been identified that regulate expression of neighboring genes acting in *cis* configuration [150, 271]. Therefore, we searched for DE lncRNAs that were partially or completely overlapping with protein-coding genes in the trout genome. Out of 556 DE lncRNAs, 92 overlapped with protein-coding loci in sense or antisense orientation (Supplementary Dataset 3) [272]. Out of the 92 overlapped genes, 36 genes had hits to KEGG pathways, of them 8 different genes were involved in immunity pathways (such as TNF and mTOR signaling pathways) and 4 genes were associated with microbial diseases (such as *Staphylococcus aureus* infection and viral carcinogenesis). There were 3 genes common in both sets of these pathways which are Complement component 5, Signal transducer and activator of transcription 3 and Cyclic AMP-responsive element-binding protein 5.

In order to identify possible relationships between DE lncRNAs and protein-coding genes that physically overlap with them, we compared their expression patterns across 24 different samples that included different genetic lines and infection statuses. Normalized expressions values of the transcripts used to generate clusters are provided in Supplementary Dataset 4 [273]. The DE lncRNAs and their overlapping protein-coding genes with a strong expression correlation are listed in Table 15. Overall, we identified 13 protein-coding genes that had strong expression correlation ($R^2 \geq 0.70$) with their overlapping lncRNAs and 6 of those protein-coding genes had already described role in immune system. Consistent with this observation, previous studies suggested overlapped genomic localization of immunity associated lncRNAs with protein coding genes of immune system [274].

Table 15: Correlation between expression patterns of lncRNAs and their overlapping protein-coding genes (R2> 0.70).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Protein-coding genes with known immune function or association with microbial infection** | | | | | | | |
| Omy100063056 | 1,263 | GSONMT00040216001 | Intronic | Antisense | Positive (0.73) | Interferon-induced guanylate-binding protein 1-like | [233, 234] |
| Omy200083892 | 1,294 | GSONMT00050654001 | Intronic | Antisense | Positive (0.84) | Tumor necrosis factor receptor superfamily member 9-like (Tnfrsf9) | [200] |
| Omy200080884 | 1,512 | GSONMT00034829001 | Exonic | Antisense | Positive (0.93) | Response gene to complement 32 protein (rgc32) | [235] |
| Omy200206941 | 537 | GSONMT00021084001 | Intronic | Unknown | Positive (0.83) | Lysozyme C II precursor | [216] |
| Omy200107012 | 885 | GSONMT00019341001 | Intronic | Unknown | Positive (0.89) | Stromal interaction molecule 2-like | [236] |
| Omy100228715 | 297 | GSONMT00079494001 | Exonic | Unknown | Positive (0.83) | Unnamed protein product/transcobalamin-1 like | [237] |
| **Protein-coding genes with no previously described immunity function** | | | | | | | |
| Omy400008156 | 668 | GSONMT00041383001 | Intronic | Unknown | Positive (0.87) | Reticulon-2 like | |
| Omy300038945 | 596 | GSONMT00049537001 | Intronic | Antisense | Positive (0.81) | Cytochrome P450 7B1 | |
| Omy400006181 | 248 | GSONMT00049631001 | Intronic | Unknown | Positive (0.77) | Collagen alpha-1(IX) chain-like | |
| Omy400003716 | 725 | GSONMT00061535001 | Intronic | Sense | Positive (0.87) | Protocadherin 8 (pcdh8) | |
| Omy100224015 | 683 | GSONMT00065518001 | Intronic | Antisense | Positive (0.71) | Fatty acyl-CoA reductase 1 (facr1) | |
| Omy200181316 | 604 | GSONMT00071779001 | Exonic | Unknown | Positive (0.82) | Muscular LMNA-interacting protein | |
| Omy100171980 | 1,292 | GSONMT00073108001 | Exonic | Unknown | Positive (0.82) | Immunoglobulin-like and fibronectin type III domain-containing protein 1 | |

Some lncRNAs showed interesting correlated expression pattern with immune-related protein coding genes post Fp challenge and were selected for the following further discussion:

LncRNA Omy100063056 partially overlapped with intron 6 of interferon induced guanylate binding protein-1 like (gbp1) (GSONMT00040216001) in antisense orientation and their expression pattern was positively correlated (R2=0.80) (Figure 19, A-C). RPKM (reads per kilobase per million) count showed that both Omy100063056 and gbp1 gene transcript were upregulated on day 1 and 5 post-challenge. Upregulation on day 5 was greater in the susceptible line relative to control and resistant lines. GPB1 gene transcript also shows correlated expression with lncRNA in human [275]. Previous reports suggested that gbp1 is one of the differentially regulated immune response genes against microbial pathogens in salmon and trout [257, 276].

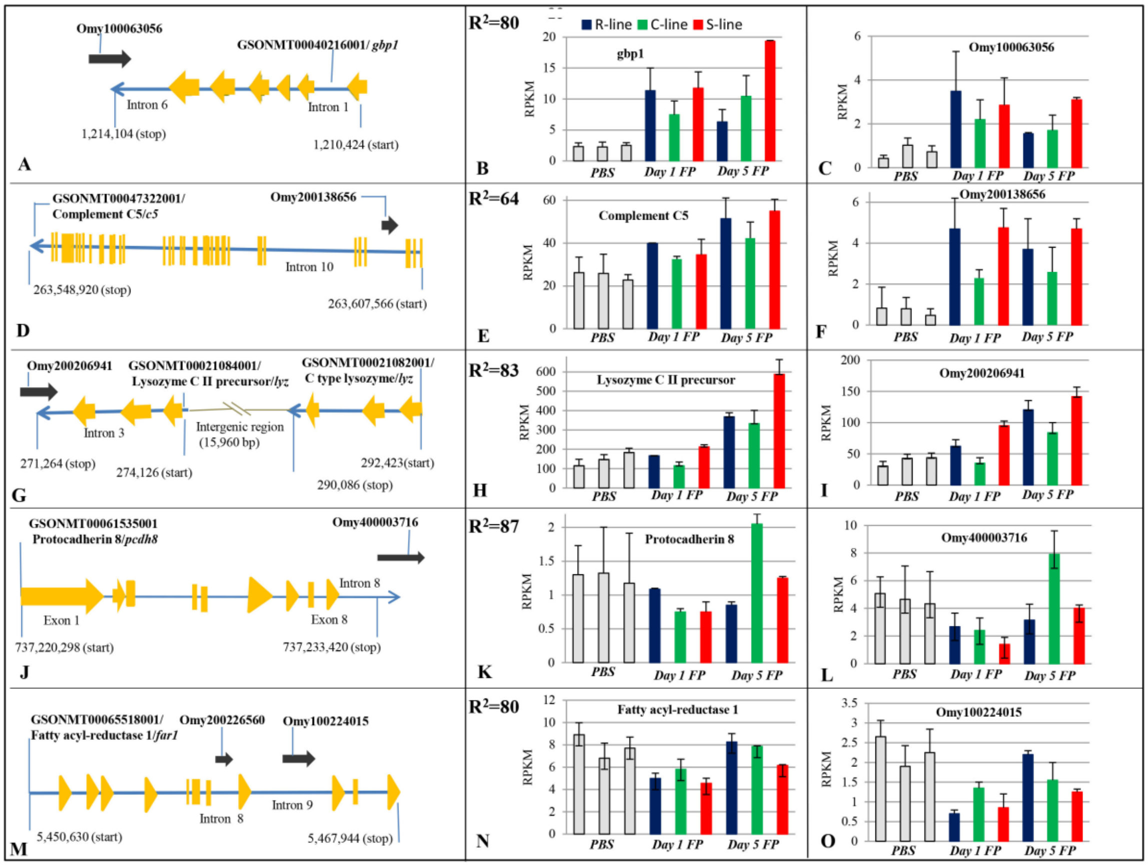


Figure 19: Genomic location of selected differentially expressed lncRNAs relative to protein-coding genes with immune-related functions and their expression patterns among PBS injected and day 1 and day 5 post-*Fp* challenged fish of different genetic lines. Omy100063056 is partially overlapped with intron 6 of *gbp1* in antisense orientation (A) and their expression is positively correlated ($R^2$=0.80) (B and C). Omy2001386656 is within intron of complement C5 in antisense orientation (D) and they show correlated expression pattern between PBS and *Fp* injected fish ($R^2$=0.64) (E and F). Omy200206941 partially overlaps with intron of lysozyme CII precursor in antisense orientation (G) and shows correlated expression pattern with the lysozyme CII precursor ($R^2$=0.83) (H and I). Omy400003716 partially overlaps with intron of protocadherin 8 in sense orientation (J) and shows strong positive expression correlation with the protocadherin 8 ($R^2$=0.87) (K and L). Fatty acyl-reductase 1 has one sense lncRNA in each intron 8 (Omy200226560) and 9 (Omy100224015) (M) and shows positive expression correlation with both the lncRNAs. Expression pattern of fatty acyl reductase 1 and Omy100224015 is given in figure (N and O).

LncRNA Omy200128656 was located in intron 11 of complement C5 (c5) (GSONMT00047322001) gene in antisense orientation and their expression was positively correlated (R2=0.64) (Figure 19, D-F). Expression of c5 gene transcript was increased by day 5 post-infection and expression of Omy200128656 was upregulated on days 1 and 5 post-challenge. In human, lncRNA C5T1lncRNA, located in 3'UTR of the C5, showed upregulated expression upon immune stimulation and its knockdown showed corresponding decrease in transcript level of C5 mRNA [277]. However, unlike in human, lncRNA Omy200128656 in trout is located in intron 11 of the c5 gene.

LncRNA Omy200206941 was partially overlapped with intron 4 of lysozyme CII precursor (lyz) (GSONMT00021084001) gene in antisense orientation and the expression was positively correlated (R2=0.83) (Figure 19, G-I). Its expression was also positively correlated with another C type lysozyme (lyz) (GSONMT00021082001) gene transcript located about 18 kb away in the same chromosome (R2=0.88). All these three transcripts showed upregulation on day 5 post-challenge. Consistent with this upregulated expression post challenge, it has been established that C type lysozyme is an important component of innate immune system in salmonid fish [278] . In addition, a neighboring antisense non-coding RNA, LINoCR, is involved in induction of lysozyme locus upon lipopolysaccharide stimulation in chicken [279].

LncRNA Omy400003716 partially overlapped with intron 8 of protocadherin 8 (pcdh8) (GSONMT00061535001) in sense orientation and the expression was highly positively correlated (R2=0.87) (Figure 19, J-L). RPKM count between PBS and Fp challenged fish showed that both Omy400003716 and pcdh8 gene transcript were downregulated in day 1 post infection relative to naïve and day 5 post-challenged fish.

Two lncRNAs Omy200226560 and Omy100224015 were in intron 8 and 9 of fatty acyl-reductase 1 (far1) (GSONMT00065518001) respectively and they positively correlated with the far1 gene transcript with correlation coefficient (R2) of 0.36 and 0.80 respectively (Figure 19, M-O). These three transcripts showed downregulation on day 1, post challenge relative to PBS injected, and day 5, post-Fp challenged fish.

Strand orientation of Omy200138656, Omy200206941 and Omy300084989 lncRNAs transcripts were confirmed by strand specific PCR relative to their counterpart protein coding loci (Supplementary Dataset 1E).

### Differentially expressed lncRNAs that neighbor and correlate with expression of immune-related protein-coding genes

Out of 556 DE lncRNAs, 464 were intergenic without overlap with protein-coding loci in the trout genome. In order to identify the immune-relevant protein-coding genes that were clustered around DE lncRNAs in the genome, we chose protein-coding genes within a 50 kb distance on both sides of DE lncRNAs and performed KEGG pathway analysis of the neighboring protein-coding genes [280]. Out of 464 DE intergenic lncRNAs, 371 had protein-coding genes within 50 kb distance in the genome. A total of 290 genes neighboring to DE lncRNAs had hits to KEGG pathways, of them 51 different genes were related to immunity pathways, 49 genes were involved in microbial infection processes and 28 genes were common in both sets of these pathways (Supplementary Dataset 5) [281].

In the immune system category, most of the KEGG hits were involved in chemokine signaling, platelet activation, complement system, TNF signaling, T cell receptor signaling, Fc gamma R-mediated phagocytosis, Toll-like receptor signaling, phagosome, cytokine-cytokine receptor interaction, NOD-like receptor signaling, leukocyte trans-endothelial migration and others (Supplementary Dataset 5) [281]. Similarly, in the microbial pathogenesis category, hits were involved in the pathogenesis of various viral, bacterial and protozoal infections like tuberculosis, influenza A, herpes simplex infection, amoebiasis, bacterial invasion of epithelial cell, and other microbial infections. Interestingly, almost half of the hits to immune system were involved in signal transduction pathways. Among the neighboring protein-coding genes, expression patterns of 9 were highly positively correlated with that of lncRNA ($R^2 \geq 0.70$) (Table 16). About half of the protein-coding genes with high correlation in expression patterns with their neighboring lncRNAs were from components of immune system like suppressor of cytokine signaling 3 (SOCS3), complement factor D, ninjurin-1 and ceramide-1 phosphate transfer protein. Previous studies also indicated that many immune relevant lncRNAs are in 5’ or 3’ close proximity of neighboring protein-coding genes [261, 274].

Table 16: Correlation between expression patterns of lncRNAs and their intergenic neighboring protein-coding genes (within <50 kb and $R^2$> 0.70). References are provided for some of the protein-coding genes with previously described functions in immunity or association with microbial infection/pathogenesis.

| LncRNA | Size | Neighboring protein-coding genes (ID) | Distance from LncRNA (KB) | Direction relative to LncRNA | Expression correlation type ($R^2$) | Annotation of coding gene | Reference to immune or pathogenesis function |
|---|---|---|---|---|---|---|---|
| **Protein-coding genes with known immune function or association with microbial infection** | | | | | | | |
| Omy200174653 | 519 | GSONMT00031633001 | 5 | Unknown/ Intergenic | Positive (0.92) | Complement factor D-like | [238] |
| Omy300084989 | 596 | GSONMT00013116001 | 2.6 | Antisense /Intergenic | Positive (0.71) | Suppressor of cytokine signaling | [239] |
| Omy300074800 | 493 | GSONMT00003195001 | 1.1 | Unknown/ Intergenic | Positive (0.79) | Ninjurin-1 | [240] |
| Omy200206941 | 537 | GSONMT00021082001 | 18.6 | Unknown/ Intergenic | Positive (0.88) | C type lysozyme | [216] |
| Omy200073559 | 2,093 | GSONMT00017721001 | 3.5 | Unknown/ Intergenic | Positive (0.77) | Ceramide-1-phosphate transfer protein-like | [241] |
| **Protein-coding genes with no previously described immunity function** | | | | | | | |
| Omy200061208 | 1057 | GSONMT00041695001 | 0.3 | Unknown/ Intergenic | Positive (0.90) | Coiled-coil transcriptional | |
| Omy200112536 | 1,059 | GSONMT00001821001 | 18.2 | Unknown/ Intergenic | Positive (0.88) | Serum albumin 1 | |
| Omy300087476 | 619 | GSONMT00010387001 | 0.9 | Unknown/ Intergenic | Positive (0.83) | Neutral amino acid transporter B(0) | |
| Omy200075445 | 745 | GSONMT00008107001 | 1.3 | Unknown/ Intergenic | Positive (0.70) | Hepatocyte nuclear factor 4-beta-like | |

### Differentially expressed lncRNAs that correlate with expression of immune-related protein-coding genes

LncRNAs have ability to work in *cis* as well as in *trans* configuration [162-164] and can regulate protein-coding genes that are distant in position on the same or different chromosome. In order to identify possible expression correlation of lncRNAs with such protein coding genes, we performed clustering of DE lncRNAs and protein-coding genes based on their expression pattern across 24 samples. This clustering identified several protein-coding genes with correlated expression with DE lncRNAs that were distantly located in the genome (Table 17). Most of the proteins in these clusters were related to the innate immune system, mainly the complement system, cytokines and chemokines, and receptors and transcription factors of the innate immune system signal transduction pathways. The list included chemokine CK1, NF-kappa B inhibitor alpha, c-c motif chemokine 19, and several proteins of the complements system such as factor B, properdin, component C7 and C4b-binding protein alpha (Table 17).

Table 17: Correlation between expression patterns of lncRNAs and some distantly located (>50 kb or different chromosome) immune-relevant protein-coding genes. References are provided for some of the protein-coding genes with previously described functions in immunity or association with microbial infection/pathogenesis.

| **LncRNA** | **Size** | **Protein-coding genes (ID)** | **Expression correlation type ($R^2$)** | **Annotation of coding gene** | **Reference to immunity or pathogenesis function** |
|---|---|---|---|---|---|
| Omy100104455 | 587 | GSONMT00024124001 | Positive (0.96) | Chemokine CK1 | [242] |
| Omy200174653 | 357 | GSONMT00051250001 | Positive (0.92) | C-C motif chemokine 19 precursor | [243] |
| Omy300084989 | 596 | GSONMT00062775001 | Positive (0.83) | C4b-binding protein alpha chain precursor | [244] |
| Omy300041057 | 448 | GSONMT00042009001 | Positive (0.80) | Caspase-8 | [245] |
| Omy300043066 | 715 | GSONMT00001792001 | Positive (0.82) | Properdin | [246] |
| Omy300043066 | 715 | GSONMT00027840001 | Positive (0.89) | Complement factor b-like | [247] |
| Omy200100893 | 357 | GSONMT00051250001 | Positive (0.92) | C-C motif chemokine 19 precursor | [243] |
| Omy200107378 | 522 | GSONMT00016681001 | Positive (0.92) | Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 | [229] |
| Omy200107535 | 948 | GSONMT00016681001 | Positive (0.92) | Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 | [229] |
| Omy200117486 | 529 | GSONMT00005714001 | Positive (0.82) | Interferon-induced guanylate-binding protein 1 | [233, 234] |
| Omy100066751 | 326 | GSONMT00080410001 | Positive (0.84) | Tumor necrosis factor, alpha-induced protein 2 (tnfaip2) | [200] |
| Omy100128008 | 1,232 | GSONMT00070499001 | Positive (0.82) | Complement protein component C7-1 (c7-1) | [248] |
| Omy100063056 | 1,263 | GSONMT00075049001 | Positive (0.85) | Tumor necrosis factor, alpha-induced protein 3 (tnfaip3) | [200] |
| Omy200053140 | 1,510 | GSONMT00071335001 | Negative (0.84) | NF-kappa-B inhibitor alpha | [200] |

### Differentially expressed lncRNAs that correlate with expression of several immune-related protein coding genes

Clustering of DE lncRNAs with protein coding genes based on their expression value identified several protein-coding genes of the immune system correlated with one lncRNA. As an example, lncRNA Omy200107378 was upregulated post *Fp* challenge and its expression was strongly positively correlated with six different protein coding genes, some of which have already established function in immune system ($R^2$>0.98) (Figure 20). Similarly, expression of Omy100124197 was strongly correlated with 8 different proteins including matrix metallo-proteinase (Astacin) (GSONMT00014156001), elastase-1 (GSONMT00002714001), nattectin (GSONMT00024075001), phospholipase-A2 (GSONMT00073599001), and syncollin (GSONMT00034810001) ($R^2$>0.98) (Figure 20). Role of these correlated proteins in the immune system has already been characterized in different species [282-288]. Several studies have also reported correlated expression of several immune related protein-coding genes with a single lncRNA [259].

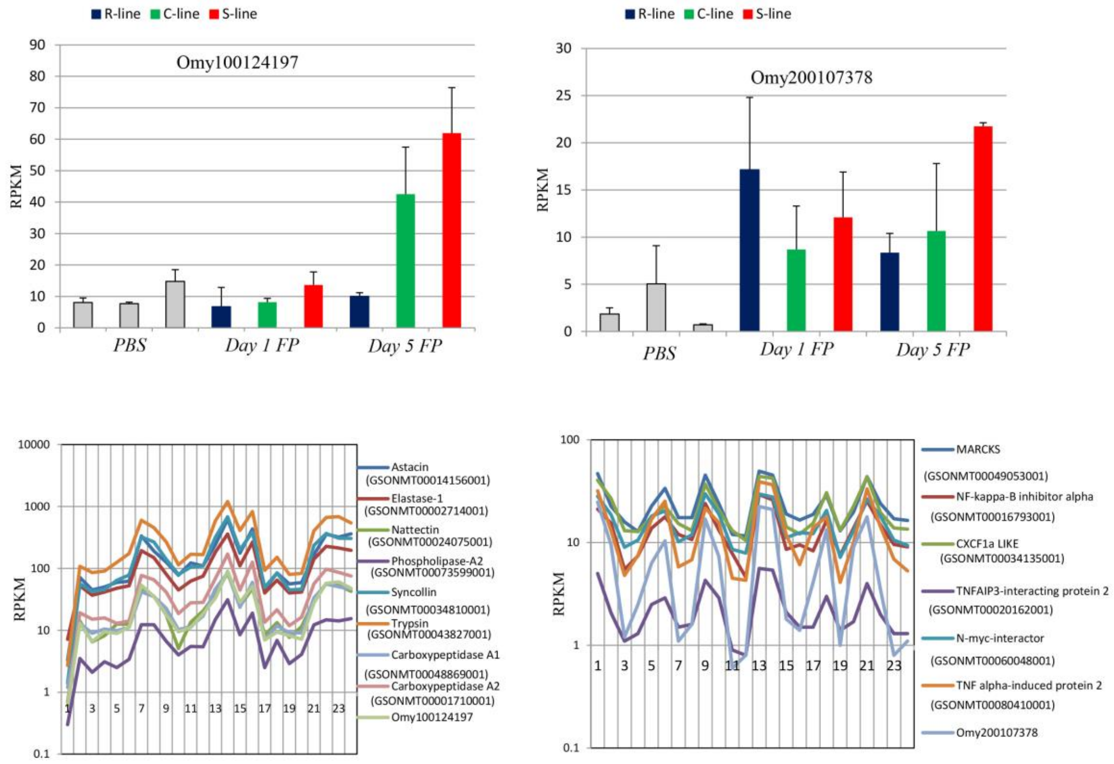


Figure 20: Top two bar graphs show expression patterns of lncRNAs Omy100124197 and Omy200107378 among PBS injected, and day 1 and day 5 post-*Fp* challenged fish in three genetic lines. Respective bottom expression line graphs show expression level of these lncRNAs with different protein-coding genes across 24 samples consisting of different genetic lines and infection statuses. Expression clusters were generated by the Multi-experiment Viewer (MeV) program using a cut off $R^2$ minimum of 0.98.

## LncRNAs expression of naïve fish in different genetic lines

Three genetic lines of rainbow trout used in this study had significant differences in infection susceptibility to *Fp* as a result of selective breeding [257]. To investigate differences in transcription between lines, we quantified the DE lncRNAs among genetic lines on day 1 following PBS injection. Pairwise comparison identified 32 DE lncRNAs among different genetic lines. Two lncRNAs were DE between the resistant and control lines, 6 lncRNAs between control and susceptible lines, and 24 lncRNAs were DE between resistant and susceptible lines (Supplementary Dataset 2) [269]. In our previous study, we identified differences in transcriptome abundance of protein-coding genes among naïve genetic lines [257]. The numbers of DE lncRNAs were smaller but consistent with the numbers of DE protein-coding genes among different naïve genetic lines (Table 13). Expression analysis identified an interesting pattern of transcriptome differences among genetic lines, which correlated with infection susceptibility. LncRNAs Omy200019549, Omy200132559, Omy200160814, Omy200075485 and Omy300048239 were most highly expressed in the resistant line, followed by control and susceptible lines. In contrast, Omy300052204, Omy200142923, Omy200118054 and Omy200165975 were upregulated in the susceptible line relative to the resistant and control lines (Figure 21). These DE lncRNAs between genetic lines may contribute to differences in infection susceptibility among genetic lines. In consistent with our findings, genetic variation in lncRNAs was shown to be associated with human disease resistance/susceptibility [289, 290].

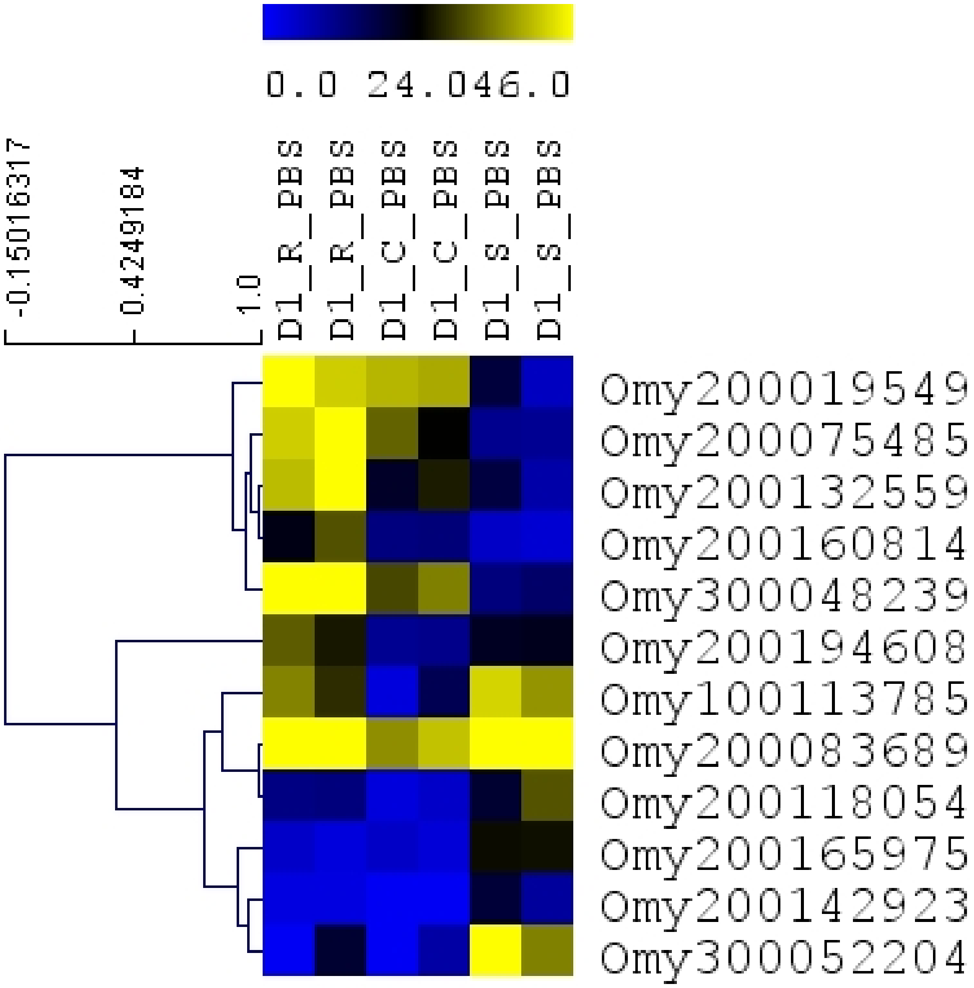


Figure 21: Comparison of transcriptome abundance of selected lncRNAs among naïve fish in all genetic lines. Genes are hierarchically clustered based on their expression pattern. D1 indicates day 1 post challenge and PBS indicates PBS injection. C, R and S represent control, resistant and susceptible genetic lines of the fish.

**Difference in transcriptome abundance of lncRNAs among genetic lines after infection**

Induction and activation of adaptive and some of the innate immune components requires pathogen entry into the host suggesting that basal naïve transcriptome level may not be sufficient enough to explain the differences in the ability of the control, susceptible, and resistant fish lines to clear the pathogen. Therefore, we reasoned that, in addition to differences in naïve lncRNA abundance, the genetic lines had altered ability to express immune-relevant transcripts following pathogen challenge. To investigate this point, we quantified DE lncRNAs among genetic lines on days 1 and 5 following *Fp* infection. Pairwise comparison identified 149 DE lncRNAs between genetic lines combined from the 1st and 5th days of infection (Table 13 and Supplementary Dataset 2). On 5th day of infection, there were 83 lncRNAs DE between resistant and susceptible lines; 21 lncRNAs between resistant and control lines, and 5 lncRNAs between control and susceptible lines. On 1st day of infection, these numbers were 15, 12 and 13 respectively (Supplementary Dataset 2). Similarly, on the 1st day of infection majority of the lncRNAs were upregulated on susceptible line relative to two other genetic lines. The expression number of DE's correlate with the gradient of bacterial load between the three genetic lines: S>C>R. Previous report also indicated correlation of lncRNAs expression with microbial load [291]. Figure 22 shows abundance of selected hierarchically clustered lncRNAs among genetic lines after infection with *Fp*. On the 5th day of infection, most of the lncRNAs were upregulated in the susceptible line compared to control and resistant lines, with only Omy200112846, Omy200075161, Omy200194608 and Omy100199114 exhibiting opposite trend in expression level (Figure 22).

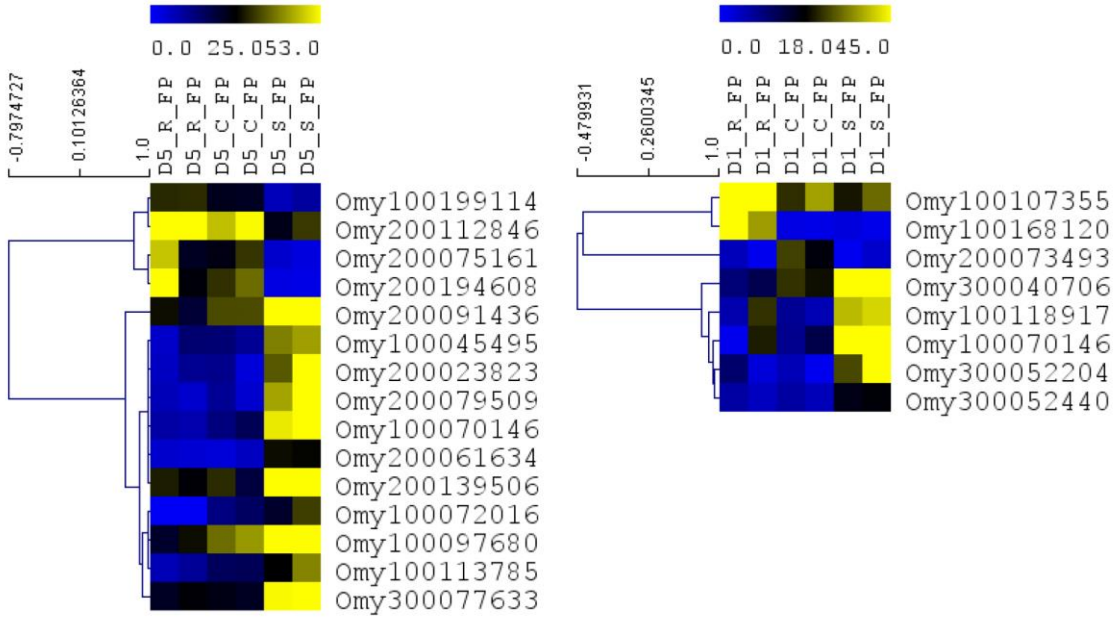


Figure 22: Comparison of transcriptome abundance of selected lncRNAs among genetic lines after infection with *Fp.* Genes are hierarchically clustered based on their expression pattern. D1 and D5 indicate day 1 and day 5 of sampling after injection. *Fp* indicates *Fp* injection. C, R and S represent control, resistant and susceptible genetic lines of the fish.

**LncRNA transcriptome change as the disease progress from day 1 to day 5**

During the course of infection, the host can utilize different immune components at different stages of disease, which requires change in expression of immune-relevant genes. We reasoned that if lncRNAs regulate the immune system, their transcriptome changes, like that of protein-coding genes, would change as the disease progresses. Pairwise comparison between day 1 and day 5 post-*Fp* challenge identified 137 lncRNAs whose expression was significantly changed during two time points (Supplementary Dataset 2). This finding is consistent with previous report demonstrating change in the number of differentially regulated lncRNAs at different ISAV infection time points in Atlantic salmon [265]. Figure 23 shows abundance of selected hierarchically clustered lncRNAs between day 1 and day 5 of *Fp* injection in each genetic line. As expected, some of the lncRNAs that showed altered expression between day 1 and day 5 post-challenges had strong expression correlation with immune relevant protein coding genes. LncRNAs Omy200174653 had altered expression on day 5 relative to day 1 post challenge in susceptible lines and a strong positive correlation with complement factor D (Table 16). Similarly, Omy100066751 and Omy200107535 exhibited a strong positive expression correlation with tumor necrosis factor alpha-induced protein 2 (tnfaip2) and nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (NFKB2) ($R^2$=0.92), respectively (Table 17). NFKB2 is a transcription factor required to maintain normal level of antigen specific antibody production in response to antigen challenge [292]. It is noteworthy that Omy200107535 was one of the 12 lncRNAs that were upregulated on day 5 post challenge relative to naïve fish in all three genetic lines (Table 14). This change in expression pattern of lncRNAs during the course of infection suggests that these lncRNAs may play a role in adjustment of immunity depending on severity and stage of the disease. In addition, these DE lncRNAs might play a role in host pathogen interaction or pathogen life cycle during the course of infection as suggested in previous studies [293].

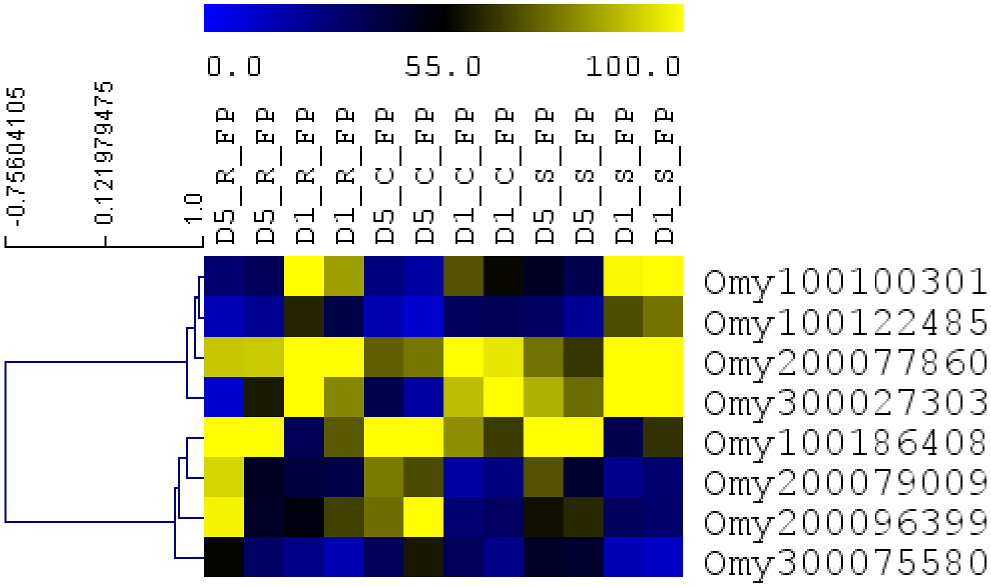


Figure 23: Comparison of transcriptome abundance of selected lncRNAs between day 1 and day 5 of *Fp* injection in each genetic line. Genes were hierarchically clustered based on their expression pattern. D1 and D5 indicate day 1 and day 5 of sampling after injection and Fp indicates *Fp* injection. C, R and S represent control, resistant and susceptible genetic lines of the fish.

### Sequence homology with lncRNAs in Atlantic salmon

Recently differentially regulated lncRNAs in response to infectious salmon anemia virus (ISAV) has been characterized in Atlantic salmon [265]. Out of 556 DE lncRNA in trout genetic lines in various comparisons, 23 showed significant sequence homology with Atlantic salmon lncRNAs that were associated with ISAV infection (query cover > 50%, sequence identity > 90% and E value < 1e-10) (Supplementary dataset 6) [294]. Interestingly, out of 23 conserved lncRNA, 17 showed regulated expression in *Fp* injected fish relative to PBS injected naïve animals; and remaining 6 were differently regulated between genetic lines and time points of infection comparison (Supplementary dataset 2). It is worth mentioning that one of the conserved lncRNA, Omy300043066 had strong positive expression correlation with properdin and complement factor b like protein in trout (Table 17) and was one of the 12 lncRNAs that were upregulated during infection in all three genetic lines relative to their PBS injected fish (Table 14). All of the 23-conserved lncRNA were regulated in salmon in response to ISAV, indicating potential role in general immunity rather than being bacterial or virus specific.

### Novel lncRNAs in resistant and susceptible genetic lines

Novel lncRNAs were detected in each genetic line separately by running sequence reads through our previously described lncRNA discovery pipeline [115]. 589 susceptible-specific and 631 resistant-specific novel lncRNAs were predicted. FASTA files are available at http://www.animalgenome.org/repository/pub/MTSU2015.1014/. Correlation analyses of gene expression showed only 9 lncRNAs in moderate correlation ($R^2 \geq 0.70$) with protein coding genes. However, none of these proteins was overlapped with lncRNA or had previously described role in immune system (Supplementary dataset 7) [295]. While identification of these lncRNAs were limited to each genetic line, their multiple group ANOVA analysis of gene expression (genetic line X infection status X time point) showed a complex expression pattern (Supplementary dataset7-PCA). Interestingly, two lncRNA (dis_R_00048342 and dis_R_00050098) showed resistant-line specific gene expression regardless of the infection status or the time points. Similarly, three lncRNA (dis_S_00030301, dis_S_00043616 and dis_S_00083595) were susceptible-line specific (Supplementary dataset 7). On the other hand, 20 lncRNAs showed explicit expression after *Fp* infection, regardless of the time of infection or the genetic line. In addition, three lncRNA showed explicit expression between day1 and day5 of infection (Supplementary dataset 7). These finding may suggests that genomic selection for BCWD over three generations may have introduced novel genomic variations or genomic reorganization of some lncRNA loci and altered expressions of lncRNAs.

## Conclusion

Thus far, studies on host response to microbial infection in salmonids have given significant attention to changes in protein-coding gene expression. However, lncRNAs have emerged as key regulators of host defense against a wide variety of pathological processes including microbial infection [170, 258-263, 265]. Manipulation of individual lncRNAs is sufficient to change the expression of hundreds of immune response genes [259], and variation in expression of other lncRNA's alter host susceptibility to different microbial pathogens [263]. In the present study, we quantified DE lncRNAs in response to *Fp* infection, which is an important cause of morbidity and mortality in salmon and trout [250]. This study is novel as we characterized the expression signature of lncRNAs on a genome-wide scale in response to one of the major bacterial infection of a salmonid fish. To our knowledge, regulation of lncRNA during bacterial pathogen challenge has not previously been studied in any aquaculture/fish species.

Using transcriptome-wide datasets of protein-coding genes and lncRNAs across 24 samples, we were able to identify potential immune-relevant and other protein-coding genes correlating with DE lncRNAs. This study identified correlation between the genomic physical proximity and coordinated expression of a large number of immune related and other protein coding genes with that of lncRNAs during BCWD in rainbow trout. In this study, most of the DE lncRNAs (sense and antisense) had significant positive expression correlation ($R^2$> 0.70) with their overlapped and/or neighboring protein coding genes. These results are consistent with human ENCODE project results that showed particularly striking positive correlation of lncRNAs with the expression of antisense coding genes [43]. In trans-acting lncRNAs, the ENCODE project observed that lncRNAs are more positively than negatively correlated with protein-coding genes, a finding consistent with our observation of more frequent positive than negative correlation with distantly located protein coding genes. The positive correlation between lncRNA and protein coding genes suggest potential for co-expression [151].

This study has characterized DE lncRNAs in response an initial phase of BCWD (day 1 and 5 post-challenge) and has explored expression correlation of lncRNAs with immune relevant protein coding gene that may play crucial role in pathogenesis or immunity during the early phase of the disease in rainbow trout. Further mechanistic study of the underlying biological relationship between correlated DE lncRNAs and proteins of innate immune system will help understand regulation of pathogenesis/ immunity at this crucial phase of infection in juvenile rainbow trout.

**Acknowledgments**

We acknowledge technical contributions from Travis Moreland and Joel Caren. This work was supported by Agricultural Research Service CRIS Project 1930-32000-005 “Integrated Approaches for Improving Aquatic Animal Health in Cool and Cold Water Aquaculture,” CRIS Project 1930-32000-009 “Utilization of Genomics for Improving Production Traits in Cool and Cold Water Aquaculture” and Agriculture and Food Research Initiative competitive grants no. 2012-67015-30217 (to Gregory D. Wiens) and No. 2014-67015-

21602 (to Mohamed Salem) from the USDA National Institute of Food and Agriculture. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US, Department of Agriculture.  USDA is an equal opportunity employer.

# CHAPTER VI: SUMMARY OF THE DISSERTATION

This dissertation work was mainly aimed at improving genome annotation, identifying long non-coding RNA repertoire and identifying coding/functional single nucleotide polymorphisms (cSNPs) associated with growth, disease resistance and muscle fillet quality traits in rainbow trout. Using deep sequencing RNA-seq approach, we were able to identify ~14,800 protein coding genes missing in the rainbow trout draft reference genome, including 710 full length sequence. These additional sequences will help annotate the genome reference towards its completion. In addition, we identified long non coding RNA repertoire in trout for the first time that will provide important genomic resources for functional genomics research in future. LncRNA database used to identify differentially expressed lncRNAs between genetic lines of naive animals and in response to infection with *Flavobacterium psychrophilum*, a causative agent of Bacterial Cold Water Disease. In addition, we identified the lncRNAs genomic co-localization relative to immune-relevant protein-coding genes, and explored their co-expression relationships to suggest possible regulation of immune-relevant protein-coding genes by lncRNAs.

This study also utilized combination of SNP calling algorithms in RNA-Seq data from large population of selectively bred trout population to identify cSNPs to design a cSNP array for functional genomics research. Functional validation of a subset of the cSNPs used in cSNP array design showed high accuracy of the algorithms in cSNP identification. This study will provide valuable genomic information for research scientists and industries to conduct functional genomics research to improve production traits in commercially important rainbow trout.

# REFERENCES

1. Losos JB, Arnold SJ, Bejerano G, Brodie ED, Hibbett D, Hoekstra HE, Mindell DP, Monteiro A, Moritz C, Orr HA *et al*: **Evolutionary biology for the 21st century**. *PLoS Biol* 2013, **11**(1):e1001466.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
3. Illumina: **Sequencing Cost and Data Output Since 2000**. In. Edited by Technology AItNGS: Illumina; 2015.
4. White RR, Brady M, Capper JL, McNamara JP, Johnson KA: **Cow-calf reproductive, genetic, and nutritional management to improve the sustainability of whole beef production systems**. *J Anim Sci* 2015, **93**(6):3197-3211.
5. Reimundo P, Rivas AJ, Osorio CR, Méndez J, Pérez-Pascual D, Navais R, Gómez E, Sotelo M, Lemos ML, Guijarro JA: **Application of suppressive subtractive hybridization to the identification of genetic differences between two Lactococcus garvieae strains showing distinct differences in virulence for rainbow trout and mouse**. *Microbiology* 2011, **157**(Pt 7):2106-2119.
6. McParland S, Kennedy E, Lewis E, Moore SG, McCarthy B, O'Donovan M, Berry DP: **Genetic parameters of dairy cow energy intake and body energy status predicted using mid-infrared spectrometry of milk**. *J Dairy Sci* 2015, **98**(2):1310-1320.
7. Boligon AA, Carvalheiro R, Albuquerque LG: **Evaluation of mature cow weight: genetic correlations with traits used in selection indices, correlated responses, and genetic trends in Nelore cattle**. *J Anim Sci* 2013, **91**(1):20-28.
8. Yang ZQ, Qing Y, Zhu Q, Zhao XL, Wang Y, Li DY, Liu YP, Yin HD: **Genetic effects of polymorphisms in myogenic regulatory factors on chicken muscle fiber traits**. *Asian-Australas J Anim Sci* 2015, **28**(6):782-787.
9. Wang Y, Lupiani B, Reddy SM, Lamont SJ, Zhou H: **RNA-seq analysis revealed novel genes and signaling pathway associated with disease resistance to avian influenza virus infection in chickens**. *Poult Sci* 2014, **93**(2):485-493.
10. Smith J, Sadeyen JR, Butter C, Kaiser P, Burt DW: **Analysis of the early immune response to infection by infectious bursal disease virus in chickens differing in their resistance to the disease**. *J Virol* 2015, **89**(5):2469-2482.
11. Singh RP, Hodson DP, Jin Y, Lagudah ES, Ayliffe MA, Bhavani S, Rouse MN, Pretorius ZA, Szabo LJ, Huerta-Espino J *et al*: **Emergence and Spread of New Races of Wheat Stem Rust Fungus: Continued Threat to Food Security and Prospects of Genetic Control**. *Phytopathology* 2015:PHYTO01150030FI.
12. Novoselskaya-Dragovich AY, Bespalova LA, Shishkina AA, Melnik VA, Upelniek VP, Fisenko AV, Dedova LV, Kudryavtsev AM: **[Genetic diversity of common wheat varieties at the gliadin-coding loci]**. *Genetika* 2015, **51**(3):324-333.
13. Ma J, Zhang CY, Yan GJ, Liu CJ: **Improving yield and quality traits of durum wheat by introgressing chromosome segments from hexaploid wheat**. *Genet Mol Res* 2013, **12**(4):6120-6129.

14. Tuncel A, Okita TW: **Improving starch yield in cereals by over-expression of ADPglucose pyrophosphorylase: expectations and unanticipated outcomes**. *Plant Sci* 2013, **211**:52-60.
15. Parry MA, Hawkesford MJ: **Food security: increasing yield and improving resource use efficiency**. *Proc Nutr Soc* 2010, **69**(4):592-600.
16. Du Z, Che M, Li G, Chen J, Quan W, Guo Y, Wang Z, Ren J, Zhang H, Zhang Z: **A QTL with major effect on reducing leaf rust severity on the short arm of chromosome 1A of wheat detected across different genetic backgrounds and diverse environments**. *Theor Appl Genet* 2015, **128**(8):1579-1594.
17. FAO: **STATE OF WORLD FISHERIES AND AQUACULTURE (AR.ED.)**. In. Roma, Italy: FAO.
18. FAO, 2014: **The State of World Fisheries and Aquaculture 2014. Rome. 223 pp.** In.; 2014.
19. Davidson WS: **Adaptation genomics: next generation sequencing reveals a shared haplotype for rapid early development in geographically and genetically distant populations of rainbow trout**. *Mol Ecol* 2012, **21**(2):219-222.
20. Speare D, Arsenault G, Buote M: **Evaluation of Rainbow Trout as a Model for use in Studies on Pathogenesis of the Branchial Microsporidian Loma salmonae**. *Contemporary topics in laboratory animal science / American Association for Laboratory Animal Science* 1998, **37**(2):55-58.
21. Giaquinto PC, Hara TJ: **Discrimination of bile acids by the rainbow trout olfactory system: evidence as potential pheromone**. *Biological research* 2008, **41**(1):33-42.
22. Patel M, Rogers JT, Pane EF, Wood CM: **Renal responses to acute lead waterborne exposure in the freshwater rainbow trout (Oncorhynchus mykiss)**. *Aquat Toxicol* 2006, **80**(4):362-371.
23. Thorgaard GH, Bailey GS, Williams D, Buhler DR, Kaattari SL, Ristow SS, Hansen JD, Winton JR, Bartholomew JL, Nagler JJ *et al*: **Status and opportunities for genomics research with rainbow trout**. *Comp Biochem Physiol B Biochem Mol Biol* 2002, **133**(4):609-646.
24. Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation**. *Nat Rev Genet* 2012, **13**(5):329-342.
25. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Labadie K, Alberti A *et al*: **The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates**. *Nature communications* 2014, **5**:3657.
26. Ng SH, Artieri CG, Bosdet IE, Chiu R, Danzmann RG, Davidson WS, Ferguson MM, Fjell CD, Hoyheim B, Jones SJ *et al*: **A physical map of the genome of Atlantic salmon, Salmo salar**. *Genomics* 2005, **86**(4):396-404.
27. Young WP, Wheeler PA, Coryell VH, Keim P, Thorgaard GH: **A detailed linkage map of rainbow trout produced using doubled haploids**. *Genetics* 1998, **148**(2):839-850.
28. Brent MR: **Genome annotation past, present, and future: how to define an ORF at each locus**. *Genome Res* 2005, **15**(12):1777-1786.
29. Jeffries AC, Saunders NJ, Hood DW: **Genome sequencing and annotation**. *Methods Mol Med* 2001, **67**:215-230.
30. Rouzé P, Pavy N, Rombauts S: **Genome annotation: which tools do we have for it?** *Curr Opin Plant Biol* 1999, **2**(2):90-95.
31. Stein L: **Genome annotation: from sequence to biology**. *Nat Rev Genet* 2001, **2**(7):493-503.

32. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al*: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nat Protoc* 2013, **8**(8):1494-1512.

33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

34. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.

35. Clark MB, Choudhary A, Smith MA, Taft RJ, Mattick JS: **The dark matter rises: the expanding world of regulatory RNAs**. *Essays Biochem* 2013, **54**:1-16.

36. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F *et al*: **Landscape of transcription in human cells**. *Nature* 2012, **489**(7414):101-108.

37. Zhu QH, Wang MB: **Molecular Functions of Long Non-Coding RNAs in Plants**. *Genes (Basel)* 2012, **3**(1):176-190.

38. Li L, Eichten SR, Shimizu R, Petsch K, Yeh CT, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE *et al*: **Genome-wide discovery and characterization of maize long non-coding RNAs**. *Genome Biol* 2014, **15**(2):R40.

39. Rinn JL, Chang HY: **Genome regulation by long noncoding RNAs**. *Annu Rev Biochem* 2012, **81**:145-166.

40. Wang KC, Chang HY: **Molecular mechanisms of long noncoding RNAs**. *Mol Cell* 2011, **43**(6):904-914.

41. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY: **Long noncoding RNA as modular scaffold of histone modification complexes**. *Science* 2010, **329**(5992):689-693.

42. Ginger MR, Shore AN, Contreras A, Rijnkels M, Miller J, Gonzalez-Rimbau MF, Rosen JM: **A noncoding RNA is a potential marker of cell fate during mammary gland development**. *Proc Natl Acad Sci U S A* 2006, **103**(15):5781-5786.

43. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG *et al*: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. *Genome Res* 2012, **22**(9):1775-1789.

44. Pharoah PD, Tsai YY, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, Buckley M, Fridley BL, Tyrer JP, Shen H *et al*: **GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer**. *Nat Genet* 2013, **45**(4):362-370, 370e361-362.

45. Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Crooker BA, Van Tassell CP, Yang J, Wang S, Matukumalli LK *et al*: **Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows**. *BMC Genomics* 2011, **12**:408.

46. Cheng YC, O'Connell JR, Cole JW, Stine OC, Dueker N, McArdle PF, Sparks MJ, Shen J, Laurie CC, Nelson S *et al*: **Genome-wide association analysis of ischemic stroke in young adults**. *G3 (Bethesda)* 2011, **1**(6):505-514.

47. Palti Y, Gao G, Liu S, Kent MP, Lien S, Miller MR, Rexroad CE, Moen T: **The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout**. *Mol Ecol Resour* 2015, **15**(3):662-672.

48. Sanchez CC, Smith TP, Wiedmann RT, Vallejo RL, Salem M, Yao J, Rexroad CE, 3rd: **Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library**. *BMC Genomics* 2009, **10**:559.

49. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G *et al*: **The diploid genome sequence of an individual human**. *PLoS Biol* 2007, **5**(10):e254.

50. Salem M, Vallejo RL, Leeds TD, Palti Y, Liu S, Sabbagh A, Rexroad CE, Yao J: **RNA-Seq identifies SNP markers for growth traits in rainbow trout**. *PLoS One* 2012, **7**(5):e36264.

51. Salem M, Paneru B, Al-Tobasei R, Abdouni F, Thorgaard GH, Rexroad CE, Yao J: **Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout**. *PLoS One* 2015, **10**(3):e0121778.

52. Behenke RJ: **Native trout of western North America**. *American Fisheries Society Monograph* 1992:0362-1715.

53. Papanastasiou AD, Georgaka E, Zarkadis IK: **Cloning of a CD59-like gene in rainbow trout. Expression and phylogenetic analysis of two isoforms**. *Molecular immunology* 2007, **44**(6):1300-1306.

54. Williams DE: **The rainbow trout liver cancer model: response to environmental chemicals and studies on promotion and chemoprevention**. *Comp Biochem Physiol C Toxicol Pharmacol* 2012, **155**(1):121-127.

55. Mc LB, O'Donnell DJ, Elvehjem CA: **Nutrition of rainbow trout**. *Federation proceedings* 1947, **6**(1):413.

56. Welsh PG, Lipton J, Mebane CA, Marr JC: **Influence of flow-through and renewal exposures on the toxicity of copper to rainbow trout**. *Ecotoxicology and environmental safety* 2008, **69**(2):199-208.

57. Ncbi: **National Center for Biotechnology Information; Bethesda (MD)**. In.

58. Brown GE, Harvey MC, Leduc AO, Ferrari MC, Chivers DP: **Social context, competitive interactions and the dynamic nature of antipredator responses of juvenile rainbow trout Oncorhynchus mykiss**. *J Fish Biol* 2009, **75**(3):552-562.

59. Genet C, Dehais P, Palti Y, Gao G, Gavory F, Wincker P, Quillet E, Boussaha M: **Analysis of BAC-end sequences in rainbow trout: content characterization and assessment of synteny between trout and other fish genomes**. *BMC Genomics* 2011, **12**:314.

60. Palti Y, Gahr SA, Hansen JD, Rexroad CE, 3rd: **Characterization of a new BAC library for rainbow trout: evidence for multi-locus duplication**. *Anim Genet* 2004, **35**(2):130-133.

61. Miller M, Palti Y, Luo M, Miller J, Brunelli J, Wheeler P, Rexroad C, Thorgaard G, Doe C: **Rapid and Accurate Sequencing of the Rainbow Trout Physical Map using Illumina Technology**. In. San Diego, California; 2011.

62. Palti Y, Luo MC, Hu Y, Genet C, You FM, Vallejo RL, Thorgaard GH, Wheeler PA, Rexroad CE, 3rd: **A first generation BAC-based physical map of the rainbow trout genome**. *BMC Genomics* 2009, **10**(1):462.

63. Palti Y: **Production of a Draft Reference Genome Sequence for Rainbow Trout**. In. Leetown, West Virginia: USDA-ARS; 2010.

64. Palti Y, Genet C, Luo MC, Charlet A, Gao G, Hu Y, Castano-Sanchez C, Tabet-Canale K, Krieg F, Yao J *et al*: **A first generation integrated map of the rainbow trout genome**. *BMC Genomics* 2011, **12**:180-180.

65. Palti Y, Rexroad Iii CE, Luo MC, Thorgaard GH, Doe CQ, Salem M, Yao J: **Generation of a high density SNP chip for genomic analysis in rainbow trout**. In.: USDA/NIFA grant number WVAR-2010-04523 http://cris.nifa.usda.gov/; 2011.

66. Palti Y, Genet C, Gao G, Hu Y, You FM, Boussaha M, Rexroad CE, 3rd, Luo MC: **A second generation integrated map of the rainbow trout (Oncorhynchus mykiss) genome: analysis of conserved synteny with model fish genomes**. *Mar Biotechnol (NY)* 2012, **14**(3):343-357.

67. Palti Y, Gao G, Miller MR, Vallejo RL, Wheeler PA, Quillet Yao JE, Thorgaard GH, Salem M, Rexroad Iii CE: **Single nucleotide polymorphism (SNP) discovery in rainbow trout using restriction site associated DNA (RAD) sequencing of doubled haploids and assessment of polymorphism in a population survey**. In. San Diego, California.; 2013.

68. Aussanasuwannakul A, Slider SD, Salem M, Yao J, Brett Kenney P: **Comparison of variable-blade to Allo-Kramer shear method in assessing rainbow trout (Oncorhynchus mykiss) fillet firmness**. *J Food Sci* 2012, **77**(9):S335-341.

69. Salgado LR, Koop DM, Pinheiro DG, Rivallan R, Le Guen V, Nicolas MF, de Almeida LG, Rocha VR, Magalhaes M, Gerber AL *et al*: **De novo transcriptome analysis of Hevea brasiliensis tissues by RNA-seq and screening for molecular markers**. *BMC Genomics* 2014, **15**:236.

70. Devisetty UK, Covington MF, Tat AV, Lekkala S, Maloof JN: **Polymorphism Identification and Improved Genome Annotation of Brassica rapa Through Deep RNA Sequencing**. *G3 (Bethesda)* 2014, **4**(11):2065-2078.

71. Salem M, Rexroad CE, Wang J, Thorgaard GH, Yao J: **Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches**. *BMC Genomics* 2010, **11**:564.

72. Sanchez CC, Weber GM, Gao G, Cleveland BM, Yao J, Rexroad CE, 3rd: **Generation of a reference transcriptome for evaluating rainbow trout responses to various stressors**. *BMC Genomics* 2011, **12**:626.

73. Fox SE, Christie MR, Marine M, Priest HD, Mockler TC, Blouin MS: **Sequencing and characterization of the anadromous steelhead (Oncorhynchus mykiss) transcriptome**. *Mar Genomics* 2014, **15**:13-15.

74. **Rainbow trout genome assembly draft** [http://www.animalgenome.org/repository/aquaculture/]

75. Alkan C, Sajjadian S, Eichler EE: **Limitations of next-generation genome sequence assembly**. *Nature methods* 2011, **8**(1):61-65.

76. Bailey GS, Poulter RTM, Stockwell PA: **Gene Duplication in Tetraploid Fish - Model for Gene Silencing at Unlinked Duplicated Loci**. *P Natl Acad Sci USA* 1978, **75**(11):5575-5579.

77. Mehnert JM, Brandenburger M, Grunow B: **Electrophysiological characterization of spontaneously contracting cell aggregates obtained from rainbow trout larvae with multielectrode arrays**. *Cell Physiol Biochem* 2013, **32**(5):1374-1385.

78. Smith CT, Elfstrom CM, Seeb LW, Seeb JE: **Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon**. *Mol Ecol* 2005, **14**(13):4193-4203.

79. Battiprolu PK, Rodnick KJ: **Dichloroacetate selectively improves cardiac function and metabolism in female and male rainbow trout**. *Am J Physiol Heart Circ Physiol* 2014, **307**(10):H1401-1411.

80. Christensen KA, Brunelli JP, Lambert MJ, DeKoning J, Phillips RB, Thorgaard GH: **Identification of single nucleotide polymorphisms from the transcriptome of an organism with a whole genome duplication**. *BMC bioinformatics* 2013, **14**:325.
81. Robison BD, Wheeler PA, Thorgaard GH: **Variation in development rate among clonal lines of rainbow trout (Oncorhynchus mykiss)**. *Aquaculture* 1999, **173**(1-4):131-141.
82. Young WP, Wheeler PA, Fields RD, Thorgaard GH: **DNA fingerprinting confirms isogenicity of androgenetically derived rainbow trout lines**. *J Hered* 1996, **87**(1):77-80.
83. Palti Y, Genet C, Gao G, Hu Y, You FM, Boussaha M, Rexroad CE, Luo MC: **A second Generation Integrated Map of the Rainbow Trout (Oncorhynchus mykiss) Genome: Analysis of Conserved Synteny with Model Fish Genomes**. *Mar Biotechnol (NY)* 2011, **10.1007/s1**.
84. Zhang H, Tan E, Suzuki Y, Hirose Y, Kinoshita S, Okano H, Kudoh J, Shimizu A, Saito K, Watabe S *et al*: **Dramatic improvement in genome assembly achieved using doubled-haploid genomes**. *Scientific reports* 2014, **4**:6780.
85. Watson JD HN, Roberts JW, Steitz JA, Weiner AM: **The functioning of higher eukaryotic genes**. In: *Molecular Biology of the Gene.* Edited by Gene MBot. Molecular Biology of the Gene; 1987.
86. Butte AJ, Dzau VJ, Glueck SB: **Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues"**. *Physiol Genomics* 2001, **7**(2):95-96.
87. Xiao SJ, Zhang C, Zou Q, Ji ZL: **TiSGeD: a database for tissue-specific genes**. *Bioinformatics* 2010, **26**(9):1273-1275.
88. Fowlkes CC, Hendriks CL, Keranen SV, Weber GH, Rubel O, Huang MY, Chatoor S, DePace AH, Simirenko L, Henriquez C *et al*: **A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm**. *Cell* 2008, **133**(2):364-374.
89. Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM: **Global analysis of patterns of gene expression during Drosophila embryogenesis**. *Genome Biol* 2007, **8**(7):R145.
90. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G *et al*: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proc Natl Acad Sci U S A* 2004, **101**(16):6062-6067.
91. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ *et al*: **Genome-wide atlas of gene expression in the adult mouse brain**. *Nature* 2007, **445**(7124):168-176.
92. Bamps S, Hope IA: **Large-scale gene expression pattern analysis, in situ, in Caenorhabditis elegans**. *Brief Funct Genomic Proteomic* 2008, **7**(3):175-183.
93. Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG: **Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans**. *PLoS Comput Biol* 2009, **5**(6):e1000417.
94. Kudoh T, Tsang M, Hukriede NA, Chen X, Dedekian M, Clarke CJ, Kiang A, Schultz S, Epstein JA, Toyama R *et al*: **A gene expression screen in zebrafish embryogenesis**. *Genome Res* 2001, **11**(12):1979-1987.
95. Henrich T, Ramialison M, Wittbrodt B, Assouline B, Bourrat F, Berger A, Himmelbauer H, Sasaki T, Shimizu N, Westerfield M *et al*: **MEPD: a resource for medaka gene expression patterns**. *Bioinformatics* 2005, **21**(14):3195-3197.

96. Allendorf FW, Thorgaard GH: **Tetraploidy and the evolution of salmonid fishes**. In. Edited by Bj T: Plenum Press, New York; 1984: 1-53.

97. Scheerer PD, Thorgaard GH, Allendorf FW: **Genetic analysis of androgenetic rainbow trout**. *The Journal of experimental zoology* 1991, **260**(3):382-390.

98. Scheerer PD TG, Allendorf FW, Knudsen KL: **Androgenetic rainbow trout produced from inbred and outbred sperm show similar survival**. *Aquaculture* 1986, **57**:289-298.

99. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nat Biotechnol* 2011, **29**(7):644-652.

100. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences**. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* 1999:138-148.

101. Lottaz C, Iseli C, Jongeneel CV, Bucher P: **Modeling sequencing errors by combining Hidden Markov models**. *Bioinformatics* 2003, **19 Suppl 2**:ii103-112.

102. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite**. *Nucleic Acids Res* 2008, **36**(10):3420-3435.

103. Gene Ontology C: **The Gene Ontology project in 2008**. *Nucleic acids research* 2008, **36**(Database issue):D440-444.

104. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

105. Jongeneel CV, Delorenzi M, Iseli C, Zhou D, Haudenschild CD, Khrebtukova I, Kuznetsov D, Stevenson BJ, Strausberg RL, Simpson AJ *et al*: **An atlas of human gene expression from massively parallel signature sequencing (MPSS)**. *Genome Res* 2005, **15**(7):1007-1014.

106. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nature protocols* 2012, **7**(3):562-578.

107. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature methods* 2012, **9**(4):357-359.

108. Min XJ, Butler G, Storms R, Tsang A: **OrfPredictor: predicting protein-coding regions in EST-derived sequences.** . *Nucleic acids research* 2005, **Web Server Issue W677-W680**.

109. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.

110. Martin JA, Wang Z: **Next-generation transcriptome assembly**. *Nat Rev Genet* 2011, **12**(10):671-682.

111. Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P: **Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study**. *BMC Bioinformatics* 2011, **12 Suppl 14**:S2.

112. Rexroad CE, 3rd, Lee Y, Keele JW, Karamycheva S, Brown G, Koop B, Gahr SA, Palti Y, Quackenbush J: **Sequence analysis of a rainbow trout cDNA library and creation of a gene index**. *Cytogenet Genome Res* 2003, **102**(1-4):347-354.

113. Wang S, Peatman E, Abernathy J, Waldbieser G, Lindquist E, Richardson P, Lucas S, Wang M, Li P, Thimmapuram J *et al*: **Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies**. *Genome Biol* 2010, **11**(1):R8.

114. Shin SC, Kim SJ, Lee JK, Ahn do H, Kim MG, Lee H, Lee J, Kim BK, Park H: **Transcriptomics and comparative analysis of three antarctic notothenioid fishes**. *PLoS One* 2012, **7**(8):e43762.

115. Al-Tobasei R, Paneru B, Salem M: **Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout**. *PLoS One* 2016, **11**(2):e0148940.

116. Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B: **Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes**. *Molecular biology and evolution* 2011, **28**(3):1205-1215.

117. Ravi V, Venkatesh B: **Rapidly evolving fish genomes and teleost diversity**. *Current opinion in genetics & development* 2008, **18**(6):544-550.

118. Koop BF, von Schalburg KR, Leong J, Walker N, Lieph R, Cooper GA, Robb A, Beetz-Sargent M, Holt RA, Moore R *et al*: **A salmonid EST genomic study: genes, duplications, phylogeny and microarrays**. *BMC Genomics* 2008, **9**:545.

119. Ji P, Liu G, Xu J, Wang X, Li J, Zhao Z, Zhang X, Zhang Y, Xu P, Sun X: **Characterization of common carp transcriptome: sequencing, de novo assembly, annotation and comparative genomics**. *PLoS One* 2012, **7**(4):e35152.

120. **S1 Table transcriptome**. In. Edited by Table S, vol. 2017. http://journals.plos.org/plosone/article/file?type=supplementary&id=info:doi/10.1371/journal.pone.0121778.s001: PLOS one; 2015.

121. Ramskold D, Wang ET, Burge CB, Sandberg R: **An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data**. *PLoS Comput Biol* 2009, **5**(12):e1000598.

122. **S2 Table**. In., vol. 2017. http://journals.plos.org/plosone/article/file?type=supplementary&id=info:doi/10.1371/journal.pone.0121778.s002: PLOS one; 2015.

123. Conn PJ, JP. P: **Physiology and pharmacology of metabotropic glutamate receptors.** *Annual review of pharmacology and toxicology*, **37**:205-237.

124. Spooren WP, Gasparini F, Salt TE, Kuhn R: **Novel allosteric antagonists shed light on mglu(5) receptors and CNS disorders**. *Trends in pharmacological sciences* 2001, **22**(7):331-337.

125. Lamp K, Humeny A, Nikolic Z, Imai K, Adamski J, Schiebel K, Becker CM: **The murine GABA(B) receptor 1: cDNA cloning, tissue distribution, structure of the Gabbr1 gene, and mapping to chromosome 17**. *Cytogenetics and cell genetics* 2001, **92**(1-2):116-121.

126. Suzuki Y, Tasumi S, Tsutsui S, Okamoto M, Suetake H: **Molecular diversity of skin mucus lectins in fish**. *Comp Biochem Physiol B Biochem Mol Biol* 2003, **136**(4):723-730.

127. Tsutsui S, Tasumi S, Suetake H, Suzuki Y: **Lectins homologous to those of monocotyledonous plants in the skin mucus and intestine of pufferfish, Fugu rubripes**. *J Biol Chem* 2003, **278**(23):20882-20889.

128. Hastie ND, Bishop JO: **The expression of three abundance classes of messenger RNA in mouse tissues**. *Cell* 1976, **9**(4 PT 2):761-774.

129. Axel R, Feigelson P, Schutz G: **Analysis of the complexity and diversity of mRNA from chicken liver and oviduct**. *Cell* 1976, **7**(2):247-254.

130. Bishop JO, Morton JG, Rosbash M, Richardson M: **Three abundance classes in HeLa cell messenger RNA**. *Nature* 1974, **250**(463):199-204.

131. Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M *et al*: **Conservation of core gene expression in vertebrate tissues**. *Journal of biology* 2009, **8**(3):33.

132. **Figure 1 Transcriptome**. In. http://journals.plos.org/plosone/article/file?type=supplementary&id=info:doi/10.1371/journal.pone.0121778.s005: PLOS one; 2015.

133. Ali A, Rexroad CE, Thorgaard GH, Yao J, Salem M: **Characterization of the rainbow trout spleen transcriptome and identification of immune-related genes**. *Front Genet* 2014, **5**:348.

134. **S1 DataSet Transcriptome**. In. http://journals.plos.org/plosone/article/file?type=supplementary&id=info:doi/10.1371/journal.pone.0121778.s003: PLOS one; 2015.

135. **S2 DataSet Transcriptom**. In. http://journals.plos.org/plosone/article/file?type=supplementary&id=info:doi/10.1371/journal.pone.0121778.s004: PLOS one; 2015.

136. Modrek B, Lee C: **A genomic view of alternative splicing**. *Nat Genet* 2002, **30**(1):13-19.

137. Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y: **Genome duplication, a trait shared by 22000 species of ray-finned fish**. *Genome Res* 2003, **13**(3):382-390.

138. Steinke D, Salzburger W, Braasch I, Meyer A: **Many genes in fish have species-specific asymmetric rates of molecular evolution**. *BMC Genomics* 2006, **7**:20.

139. Hoegg S, Brinkmann H, Taylor JS, Meyer A: **Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish**. *J Mol Evol* 2004, **59**(2):190-203.

140. Yu WP, Brenner S, Venkatesh B: **Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu**. *Trends Genet* 2003, **19**(4):180-183.

141. Altschmied J, Delfgaauw J, Wilde B, Duschl J, Bouneau L, Volff JN, Schartl M: **Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish**. *Genetics* 2002, **161**(1):259-267.

142. Xing Y, Lee C: **Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes**. *Nat Rev Genet* 2006, **7**(7):499-509.

143. Lu J, Peatman E, Wang W, Yang Q, Abernathy J, Wang S, Kucuktas H, Liu Z: **Alternative splicing in teleost fish genomes: same-species and cross-species analysis and comparisons**. *Mol Genet Genomics* 2010, **283**(6):531-539.

144. Gibb EA, Brown CJ, Lam WL: **The functional role of long non-coding RNA in human carcinomas**. *Mol Cancer* 2011, **10**:38.

145. Ponting CP, Oliver PL, Reik W: **Evolution and functions of long noncoding RNAs**. *Cell* 2009, **136**(4):629-641.

146. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP *et al*: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals**. *Nature* 2009, **458**(7235):223-227.

147. Guttman M, Rinn JL: **Modular regulatory principles of large non-coding RNAs**. *Nature* 2012, **482**(7385):339-346.

148. Beaulieu YB, Kleinman CL, Landry-Voyer AM, Majewski J, Bachand F: **Polyadenylation-dependent control of long noncoding RNA expression by the poly(A)-binding protein nuclear 1**. *PLoS Genet* 2012, **8**(11):e1003078.

149. Yin QF, Yang L, Zhang Y, Xiang JF, Wu YW, Carmichael GG, Chen LL: **Long noncoding RNAs with snoRNA ends**. *Mol Cell* 2012, **48**(2):219-230.

150. Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q *et al*: **Long noncoding RNAs with enhancer-like function in human cells**. *Cell* 2010, **143**(1):46-58.

151. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**. *Genes Dev* 2011, **25**(18):1915-1927.

152. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS: **Specific expression of long noncoding RNAs in the mouse brain**. *Proc Natl Acad Sci U S A* 2008, **105**(2):716-721.

153. Prasanth KV, Prasanth SG, Xuan Z, Hearn S, Freier SM, Bennett CF, Zhang MQ, Spector DL: **Regulating gene expression through RNA nuclear retention**. *Cell* 2005, **123**(2):249-263.

154. Hung T, Wang Y, Lin MF, Koegel AK, Kotake Y, Grant GD, Horlings HM, Shah N, Umbricht C, Wang P *et al*: **Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters**. *Nat Genet* 2011, **43**(7):621-629.

155. Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP: **Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor**. *Sci Signal* 2010, **3**(107):ra8.

156. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-Dinardo D, Kanduri C: **Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation**. *Mol Cell* 2008, **32**(2):232-246.

157. Yap KL, Li S, Muñoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, Gil J, Walsh MJ, Zhou MM: **Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a**. *Mol Cell* 2010, **38**(5):662-674.

158. Kotake Y, Nakagawa T, Kitagawa K, Suzuki S, Liu N, Kitagawa M, Xiong Y: **Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene**. *Oncogene* 2011, **30**(16):1956-1962.

159. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M *et al*: **A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response**. *Cell* 2010, **142**(3):409-419.

160. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL *et al*: **Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis**. *Nature* 2010, **464**(7291):1071-1076.

161. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E *et al*: **Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs**. *Cell* 2007, **129**(7):1311-1323.

162. Schmitz KM, Mayer C, Postepska A, Grummt I: **Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes**. *Genes Dev* 2010, **24**(20):2264-2269.

163. Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A: **Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript**. *Nature* 2007, **445**(7128):666-670.

164. Jeon Y, Lee JT: **YY1 tethers Xist RNA to the inactive X nucleation center**. *Cell* 2011, **146**(1):119-133.

165. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA *et al*: **The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation**. *Mol Cell* 2010, **39**(6):925-938.

166. Zong X, Tripathi V, Prasanth KV: **RNA splicing control: yet another gene regulatory role for long nuclear noncoding RNAs**. *RNA Biol* 2011, **8**(6):968-977.

167. Yoon JH, Abdelmohsen K, Gorospe M: **Posttranscriptional gene regulation by long noncoding RNA**. *J Mol Biol* 2013, **425**(19):3723-3730.

168. Tripathi V, Song DY, Zong X, Shevtsov SP, Hearn S, Fu XD, Dundr M, Prasanth KV: **SRSF1 regulates the assembly of pre-mRNA processing factors in nuclear speckles**. *Mol Biol Cell* 2012, **23**(18):3694-3706.

169. McLAREN BA, O'DONNELL DJ, ELVEHJEM CA: **Nutrition of rainbow trout**. *Federation proceedings* 1947, **6**(1):413.

170. Kambara H, Niazi F, Kostadinova L, Moonka DK, Siegel CT, Post AB, Carnero E, Barriocanal M, Fortes P, Anthony DD *et al*: **Negative regulation of the interferon response by an interferon-induced long non-coding RNA**. *Nucleic Acids Res* 2014, **42**(16):10668-10680.

171. Yang Z, Zhou L, Wu LM, Lai MC, Xie HY, Zhang F, Zheng SS: **Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation**. *Ann Surg Oncol* 2011, **18**(5):1243-1250.

172. Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, Lee CS, Flockhart RJ, Groff AF, Chow J *et al*: **Control of somatic tissue differentiation by the long non-coding RNA TINCR**. *Nature* 2013, **493**(7431):231-235.

173. Luo M, Li Z, Wang W, Zeng Y, Liu Z, Qiu J: **Long non-coding RNA H19 increases bladder cancer metastasis by associating with EZH2 and inhibiting E-cadherin expression**. *Cancer Lett* 2013, **333**(2):213-221.

174. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C *et al*: **Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat**. *Nature* 2012, **491**(7424):454-457.

175. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A *et al*: **Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis**. *Genome Res* 2012, **22**(3):577-591.

176. Narum SR, Campbell NR: **Transcriptomic response to heat stress among ecologically divergent populations of redband trout**. *BMC Genomics* 2015, **16**:103.

177. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G: **CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine**. *Nucleic acids research* 2007, **35**(Web Server issue):W345-349.

178. Chan PP, Lowe TM: **GtRNAdb: a database of transfer RNA genes detected in genomic sequence**. *Nucleic acids research* 2009, **37**(Database issue):D93-97.

179. Wuyts J, Van de Peer Y, Winkelmans T, De Wachter R: **The European database on small subunit ribosomal RNA**. *Nucleic acids research* 2002, **30**(1):183-185.

180. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools**. *Nucleic acids research* 2013, **41**(Database issue):D590-596.

181. Van Peer G, Lefever S, Anckaert J, Beckers A, Rihani A, Van Goethem A, Volders PJ, Zeka F, Ongenaert M, Mestdagh P *et al*: **miRBase Tracker: keeping track of microRNA annotation changes**. *Database (Oxford)* 2014, **2014**.

182. Bu D, Yu K, Sun S, Xie C, Skogerbo G, Miao R, Xiao H, Liao Q, Luo H, Zhao G *et al*: **NONCODE v3.0: integrative annotation of long noncoding RNAs**. *Nucleic acids research* 2012, **40**(Database issue):D210-215.

183. Christensen KA, Brunelli JP, Wheeler PA, Thorgaard GH: **Antipredator behavior QTL: differences in rainbow trout clonal lines derived from wild and hatchery populations**. *Behav Genet* 2014, **44**(5):535-546.

184. Galt NJ, Froehlich JM, Remily EA, Romero SR, Biga PR: **The effects of exogenous cortisol on myostatin transcription in rainbow trout, Oncorhynchus mykiss**. *Comp Biochem Physiol A Mol Integr Physiol* 2014, **175**:57-63.

185. Cleveland BM: **In vitro and in vivo effects of phytoestrogens on protein turnover in rainbow trout (Oncorhynchus mykiss) white muscle**. *Comparative biochemistry and physiology Toxicology & pharmacology : CBP* 2014, **165**:9-16.

186. Sovadinová I, Liedtke A, Schirmer K: **Effects of clofibric acid alone and in combination with 17β-estradiol on mRNA abundance in primary hepatocytes isolated from rainbow trout**. *Toxicol In Vitro* 2014, **28**(6):1106-1116.

187. **Annotation of Non-Coding RNAs** [http://useast.ensembl.org/info/genome/genebuild/ncrna.html?redirect=no]

188. Howe EA, Sinha R, Schlauch D, Quackenbush J: **RNA-Seq analysis in MeV**. *Bioinformatics* 2011, **27**(22):3209-3210.

189. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M *et al*: **TM4: a free, open-source system for microarray data management and analysis**. *Biotechniques* 2003, **34**(2):374-378.

190. Louro R, Smirnova AS, Verjovski-Almeida S: **Long intronic noncoding RNA transcription: expression noise or expression choice?** *Genomics* 2009, **93**(4):291-298.

191. Zhang K, Huang K, Luo Y, Li S: **Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data**. *BMC Genomics* 2014, **15**:845.

192. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM *et al*: **Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome**. *Genome Res* 2006, **16**(1):11-19.

193. Mercer TR, Dinger ME, Mattick JS: **Long non-coding RNAs: insights into functions**. *Nat Rev Genet* 2009, **10**(3):155-159.

194. Villegas VE, Zaphiropoulos PG: **Neighboring gene regulation by antisense long non-coding RNAs**. *Int J Mol Sci* 2015, **16**(2):3251-3266.

195. Li R: **Transient transfection of CHO cells using linear polyethylenimine is a simple and effective means of producing rainbow trout recombinant IFN-γ protein**. *Cytotechnology* 2014.

196. Long A, Call DR, Cain KD: **Investigation of the link between broodstock infection, vertical transmission, and prevalence of Flavobacterium psychrophilum in eggs and progeny of Rainbow Trout and Coho Salmon**. *J Aquat Anim Health* 2014, **26**(2):66-77.

197. Eya JC, Yossa R, Ashame MF, Pomeroy CF, Gannam AL: **Effects of dietary lipid levels on mitochondrial gene expression in low and high-feed efficient families of rainbow trout Oncorhynchus mykiss**. *J Fish Biol* 2014, **84**(6):1708-1720.

198. Ulitsky I, Bartel DP: **lincRNAs: genomics, evolution, and mechanisms**. *Cell* 2013, **154**(1):26-46.

199. Al-Tobasei R, Ali A, Leeds T, Liu S, Palti Y, Kenney B, Salem M: **Identification of SNPs Associated with Muscle Yield and Quality Traits Using Allelic-Imbalance Analysis in Pooled RNA-Seq Samples in Rainbow Trout**. In. submitted to BMC Genomics (under review); 2017.

200. Salem M, Kenney PB, Rexroad CE, Yao J: **Molecular characterization of muscle atrophy and proteolysis associated with spawning in rainbow trout**. *Comp Biochem Physiol Part D Genomics Proteomics* 2006, **1**(2):227-237.

201. Gjedrem T: **Selection and Breeding Programs in Aquaculture.** . New York: Springer; 2008.

202. Rexroad CE, Palti Y, Gahr SA, Vallejo RL: **A second generation genetic map for rainbow trout (Oncorhynchus mykiss)**. *BMC Genet* 2008, **9**:74-74.

203. Wang R, Sun L, Bao L, Zhang J, Jiang Y, Yao J, Song L, Feng J, Liu S, Liu Z: **Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish**. *BMC Genomics* 2013, **14**:929.

204. Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridhivej B, Peatman E, Kucuktas H, Liu Z: **Quality assessment parameters for EST-derived SNPs from catfish**. *BMC Genomics* 2008, **9**:450.

205. Gonzalez-Pena D, Gao G, Baranski, M., Moen, T., , Cleveland B, Kenney P, Vallejo R, Palti Y, Leeds T: **Genome-Wide Association Study for Identifying Loci that Affect Fillet Yield, Carcass, and Body Weight Traits in Rainbow Trout (Oncorhynchus mykiss).** . *Front Genet* 2016, **7**.

206. Tsai HY, Hamilton A, Tinch AE, Guy DR, Gharbi K, Stear MJ, Matika O, Bishop SC, Houston RD: **Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array**. *BMC Genomics* 2015, **16**:969.

207. Carlton VE, Ireland JS, Useche F, Faham M: **Functional single nucleotide polymorphism-based association studies**. *Hum Genomics* 2006, **2**(6):391-402.

208. Brookes AJ: **Single Nucleotide Polymorphism (SNP)**. In: *ENCYCLOPEDIA OF LIFE SCIENCES (els).* Edited by John Wiley & Sons Lwen, http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0005006.pub2/pdf, accessed April 4, 2012; 2007.

209. Villanueva B, Dekkers JC, Woolliams JA, Settar P: **Maximizing genetic gain over multiple generations with quantitative trait locus selection and control of inbreeding**. *J Anim Sci* 2004, **82**(5):1305-1314.

210. Dekkers JC: **Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons**. *J Anim Sci* 2004, **82 E-Suppl**:E313-E328.

211. Pang Y, Wang J, Zhang C, Lei C, Lan X, Yue W, Gu C, Chen D, Chen H: **The polymorphisms of bovine VEGF gene and their associations with growth traits in Chinese cattle**. *Mol Biol Rep* 2010.

212. Tsai HY, Hamilton A, Guy DR, Tinch AE, Bishop SC, Houston RD: **The genetic architecture of growth and fillet traits in farmed Atlantic salmon (Salmo salar)**. *BMC Genet* 2015, **16**:51.

213. Harvey DJ: **Aquaculture Outlook**. In: *Electronic Outlook Report from the Economic Research Service wwwersusdagov.* 2006.

214. Salem M, Kenney PB, Rexroad CE, 3rd, Yao J: **Microarray gene expression analysis in atrophying rainbow trout muscle: a unique nonmammalian muscle degradation model**. *Physiological genomics* 2006, **28**(1):33-45.

215. Cirulli ET, Singh A, Shianna KV, Ge D, Smith JP, Maia JM, Heinzen EL, Goedert JJ, Goldstein DB, Center for HIVAVI: **Screening the human exome: a comparison of whole genome and whole transcriptome sequencing**. *Genome Biol* 2010, **11**(5):R57.

216. Heap GA, Yang JH, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ *et al*: **Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing**. *Hum Mol Genet* 2010, **19**(1):122-134.

217. Vidal RO, do Nascimento LC, Mondego JM, Pereira GA, Carazzolle MF: **Identification of SNPs in RNA-seq data of two cultivars of Glycine max (soybean) differing in drought resistance**. *Genet Mol Biol* 2012, **35**(1 (suppl)):331-334.

218. Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JF, Jung HJ, Vance CP, Gronwald JW: **Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems**. *BMC Genomics* 2011, **12**:199.

219. Leeds TD, Vallejo RL, Weber GM, Pena DG, Silverstein JS: **Response to five generations of selection for growth performance traits in rainbow trout (Oncorhynchus mykiss)**. *Aquaculture* 2016, **465**:341-351.

220. Manor ML, Cleveland BM, Kenney PB, Yao J, Leeds T: **Differences in growth, fillet quality, and fatty acid metabolism-related gene expression between juvenile male and female rainbow trout**. *Fish Physiology and Biochemistry* 2015, **41**(2):533-547.

221. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.

222. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

223. Kofler R, Pandey RV, Schlotterer C: **PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq)**. *Bioinformatics* 2011, **27**(24):3435-3436.

224. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J *et al*: **From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline**. *Curr Protoc Bioinformatics* 2013, **43**:11 10 11-33.

225. O'Connell JR, Weeks DE: **PedCheck: A program for identification of genotype incompatibilities in linkage analysis**. *Am J Hum Genet* 1998, **63**(1):259-266.

226. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic acids research* 2012, **40**(Database issue):D109-114.

227. Leeds TD, Kenney PB, Manor M: **Genetic parameter estimates for feed intake, body composition, and fillet quality traits in a rainbow trout population selected for improved growth**. In: *International Symposium on Genetics in Aquaculture 2012; Auburn, AL*; 2012: 259.

228. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.

229. Raineri E, Ferretti L, Esteve-Codina A, Nevado B, Heath S, Pérez-Enciso M: **SNP calling by sequencing pooled samples**. *BMC bioinformatics* 2012, **13**:239.

230. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

231. Palti Y, Gao G, Miller MR, Vallejo RL, Wheeler PA, Quillet E, Yao J, Thorgaard GH, Salem M, Rexroad CE, 3rd: **A resource of single-nucleotide polymorphisms for rainbow trout generated by restriction-site associated DNA sequencing of doubled haploids**. *Mol Ecol Resour* 2014, **14**(3):588-596.

232. Seeb JE, Pascal CE, Grau ED, Seeb LW, Templin WD, Harkins T, Roberts SB: **Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids**. *Mol Ecol Resour* 2011, **11**(2):335-348.

233. Ryynanen HJ, Primmer CR: **Single nucleotide polymorphism (SNP) discovery in duplicated genomes: intron-primed exon-crossing (IPEC) as a strategy for avoiding amplification of duplicated loci in Atlantic salmon (Salmo salar) and other salmonid fishes**. *BMC Genomics* 2006, **7**:192.

234. Konczal M, Koteja P, Stuglik MT, Radwan J, Babik W: **Accuracy of allele frequency estimation using pooled RNA-Seq**. *Molecular Ecology Resources* 2014, **14**(2):381-392.

235. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W *et al*: **Chromatin architecture reorganization during stem cell differentiation**. *Nature* 2015, **518**(7539):331-336.

236. Eckersley-Maslin MA, Spector DL: **Random monoallelic expression: regulating gene expression one allele at a time**. *Trends in genetics : TIG* 2014, **30**(6):237-244.

237. Chen A, Wang R, Liu S, Peatman E, Sun L, Bao L, Jiang C, Li C, Li Y, Zeng Q *et al*: **Ribosomal protein genes are highly enriched among genes with allele-specific expression in the interspecific F1 hybrid catfish**. *Molecular genetics and genomics : MGG* 2016, **291**(3):1083-1093.

238. Deng Q, Ramskold D, Reinius B, Sandberg R: **Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells**. *Science* 2014, **343**(6167):193-196.

239. Piskol R, Ramaswami G, Li JB: **Reliable identification of genomic variants from RNA-seq data**. *Am J Hum Genet* 2013, **93**(4):641-651.

240. Danzmann RG, Kocmarek AL, Norman JD, Rexroad CE, Palti Y: **Transcriptome profiling in fast versus slow-growing rainbow trout across seasonal gradients**. *Bmc Genomics* 2016, **17**.

241. Wringe BF, Devlin RH, Ferguson MM, Moghadam HK, Sakhrani D, Danzmann RG: **Growth-related quantitative trait loci in domestic and wild rainbow trout (Oncorhynchus mykiss)**. *Bmc Genetics* 2010, **11**.

242. Salem M, Kenney PB, Rexroad CE, 3rd, Yao J: **Proteomic signature of muscle atrophy in rainbow trout**. *J Proteomics* 2010, **73**(4):778-789.

243. Salem M, Kenney PB, Rexroad CE, III, Yao J: **Development of a 37 k high-density oligonucleotide microarray: a new tool for functional genome research in rainbow trout**. *Journal of Fish Biology* 2008, **72**(9):2187-2206.

244. Rescan PY, Montfort J, Ralliere C, Le Cam A, Esquerre D, Hugot K: **Dynamic gene expression in fish muscle during recovery growth induced by a fasting-refeeding schedule**. *Bmc Genomics* 2007, **8**.

245. Fuentes EN, Ruiz P, Valdes JA, Molina A: **Catabolic signaling pathways, atrogenes, and ubiquitinated proteins are regulated by the nutritional status in the muscle of the fine flounder**. *PLoS One* 2012, **7**(9):e44256.

246. Malila Y, Carr KM, Ernst CW, Velleman SG, Reed KM, Strasburg GM: **Deep transcriptome sequencing reveals differences in global gene expression between normal and pale, soft, and exudative turkey meat**. *Journal of Animal Science* 2014, **92**(3):1250-1260.

247. Fan H, Wu Y, Zhou X, Xia J, Zhang W, Song Y, Liu F, Chen Y, Zhang L, Gao X *et al*: **Pathway-Based Genome-Wide Association Studies for Two Meat Production Traits in Simmental Cattle**. *Scientific reports* 2015, **5**:18389.

248. Paneru B, Al-Tobasei R, Palti Y, Wiens GD, Salem M: **Differential expression of long non-coding RNAs in three genetic lines of rainbow trout in response to infection with Flavobacterium psychrophilum**. *Scientific reports* 2016, **6**:36032.

249. Asche F, Håvard H, Ragnar T, Sigbjørn T: **The salmon disease crisis in Chile**. *Marine Resource Economics* 2009, **24**( 4):405-411.

250. Nematollahi A, Decostere A, Pasmans F, Haesebrouck F: **Flavobacterium psychrophilum infections in salmonid fish**. *J Fish Dis* 2003, **26**(10):563-574.

251. Kent L, Groff J, Morrison J, Yasutake W, Holt R: **Spiral swimming behavior due to cranial and vertebral lesions associated with Cytophaga psychrophila infections in salmonid fishes**. *Diseases of Aquatic Organisms* 1989, **6**(1):11-16.

252. Carson L, Schmidtke J: **Characteristics of Flexibacter psychrophilus isolated from Atlantic salmon in Australia**. *Diseases of Aquatic Organisms* 1995, **21**:157-161.

253. Brown L, Cox W, Levine R: **Evidence that the causal agent of bacterial cold-water disease Flavobacterium psychrophilum is transmitted within salmonid eggs**. *Diseases of Aquatic Organisms* 1997, **29**(3):213-218.

254. Gómez E, Méndez J, Cascales D, Guijarro JA: **Flavobacterium psychrophilum vaccine development: a difficult task**. *Microb Biotechnol* 2014, **7**(5):414-423.

255. Gjedrem T: ***Selection and breeding programs in aquaculture***. *Dordrecht: Springer* 2005.

256. Wiens GD, Scott E L, Timothy J W, Jason P E, Caird E R, Timothy D L: **On-farm performance of rainbow trout (Oncorhynchus mykiss) selectively bred for resistance to bacterial cold water disease: effect of rearing environment on survival phenotype**. *Aquaculture* 2013, **388**:128-136.

257. Marancik D, Gao G, Paneru B, Ma H, Hernandez AG, Salem M, Yao J, Palti Y, Wiens GD: **Whole-body transcriptome of selectively bred, resistant-, control-, and susceptible-line rainbow trout following experimental challenge with Flavobacterium psychrophilum**. *Front Genet* 2014, **5**:453.

258. Peng X, Gralinski L, Armour CD, Ferris MT, Thomas MJ, Proll S, Bradel-Tretheway BG, Korth MJ, Castle JC, Biery MC *et al*: **Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling**. *MBio* 2010, **1**(5).

259. Carpenter S, Aiello D, Atianand MK, Ricci EP, Gandhi P, Hall LL, Byron M, Monks B, Henry-Bezy M, Lawrence JB *et al*: **A long noncoding RNA mediates both activation and repression of immune response genes**. *Science* 2013, **341**(6147):789-792.

260. Wang P, Xue Y, Han Y, Lin L, Wu C, Xu S, Jiang Z, Xu J, Liu Q, Cao X: **The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation**. *Science* 2014, **344**(6181):310-313.

261. Hu G, Tang Q, Sharma S, Yu F, Escobar TM, Muljo SA, Zhu J, Zhao K: **Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation**. *Nat Immunol* 2013, **14**(11):1190-1198.

262. Xia F, Dong F, Yang Y, Huang A, Chen S, Sun D, Xiong S, Zhang J: **Dynamic transcription of long non-coding RNA genes during CD4+ T cell development and activation**. *PLoS One* 2014, **9**(7):e101588.

263. Gomez JA, Wapinski OL, Yang YW, Bureau JF, Gopinath S, Monack DM, Chang HY, Brahic M, Kirkegaard K: **The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus**. *Cell* 2013, **152**(4):743-754.

264. Collier SP, Collins PL, Williams CL, Boothby MR, Aune TM: **Cutting edge: influence of Tmevpg1, a long intergenic noncoding RNA, on the expression of Ifng by Th1 cells**. *J Immunol* 2012, **189**(5):2084-2088.

265. Boltana S, Valenzuela-Miranda D, Aguilar A, Mackenzie S, Gallardo-Escarate C: **Long noncoding RNAs (lncRNAs) dynamics evidence immunomodulation during ISAV-Infected Atlantic salmon (Salmo salar)**. *Scientific reports* 2016, **6**:22698.

266. Schmittgen TD, Livak KJ: **Analyzing real-time PCR data by the comparative C(T) method**. *Nature protocols* 2008, **3**(6):1101-1108.

267. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

268. **DataSet 1ALNCRNA**. In: *Scientific reports.* http://www.nature.com/article-assets/npg/srep/2016/161027/srep36032/extref/srep36032-s1.doc; 2016.

269. **DataSet 2LNCRNA**. In: *Scientific Report.* http://www.nature.com/article-assets/npg/srep/2016/161027/srep36032/extref/srep36032-s2.xls: Scientific Report; 2016.

270. Madetoja J, Nyman P, Wiklund T: **Flavobacterium psychrophilum, invasion into and shedding by rainbow trout Oncorhynchus mykiss**. *Diseases of Aquatic Organisms* 2000, **43**(1):27-38.

271. Tian D, Sun S, Lee JT: **The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation**. *Cell* 2010, **143**(3):390-403.

272. **DataSet 3 LNCRNA**. In. http://www.nature.com/article-assets/npg/srep/2016/161027/srep36032/extref/srep36032-s3.xls: Scientific Report; 2016.

273. **DataSet 4 LNCRNA**. In: *Scientific Report.* http://www.nature.com/article-assets/npg/srep/2016/161027/srep36032/extref/srep36032-s4.xls; 2016.

274. NE II, Heward JA, Roux B, Tsitsiou E, Fenwick PS, Lenzi L, Goodhead I, Hertz-Fowler C, Heger A, Hall N *et al*: **Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes**. *Nature communications* 2014, **5**:3979.

275. Barriocanal M, Carnero E, Segura V, Fortes P: **Long Non-Coding RNA BST2/BISPR is Induced by IFN and Regulates the Expression of the Antiviral Factor Tetherin**. *Front Immunol* 2014, **5**:655.

276. Jeffries KM, Hinch SG, Gale MK, Clark TD, Lotto AG, Casselman MT, Li S, Rechisky EL, Porter AD, Welch DW *et al*: **Immune response genes and pathogen presence predict migration survival in wild salmon smolts**. *Molecular ecology* 2014, **23**(23):5803-5815.

277. Messemaker TC, Frank-Bertoncelj M, Marques RB, Adriaans A, Bakker AM, Daha N, Gay S, Huizinga TW, Toes RE, Mikkers HM *et al*: **A novel long non-coding RNA in the rheumatoid arthritis risk locus TRAF1-C5 influences C5 mRNA levels**. *Genes Immun* 2016, **17**(2):85-92.
278. Saurabh S, Sahoo PK: **Lysozyme: an important defence molecule of fish innate immune system**. *Aquaculture Research* 2008, **39**(3):223-239.
279. Lefevre P, Witham J, Lacroix CE, Cockerill PN, Bonifer C: **The LPS-induced transcriptional upregulation of the chicken lysozyme locus involves CTCF eviction and noncoding RNA transcription**. *Mol Cell* 2008, **32**(1):129-139.
280. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W182-185.
281. **DataSet 5 LNCRNA**. In: *Scientific Report.* http://www.nature.com/article-assets/npg/srep/2016/161027/srep36032/extref/srep36032-s5.xls; 2016.
282. Lopes-Ferreira M, Magalhães GS, Fernandez JH, Junqueira-de-Azevedo IeL, Le Ho P, Lima C, Valente RH, Moura-da-Silva AM: **Structural and biological characterization of Nattectin, a new C-type lectin from the venomous fish Thalassophryne nattereri**. *Biochimie* 2011, **93**(6):971-980.
283. Saraiva TC, Grund LZ, Komegae EN, Ramos AD, Conceição K, Orii NM, Lopes-Ferreira M, Lima C: **Nattectin a fish C-type lectin drives Th1 responses in vivo: licenses macrophages to differentiate into cells exhibiting typical DC function**. *Int Immunopharmacol* 2011, **11**(10):1546-1556.
284. Parks WC, Wilson CL, López-Boado YS: **Matrix metalloproteinases as modulators of inflammation and innate immunity**. *Nat Rev Immunol* 2004, **4**(8):617-629.
285. Bach JP, Borta H, Ackermann W, Faust F, Borchers O, Schrader M: **The secretory granule protein syncollin localizes to HL-60 cells and neutrophils**. *J Histochem Cytochem* 2006, **54**(8):877-888.
286. Belaaouaj A, McCarthy R, Baumann M, Gao Z, Ley TJ, Abraham SN, Shapiro SD: **Mice lacking neutrophil elastase reveal impaired host defense against gram negative bacterial sepsis**. *Nat Med* 1998, **4**(5):615-618.
287. Belaaouaj A, Kim KS, Shapiro SD: **Degradation of outer membrane protein A in Escherichia coli killing by neutrophil elastase**. *Science* 2000, **289**(5482):1185-1188.
288. Belaaouaj A: **Neutrophil elastase-mediated killing of bacteria: lessons from targeted mutagenesis**. *Microbes Infect* 2002, **4**(12):1259-1264.
289. Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, Almeida R, Zhernakova A, Reinmaa E, Vosa U *et al*: **Human disease-associated genetic variation impacts large intergenic non-coding RNA expression**. *PLoS Genet* 2013, **9**(1):e1003201.
290. Liu Y, Pan S, Liu L, Zhai X, Liu J, Wen J, Zhang Y, Chen J, Shen H, Hu Z: **A genetic variant in long non-coding RNA HULC contributes to risk of HBV-related hepatocellular carcinoma in a Chinese population**. *PLoS One* 2012, **7**(4):e35145.
291. Pawar K, Hanisch C, Palma Vera SE, Einspanier R, Sharbati S: **Down regulated lncRNA MEG3 eliminates mycobacteria in macrophages via autophagy**. *Scientific reports* 2016, **6**:19416.
292. Caamaño JH, Rizzo CA, Durham SK, Barton DS, Raventós-Suárez C, Snapper CM, Bravo R: **Nuclear factor (NF)-kappa B2 (p100/p52) is required for normal splenic**

**microarchitecture and B cell-mediated immune responses**. *J Exp Med* 1998, **187**(2):185-196.

293. Scaria V, Pasha A: **Long Non-Coding RNAs in Infection Biology**. *Front Genet* 2012, **3**:308.
294. **DataSet 6 LNCRNA**. In: *Scientific Report.* http://www.nature.com/article-assets/npg/srep/2016/161027/srep36032/extref/srep36032-s6.xls; 2016.
295. **DataSet 7 LNCRNA**. In: *Scientific Report.* http://www.nature.com/article-assets/npg/srep/2016/161027/srep36032/extref/srep36032-s7.xls

2016.

296. Yamada T, Goto I, Sakaki Y: **Neuron-specific splicing of the Alzheimer amyloid precursor protein gene in a mini-gene system**. *Biochem Biophys Res Commun* 1993, **195**(1):442-448.
297. Shi J, Qian W, Yin X, Iqbal K, Grundke-Iqbal I, Gu X, Ding F, Gong CX, Liu F: **Cyclic AMP-dependent protein kinase regulates the alternative splicing of tau exon 10: a mechanism involved in tau pathology of Alzheimer disease**. *The Journal of biological chemistry* 2011, **286**(16):14639-14648.
298. Rogaev EI, Sherrington R, Wu C, Levesque G, Liang Y, Rogaeva EA, Ikeda M, Holman K, Lin C, Lukiw WJ *et al*: **Analysis of the 5' sequence, genomic structure, and alternative splicing of the presenilin-1 gene (PSEN1) associated with early onset Alzheimer disease**. *Genomics* 1997, **40**(3):415-424.
299. Humphries C, Kohli MA, Whitehead P, Mash DC, Pericak-Vance MA, Gilbert J: **Alzheimer disease (AD) specific transcription, DNA methylation and splicing in twenty AD associated loci**. *Mol Cell Neurosci* 2015, **67**:37-45.
300. Zhao YJ, Han HZ, Liang Y, Shi CZ, Zhu QC, Yang J: **Alternative splicing of VEGFA, APP and NUMB genes in colorectal cancer**. *World J Gastroenterol* 2015, **21**(21):6550-6560.
301. Yuge S, Richter CA, Wright-Osment MK, Nicks D, Saloka SK, Tillitt DE, Li W: **Identification of the thiamin pyrophosphokinase gene in rainbow trout: characteristic structure and expression of seven splice variants in tissues and cell lines and during embryo development**. *Comparative biochemistry and physiology Part B, Biochemistry & molecular biology* 2012, **163**(2):193-202.
302. Yoshida T, Kim JH, Carver K, Su Y, Weremowicz S, Mulvey L, Yamamoto S, Brennan C, Mei S, Long H *et al*: **CLK2 Is an Oncogenic Kinase and Splicing Regulator in Breast Cancer**. *Cancer Res* 2015, **75**(7):1516-1526.
303. Tang JY, Li RN, Chen PH, Huang HW, Hou MF, Chang HW: **Alternative Splicing, DNA Damage and Modulating Drugs in Radiation Therapy for Cancer**. *Anticancer Agents Med Chem* 2015, **15**(6):674-680.
304. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat Genet* 2008, **40**(12):1413-1415.
305. Lee C, Wang Q: **Bioinformatics analysis of alternative splicing**. *Brief Bioinform* 2005, **6**(1):23-33.
306. Douglas, L: **Black mechanisms of alternative Pre-Messenger RNA Splicing**. In., vol. 72: Biochem; 2003: 291-336.

# APPENDIX

# APPENDIX A: Transcriptome-wide Detection of Tissue-specific Alternative Splicing in Rainbow Trout

## Introduction

Alternative splicing is a process where different RNA transcripts are generated from the same pre-mRNA. This process incorporates different exons of the same gene into mRNAs that produce structurally and functionally different proteins or isoforms. Alternative splicing is specific to tissues, mRNAs and developmental stages. Miss-regulated mRNA splicing is reported in many diseases such as human Alzheimer's [296-299] and cancer [300-303]. Determining the alternative splice variants answers important questions about the genome for any species. For example, it permits estimating the number of genes versus the number of proteins produced in a cell type, tissue and/or species. Hence, it enables interesting studies in the fields of comparative genomics and evolutionary genomics.

Over the last decade, studying alternative splicing using bioinformatics has become an important new field. Emergence of the next-generation RNA sequencing (RNA-seq) technology offers an unprecedented opportunity for genome-wide detection of alternative splicing. However, one of the most challenging processes in bioinformatics is the transcriptome-wide discovery and characterization of alternative splicing in non-model species, where a complete reference genome is not available.

According to different analyses and studies, in the human genome and other species, alternative splicing takes place in 60% to 95% of the genes [136, 304, 305].

Five different types of alternative splice are possible [306].

• Exon skipping, the most common type, in which one exon or more is either included or excluded in the formation of mRNA (see Figure 24A).

• Intron retention, where either part or all the intron is not removed from the mRNA production leading to changes in the functionality and structure of the protein as shown in (Figure 24B).

• Third and fourth types of alternative splicing are alteration of the 3’ splice site and 5’ splice site in which two 3’ ends of the exon splice with the 5’ junction and vice versa (Figure 24C and 24D).

• Mutually exclusive exons: one of two exons is retained in mRNA after splicing, but not both (Figure 24E).

In this chapter, we introduce a naïve approach to detect alternative splicing without the need to have RNA pre-assembled to the full length or the existence of a reference genome. The main concept of this “Splice Detection through Gap Existence” method is based on the fact that all types of splice variants (Figure 24) will lead to gaps when sequence reads from different tissues with splice variances are mapped back to a transcriptome reference that is assembled from reads of all tissues.

This approach requires output from an assembler such as Trinity [32] and the reads for individual tissues from a sequencer such as Illumina. To the best of our knowledge, this approach has not been used to detect alternative splicing.

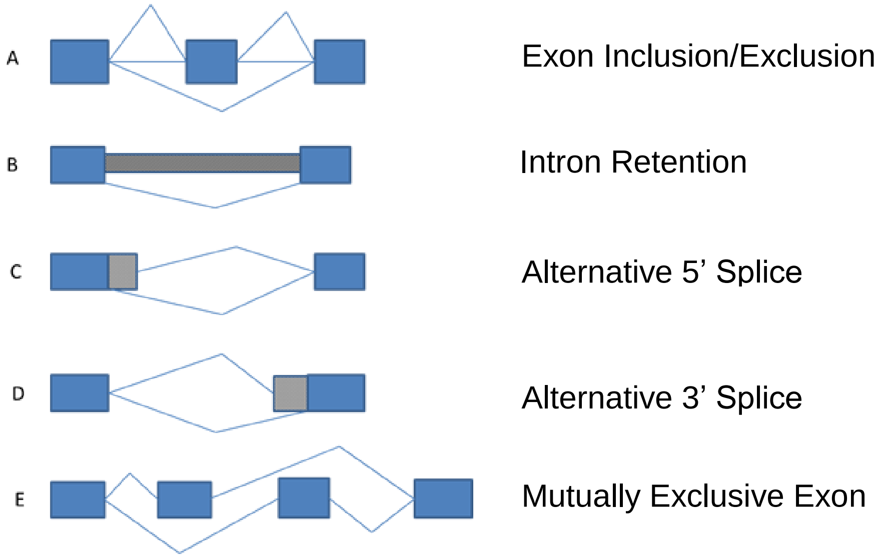


Figure 24: Splice Variant type.

## Material and Method

### Alternative splice detection through gap existence:

The Gap Existence approach requires certain pre-requisites. Data from a NGS such as Illumina must be *de novo* assembled into a transcriptome reference. Here we used Trinity transcriptome assembler [32]. After that assembly, the genes can be identified. In order to achieve this, we used gene identification techniques such as *ab initio* identification of open reading frames (ORFs) by ESTScan [100] and a BLASTx (Basic Local Alignment Search Tool) [33] nr database search. Figure 25 summaries Splice Variants Detection through Gap existence.

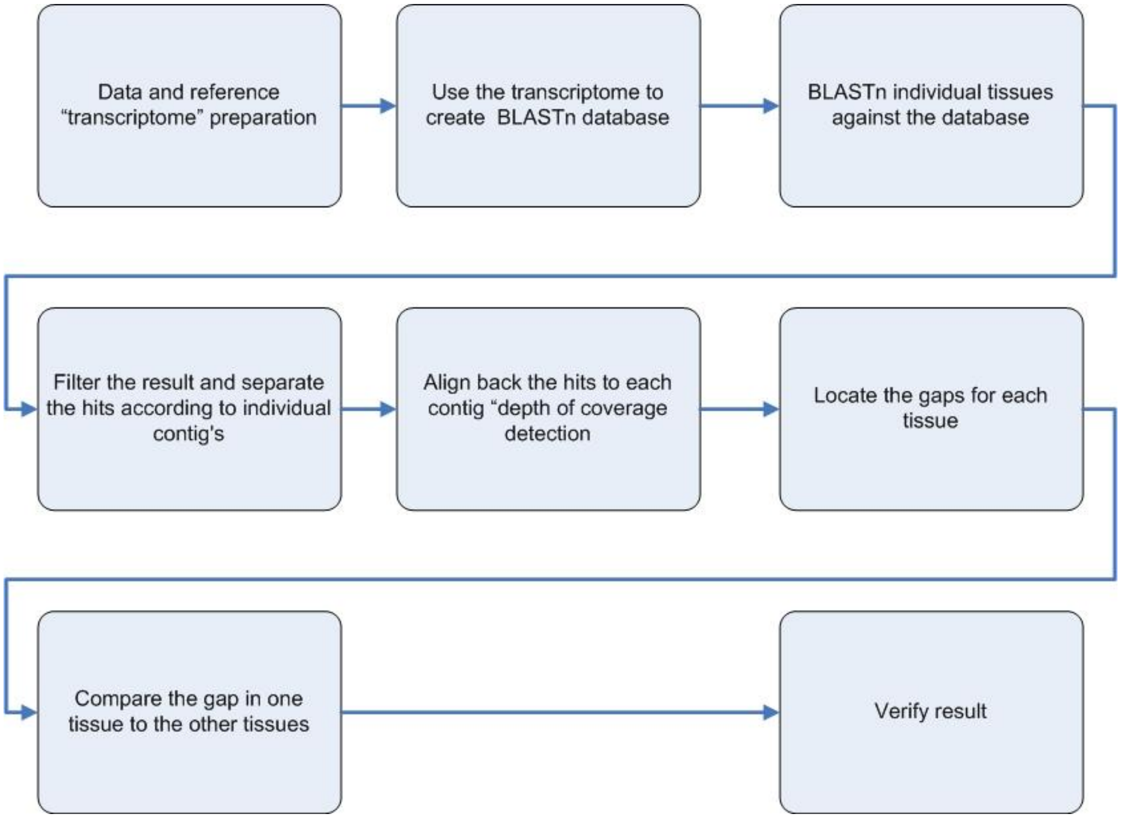

Data and reference "transcriptome" preparation
Use the transcriptome to create  BLASTn database
BLASTn individual tissues against the database
Filter the result and separate the hits according to individual contig's
Align back the hits to each contig "depth of coverage detection
Locate the gaps for each tissue
Compare the gap in one tissue to the other tissues
Verify result

Figure 25: Splice Variants Detection pipeline.

**Data mapping and alignment**

Reads from each tissue aligned against the transcriptome reference using BLASTn (Table 18). The main function of BLAST is to map each read to a transcript database. A table format will be used to retrieve the following information from the BLASTn searches: the name of the gene, E-value, how many nucleotides were matched, how many gaps (nucleotides that did not match), and the start and end of the nucleotide for both the query and the reference "transcript".

Table 18: Result of BLASTn for each tissue.

| Tissue | # of reads | # of hits befor filtering | # of hits after filtering |
|---|---|---|---|
| Brain | 84,816,430 | 39,178,259 | 27,729,233 |
| Fat | 93,546,068 | 40,882,764 | 28,823,582 |
| Gill | 92,670,670 | 43,292,677 | 29,411,937 |
| Head kidney | 92,168,818 | 43,714,406 | 31,380,348 |
| Intestine | 91,613,688 | 41,451,989 | 28,155,209 |
| Kidney | 89,642,288 | 42,822,877 | 29,736,223 |
| Liver | 85,281,910 | 47,029,190 | 37,168,536 |
| Red Muscle | 93,641,068 | 45,151,981 | 32,950,143 |
| Skin | 87,743,778 | 43,408,701 | 32,765,917 |
| Spleen | 93,532,200 | 41,794,723 | 28,770,351 |
| Stomach | 91,231,186 | 48,742,586 | 36,610,242 |
| Testis | 85,389,746 | 39,221,647 | 27,542,037 |
| White Muscle | 86,643,770 | 42,574,543 | 34,109,987 |

A C++ object-oriented program used the BLASTn result to create a linked list for each transcript. The program generates an output file that holds a vector that contains frequency information ("depth of coverage") for each transcript. This step performed on each tissue separately.

**Gap detection process**

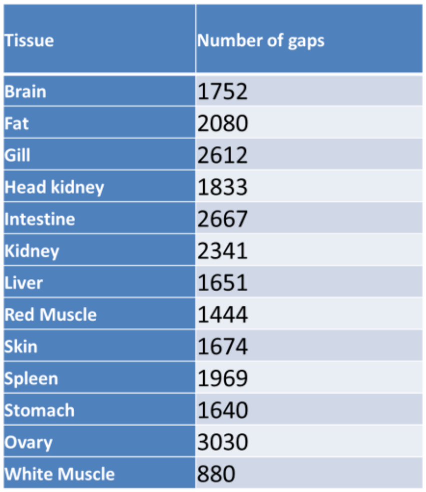
The vector file is used to determine possible gaps for each tissue (Table 18) where a sliding window is used to determine if the gap is present or not using the following criteria:

1. If the depth of coverage drop below a specific threshold (*e.g.*, 5 reads) a possible edge can be called.

2. The average of the previous 50 positions from the left of the edge and average of 20 positions to the right of the edge are calculated.
3. If the ratio of the left average to the right average of the edge is higher than 10 fold, the edge is considered as a gap start.
4. Similarly, the right edge side will be detected when the depth of coverage goes above certain threshold (*e.g.*, 5 reads).
5. The average of the previous 20 positions from the left of the edge and average of 50 positions to the right of the edge are calculated.
6. If the ratio between averages, right and left of the edge is higher than 10 fold, the edge is considered as a gap end.
7. For the gap to be considered, it has to be longer than 80 nt, shorter than 500 nt and 33% of the transcript size, where these values can be changed based on the average exon size of the studied species using input argument (Figure 26).

A total of 25,573 gaps were detected from the thirteen studied tissues [51] (Table 19).

Table 19: Number of putative gaps for each tissue.



| Tissue | Number of gaps |
|---|---|
| Brain | 1752 |
| Fat | 2080 |
| Gill | 2612 |
| Head kidney | 1833 |
| Intestine | 2667 |
| Kidney | 2341 |
| Liver | 1651 |
| Red Muscle | 1444 |
| Skin | 1674 |
| Spleen | 1969 |
| Stomach | 1640 |
| Ovary | 3030 |
| White Muscle | 880 |

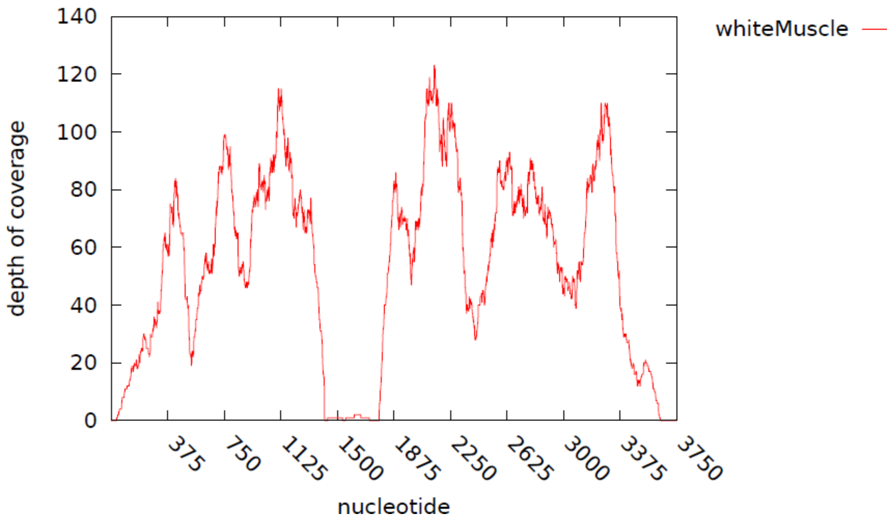


Figure 26: Example of a gap detected in white Muscle based on a *de novo* assembled reference from 13 tissues

## Gap validation

The final step was to test if the gap is valid and if it's a splice variant candidate or not by checking each gap in one tissue against other tissues (Figure 27). With a minimum depth and width of coverage to be determined, if the gap exists in one tissue but does not exist in another tissue(s), the gap considered a good candidate of a splice variance event. Some splice variant events will be validated by qPCR technique.

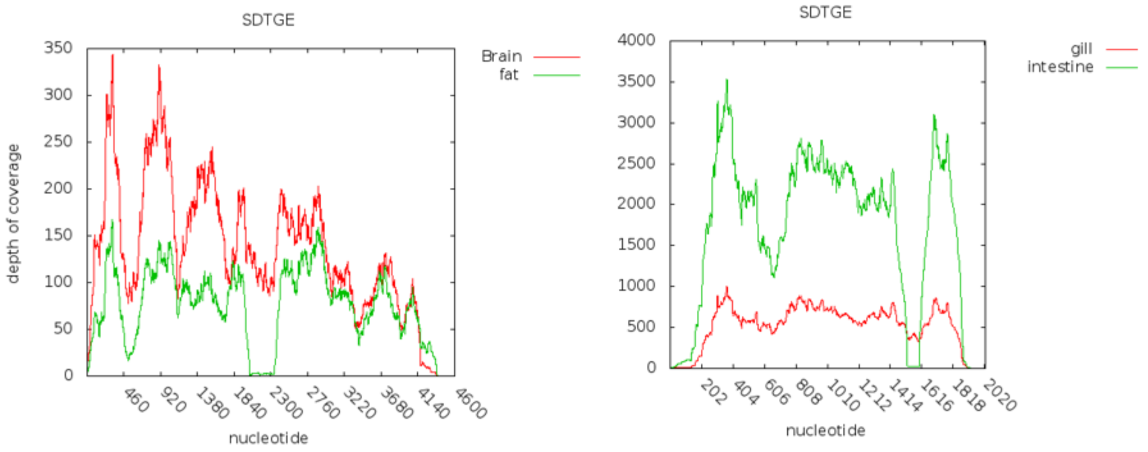


Figure 27: Left Coverage comparison between brain and fat tissue. Right Coverage comparison between gill and intestine

## Result and Conclusion

### Software Validation

To validate the functionality of the process and the program, a test data was created and tested to determine if the program can detect  gaps, can distinguish between different gap locations, and whether the gap internal, at the beginning, or at the end of the contig. Test data was created by randomly selecting 28 contigs. For each one of these contigs a depth of coverage was assigned. A gap was inserted in each of these contigs, some of the contigs have internal gap other has external “beginning or end” gaps. The program was able to detect all the gaps except two out of 28. Figure 28 show the results with different gap locations. The reason why the program couldn’t detect these two gaps is that we artificially concatenated all contigs together into one string. One of the undetected gaps was at the end of the last and the second gap was at the beginning it. The algorithm is setup to detect a gap between two boundaries of coverage only at this point so the first and last contig has no boundary from the left for the first contig in the linked list and no boundary from the right for the last contig in the linked list.

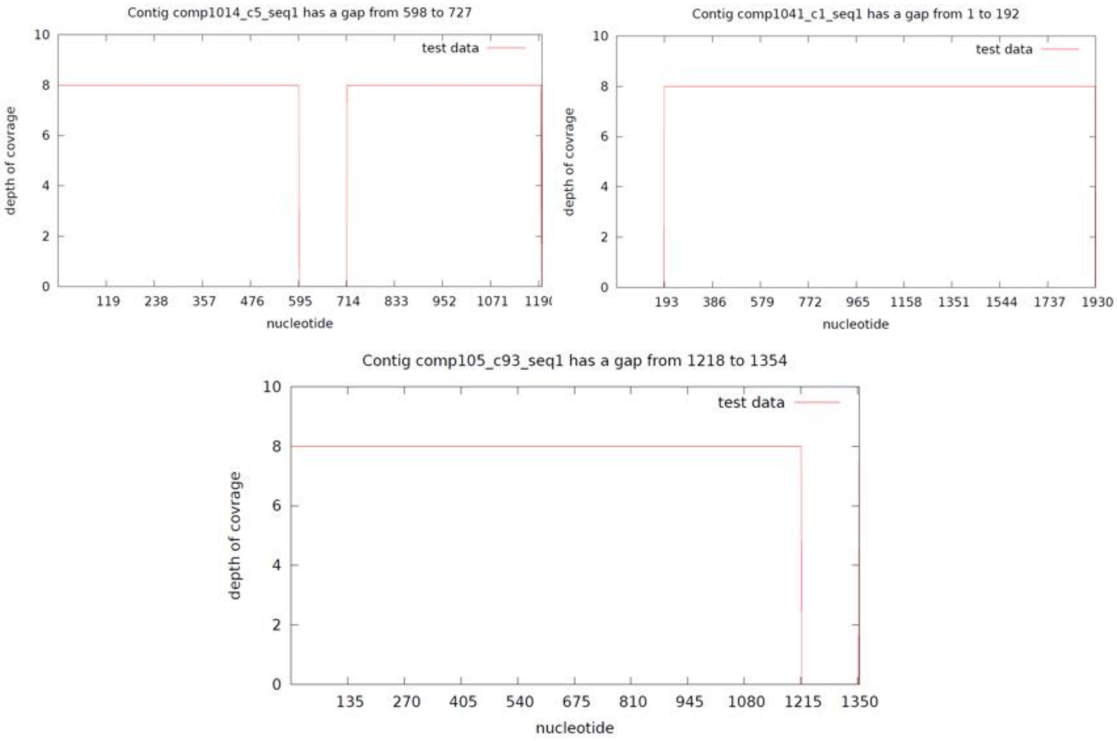


Figure 28: Gap location: upper left) internal gap. Upper right) beginning gap. Lower) end gap detected by the pipeline

**Conclusion:**

From the above results we can conclude that this method can detect splice variance. However it’s being improved upon. The next step is to run the program for the rest of the tissues and compare tissues against each other to determine possible splice variants. After this has been done, a random splice variance should be verified using qPCR.

### Future work

Software could be improved by combining the scripts into one package. For each contig that has a gap, a method need to be added to check the depth and average of coverage. The gap detection process could be improved by using individual contig instead of linked list. Visualization could be improved by using OpenGL or any similar visualization software to allow user Interaction, and the ability to select individual contig.