

**GENOME ANNOTATION AND ROLE OF NON-CODING RNAS IN DISEASE  
RESISTANCE, GROWTH AND MUSCLE QUALITY TRAITS IN RAINBOW  
TROUT**

By

Bam Dev Paneru

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy in Molecular Bioscience

Middle Tennessee State University

August 2017

Dissertation committee:

Dr. Mohamed (Moh) Salem, Chair

Dr. Anthony L. Farone

Dr. Mary Farone

Dr. Sarah Bergemann

Dr. Jason Jessen

## DECLARATION

I want to declare that all the contents presented are my original works performed in collaboration with other colleagues from Middle Tennessee State University, West Virginia University and National Center for Cool and Cold-Water Aquaculture (NCCCWA). I was the main contributor to the materials presented in chapter 3, 4 and 5; and significantly contributed to the chapters 1 and 2. Some of these works have been done in collaboration with a student from Computational science (COMS) program at MTSU, who may use these in his dissertation.

1. Salem, Mohamed, Bam Paneru, Rafet Al-Tobasei, Fatima Abdouni, Gary H. Thorgaard, Caird E. Rexroad, and Jianbo Yao. "Transcriptome assembly, gene annotation and tissue gene expression atlas of the Rainbow trout." *PloS one* 10, no. 3 (2015): e0121778.
2. Al-Tobasei, Rafet, Bam Paneru, and Mohamed Salem. "Genome-wide discovery of long non-coding RNAs in Rainbow trout." *PloS one* 11, no. 2 (2016): e0148940.
3. Paneru, Bam, Rafet Al-Tobasei, Yniv Palti, Gregory D. Wiens, and Mohamed Salem. "Differential expression of long non-coding RNAs in three genetic lines of Rainbow trout in response to infection with *Flavobacterium psychrophilum*." *Scientific Reports* 6 (2016).
4. Bam Paneru, Rafet Al-Tobasei, Brett Kenney, Timothy D. Leeds and Mohamed Salem. RNA-Seq reveals microRNA expression signature and genetic polymorphism associated growth and muscle quality traits in Rainbow trout (Manuscript submitted).

5. Bam Paneru, Ali Ali, Brett Kenney and Mohamed Salem. A closer look at the muscle degradome in Rainbow trout: Functional interplay among lncRNAs, microRNAs and protein coding genes (Manuscript in preparation).

## ACKNOWLEDGMENTS

First, I would like to acknowledge my deepest respect and indebtedness to my PhD. dissertation advisor Dr. Mohamed Salem for his continuous guidance, support and encouragement throughout my PhD. research work. I would like to thank him for training me throughout my PhD work to solve the problems and grow as a professional research scholar.

My deepest appreciation and respect also goes to my PhD. dissertation committee members Dr. Sarah Bergemann, Dr. Jason Jessen, Dr. Anthony Farone and Dr. Mary Farone for guiding me with their expertise that played important role in accomplishment of this project.

I would like to thank my lab colleagues Ali Reda Eid Ali and Rafet Al-Tobasei for working together in several projects and sharing scientific ideas. I also want to acknowledge my collaborators Dr. Gregory D. Wiens, Dr. Yniv Palti, Dr. Timothy D. Leeds and Dr. Caird E. Rexroad from National Center for Cool and Cold Water Aquaculture (NCCCWA)/USDA. My sincere thanks also go to my collaborator Dr. Brett Kenney from West Virginia University (WVU).

I also want to acknowledge entire MTSU Molecular biosciences (MOBI) family and MTSU Biology department for their help and support during my stay as a doctoral student at MTSU.

Finally, yet importantly, I am grateful to my parents, siblings and wife for their constant support and encouragement during this PhD. project.

## ABSTRACT

In Rainbow trout, effort to annotate the genome reference is ongoing. While recently published trout genome has discovered large number of protein coding genes, many protein coding genes appear to be missing. In addition, non-coding RNAs, which occupy vast majority of the transcribed portion of genome, are not investigated. In the present study, we have sequenced and assembled the transcriptomes of lncRNA (long non-coding RNA) and mRNA to facilitate gene discovery, and have investigated the role of non-coding RNAs in disease resistance, muscle atrophy, growth and muscle quality traits.

By sequencing RNA from 13 vital tissues, we identified 44,990 protein-coding and 54,503 lncRNA genes in trout. While lncRNAs were discovered for the first time, 11,843 mRNA genes reported by us were missing in the previously assembled genome reference. A total of 556 lncRNAs were differentially expressed during *F. psychrophilum* infection. There was strong correlation between lncRNA expression and infection susceptibility of different Rainbow trout genetic lines. These lncRNAs showed correlated expression with immunity related protein-coding genes. In addition, 1,198 lncRNAs showed altered expression during sexual maturation associated muscle atrophy and their expression level correlated with the extent of skeletal muscle atrophy.

We also investigated association of microRNAs with growth parameters and muscle quality traits in Rainbow trout. Twenty-eight microRNAs showed significantly altered expression during sexual maturation associated muscle atrophy. Similarly, 90 microRNAs were differentially expressed between fish families with different phenotypes for 5 fish/muscle growth and quality traits: muscle mass, muscle fat content, muscle shear force (tenderness), muscle whiteness and whole body weight (WBW). Expression of 12

DE microRNAs chosen for ‘genotype-phenotype’ association correlated significantly with the phenotypes. In addition to microRNA expression, at least 72 single nucleotide polymorphisms (SNPs) either destroying or creating novel illegitimate microRNA target sites in protein coding genes explained significant variation in growth and muscle quality phenotypes.

The present study explores role of non-coding RNAs in regulation of important aquaculture traits in Rainbow trout and suggests that non-coding RNA mediated gene regulation plays a critical role in determining these phenotypes.

Key words: Rainbow trout, disease resistance, growth, muscle quality, long non-coding RNA, microRNA

## TABLE OF CONTENTS

<b>CONTENTS</b> .....	<b>...PAGE</b>
<b>LIST OF TABLES</b> .....	<b>x</b>
<b>LIST OF FIGURES</b> .....	<b>xiii</b>
<b>LIST OF APPENDICES</b> .....	<b>xvii</b>
<b>INTRODUCTION</b> .....	<b>1</b>
AQUACULTURE AND GENOMIC SELECTION IN FISH .....	1
AQUACULTURE PRODUCTION TRAITS.....	2
NON-CODING RNAS: LONG NON-CODING RNAS (LNCRNAS) AND MICRORNAS.....	4
<b>OBJECTIVES</b> .....	<b>7</b>
GENERAL OBJECTIVES .....	7
SPECIFIC OBJECTIVES.....	7
<b>CHAPTER I: TRANSCRIPTOME ASSEMBLY, GENE ANNOTATION AND TISSUE GENE EXPRESSION ATLAS OF THE RAINBOW TROUT</b> .....	<b>8</b>
ABSTRACT.....	8
INTRODUCTION .....	9
MATERIALS AND METHODS.....	13
RESULTS AND DISCUSSION.....	19
CONCLUSION.....	46
REFERENCES .....	47
<b>CHAPTER II: GENOME-WIDE DISCOVERY OF LONG NON-CODING RNA IN RAINBOW TROUT</b> .....	<b>52</b>

ABSTRACT.....	52
INTRODUCTION .....	53
MATERIALS AND METHODS.....	55
RESULTS AND DISCUSSION.....	59
REFERENCES .....	72
APPENDICES .....	77
<b>CHAPTER III: DIFFERENTIAL EXPRESSION OF LONG NON-CODING RNAS IN THREE GENETIC LINES OF RAINBOW TROUT IN RESPONSE TO INFECTION WITH <i>FLAVOBACTERIUM PSYCHROPHILUM</i>.....</b>	<b>79</b>
ABSTRACT.....	79
INTRODUCTION .....	80
MATERIALS AND METHODS.....	82
RESULTS AND DISCUSSION.....	87
CONCLUSION.....	110
REFERENCES .....	112
APPENDICES .....	116
<b>CHAPTER IV: MICRORNA EXPRESSION AND GENETIC POLYMORPHISM ASSOCIATION WITH GROWTH AND MUSCLE QUALITY TRAITS IN RAINBOW TROUT .....</b>	<b>124</b>
ABSTRACT.....	124
INTRODUCTION .....	124
MATERIALS AND METHODS.....	127
RESULTS AND DISCUSSION.....	133



CONCLUSION.....	154
REFERENCES .....	157
APPENDICES .....	162
<b>CHAPTER V: A CLOSER LOOK AT THE MUSCLE “DEGRADOME” IN RAINBOW TROUT: FUNCTIONAL INTERPLAY AMONG LNC-RNAS, MICRORNAS AND PROTEIN CODING GENES .....</b>	<b>169</b>
ABSTRACT.....	169
INTRODUCTION .....	170
MATERIALS AND METHODS.....	173
RESULTS AND DISCUSSION.....	178
CONCLUSION.....	202
REFERENCES .....	204
APPENDICES .....	209
<b>PROJECT CONCLUSION.....</b>	<b>215</b>
<b>PROJECT REFERENCES .....</b>	<b>219</b>

## LIST OF TABLES

TABLE.....	PAGE
<b>CHAPTER I</b>	
<b>Table 1:</b> cDNA library information and summary of the high-throughput sequencing yield.....	20
<b>Table 2:</b> Assembly statistics of Illunina paired-end data. ....	21
<b>Table 3:</b> Summary of BLASTx search analysis of Rainbow trout sequences against different model fish species with known reference genomes. ....	29
<b>Table 4:</b> Number of genes expressed in 13 Rainbow trout tissues at different RPKM threshold.....	39
<b>CHAPTER II</b>	
<b>Table 1:</b> Number of lncRNA predicted in at least 2 of the 4 datasets and final numbers after merging and removal of redundant sequences. ....	62
<b>Table 2:</b> Number of exons and average length of lncRNAs in different data sets.....	63
<b>CHAPTER III</b>	
<b>Table 1:</b> Comparison of differentially expressed lncRNA and protein coding genes in response to <i>Fp</i> infection. ....	91
<b>Table 2:</b> LncRNAs upregulated in all three genetic lines (> 2 fold) on 5th day post <i>Fp</i> challenge (top). LncRNAs showing highest fold change (> 100-fold) upon <i>Fp</i> infection in at least one genetic line relative to the two other genetic lines and their associated protein coding gene in genome (bottom).....	92
<b>Table 3:</b> Correlation between expression patterns of lncRNAs and their overlapping protein-coding genes ( $R^2 > 0.70$ ). ....	95

**Table 4:** Correlation between expression patterns of lncRNAs and their intergenic neighboring protein-coding genes (within < 50 kb and  $R^2 > 0.70$ ).....99

**Table 5:** Correlation between expression patterns of lncRNAs and some distantly located (> 50 kb or different chromosome) immune-relevant protein-coding genes ..... 101

#### **CHAPTER IV**

**Table 1:** Small RNA sequencing and annotation statistics of 22 samples used in the study.....136

**Table 2:** Correlation between microRNA expression level and phenotypic variation....141

**Table 3:** Cis-regulatory transcription factor binding motifs that exist in promoter sequences of differentially expressed (DE) microRNAs and their positively correlated target genes. ....150

**Table 4:** SNPs in microRNA recognition element seed site (MRESS) of target gene, allele frequency ratio of MRESS-destroying SNPs between high vs low ranked families of different muscle traits, and correlation between the SNP and phenotype. ....153

#### **CHAPTER V**

**Table 1:** DE microRNAs between atrophying muscle of gravid fish and normal skeletal muscle of sterile fish. Positive and negative value of fold change represent upregulation and downregulation respectively in atrophying skeletal. Fold change was considered significant at cut off:  $3 >$  or  $< -3$ , FDR-p-value  $< 0.01$ .....182

**Table 2:** Selected proteolytic genes highly upregulated in atrophying skeletal muscle of gravid female Rainbow trout relative to normal skeletal muscle of same-aged sterile Rainbow trout.....184

**Table 3:** DE lncRNAs and mRNAs sharing microRNA binding sites and expression correlation between them.....195

**Table 4:** Differentially expressed (DE) lncRNA-DE mRNA physical interaction statistics and expression correlation between them (top). Physical interaction statistics between DE lncRNA and proteome of DE protein coding genes (bottom). .....198

## LIST OF FIGURES

FIGURE.....	PAGE
<b>CHAPTER I</b>	
<b>Figure 1:</b> Distribution of contig ( $\geq 500$ nt) length of a Rainbow trout Illumina/Trinity transcriptome assembly.....	21
<b>Figure 2:</b> Comparison of total number of sequenced bases (A), total number of contigs (B), number of long contigs ( $\geq 500$ bp) (C), and average length of contigs (D) obtained from Illumina, Sanger-based, and 454-pyrosequencing techniques. ....	23
<b>Figure 3:</b> Number of UniGenes of model fish species and Rainbow trout UniGenes that are available in the NCBI database. ....	26
<b>Figure 4:</b> Gene Ontology (GO) assignment (2nd level GO terms) of the Rainbow trout of 13 lanes of Illumina Trinity assembly. ....	32
<b>Figure 5:</b> Gene Ontology (2 <sup>nd</sup> level GO terms) comparison of Rainbow trout and Nile tilapia.....	34
<b>Figure 6:</b> Number of tissue-specific genes predicted in different tissues. ....	36
<b>Figure 7:</b> Distribution of gene abundance in various tissues. ....	40
<b>Figure 8:</b> Transcript abundance of tissue-specific genes in various tissues. ....	40
<b>CHAPTER II</b>	
<b>Figure 1:</b> Bioinformatics pipeline used in prediction of Rainbow trout lncRNAs. ....	60
<b>Figure 2:</b> Distribution of sequence length and number of exons in lncRNAs compared to protein-coding transcripts in Rainbow trout. ....	63
<b>Figure 3:</b> Classification of lncRNAs based on their intersection with protein-coding genes and number of lncRNAs in each class. ....	65

**Figure 4:** RPKM comparison of protein-coding genes and lncRNAs (left). Number of tissue-specific lncRNAs and protein-coding genes in various tissues (right).....66

**Figure 5:** Distribution of lncRNA expression in various tissues .....68

### CHAPTER III

**Figure 1:** 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.....97

**Figure 2:** 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. ....103

**Figure 3:** 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. ....104

**Figure 4:** Comparison of transcriptome abundance of selected lncRNAs among genetic lines after infection with *Fp* .....106

**Figure 5:** Comparison of transcriptome abundance of selected lncRNAs between day 1 and day 5 of *Fp* injection in each genetic line. ....108

### CHAPTER IV

**Figure 1:** Phenotypic difference for WBW and 4 muscle quality traits (muscle yield, crude-fat content, shear force and FWI) of top 4 high ranked and 4 low ranked families (5 fish/family) of selectively-bred trout at ca. 13 months post-hatch. A: whole body weight

(wbw), B: muscle yield, C: muscle crude fat content, D: muscle shear force and E: muscle whiteness index.....134

**Figure 2:** Heat map of fold change of differentially expressed (DE) microRNAs between high vs low ranked families of various traits (left) and Venn-diagram showing shared DE microRNAs between different traits (right).....138

**Figure 3:** Enrichment map and enriched gene pathways of predicted microRNA targets.....143

## CHAPTER V

**Figure 1:** Comparison of different muscle phenotypes between atrophying skeletal muscle from gravid diploid (2N) fish and normal skeletal muscle from sterile triploid (3N) fish.....179

**Figure 2:** Heat map of DE lncRNAs (left) and protein coding genes (right) between atrophying muscle of gravid fish and normal skeletal muscle of sterile fish. Value of color limit represents normalized expression values (Z scores). Fold change in gene expression was considered significant at: FDR-p-value < 0.01, fold change: > 3 or < -3. Darker red and lighter red colors represent higher and lower level of expression respectively. ....181

**Figure 3:** Transcript level of different classes of DE genes during pre-spawning and spawning months in skeletal muscle of diploid gravid fish: all ubiquitinating genes combined (A), all autophagy related genes combined (B), atrogen-1 isoforms (C), cathepsin D isoforms (D), all development related genes combined (E), all collagen and extracellular matrix related genes combined (F), all upregulated lncRNAs combined (G) and all downregulated lncRNAs combined (H). Note that expression level of each gene in gravid fish (2N) was normalized by expression level of respective gene in sterile fish (3N) ....187

**Figure 4:** Heat map showing tissue specific expression pattern of DE lncRNAs (left), tissue specific expression pattern of DE mRNAs (middle) and temporal expression pattern of DE lncRNAs and DE mRNAs during pre-spawning and spawning months (right). Value of color limit represents normalized expression values (Z scores). Darker red and lighter red colors represent higher and lower level of expression respectively. ....191

**Figure 5:** Gene expression network of DE lncRNAs (blue node), DE mRNAs (green node) and microRNAs (pink node) ( $R > 0.97$  or  $< -0.97$ ) .....201



## LIST OF APPENDICES

APPENDIX.....	PAGE
<b>CHAPTER II</b>	
<b>Appendix A:</b> Cluster of lncRNA and protein coding genes based on their expression values across 13 tissues. Clusters were generated at threshold of $R^2 > 0.97$ .....	78
<b>CHAPTER III</b>	
<b>Appendix A:</b> Summary statistics of 24 RNA seq libraries with different genetic lines, time, infection status and tank replicate.....	117
<b>Appendix B:</b> Fold change comparison of selected differentially expressed genes by RNA-Seq and real time PCR .....	118
<b>Appendix C:</b> Classification of DE lncRNA in response to <i>Fp</i> challenge based on their intersection with protein-coding genes and number of lncRNAs in each class.....	119
<b>Appendix D:</b> Strand specific PCR Method used in validation of strand orientation of some of lncRNAs transcripts relative to their protein coding loci counterparts.....	120
<b>Appendix E:</b> Differentially expressed (DE) lncRNAs conserved in Atlantic salmon....	121
<b>Appendix F:</b> Novel lncRNAs specific to resistant or susceptible lines, and their expression correlation with protein coding gene .....	122
<b>Appendix G:</b> Novel lncRNAs specific to resistant and susceptible lines and their relative expression between various comparisons .....	123
<b>CHAPTER IV</b>	
<b>Appendix A:</b> Correlation between different growth and muscle quality traits.....	163
<b>Appendix B:</b> Real time PCR validation of fold change of 12 microRNAs DE between high and low muscle yield group .....	164

<b>Appendix C:</b> Enriched gene pathways (in biological process categories) among target genes of DE microRNAs between high vs low muscle yield .....	165
<b>Appendix D:</b> Enriched gene pathways (in biological process categories) among target genes of DE microRNAs between high vs low crude fat content.....	166
<b>Appendix E:</b> Shear force associated microRNAs and their target genes coding for collagen and collagen regulators.....	167
<b>Appendix F:</b> Enriched gene pathways (in biological process categories) among target genes of DE microRNAs between high vs low fillet whiteness index .....	168
<b>CHAPTER V</b>	
<b>Appendix A:</b> Selected DE genes involved in fat/amino acid biosynthesis, amino acid transport/catabolism, muscle structure and myogenesis.....	210
<b>Appendix B:</b> Differentially expressed (DE) transcription factors and/or transcription regulators in skeletal muscle between gravid and sterile fish.....	211
<b>Appendix C:</b> Selected targets of downregulated microRNAs that are involved in muscle proteolysis.....	212
<b>Appendix D:</b> Selected overlapping differentially expressed lncRNA-protein coding gene pairs and their expression correlation .....	213
<b>Appendix E:</b> Selected neighboring (< 50 KB distance) differentially expressed lncRNA-protein coding gene pairs and their expression correlation .....	214

## INTRODUCTION

### AQUACULTURE AND GENOMIC SELECTION IN FISH

In the United States, seafood imports have been increasing continuously. From 1990 to 2014, cost of the US seafood import have risen from \$5 billion to \$20.2 billion (NOAA 2014), but there was almost no change in seafood export during the period. Due to this unbalanced trend in export and import, US trade deficit in seafood has been expanding rapidly in last 3 decades. In 2014, the US seafood trade deficit was \$14.9 billion, (NOAA 2014). Seafood production with aquaculture is necessary to meet the increasing demand of growing population as natural fresh water and marine fisheries in North America has been depleted due to overfishing, exploitation and habitat impact. Food production through aquaculture is rapidly growing worldwide in recent years with net production of 73.8 million tonnes worth of \$160.2 billion in 2014 (FAO 2016). In 1980, only 9% of the fish consumed worldwide came from aquaculture production but it reached ~50% in 2014 (FAO 2016). However, for efficient production there is a need of high quality, fast growing and disease resistant fish.

The lack of genetically improved fish is the major hindrance to boost aquaculture production (Gjedrem 2008b). Unlike farm animals such as chicken and cow, most of the seafood production (> 90 %) comes from genetically unimproved strains (Gjedrem 2008a). In fish, the rate of genetic variation in growth is 20–35% compared to 7–10% in terrestrial food animals (Gjedrem 1997). In selectively bred Rainbow trout population developed at National Center for Cool and Cold Water Aquaculture (NCCCWA) (to be used in this study), genetic variation has been reported for several traits including disease resistance (Vallejo et al. 2016) and growth (Salem et al. 2012). These findings give hope for

improvement of fish growth and reproduction through genetic selection approach (Gjedrem 2010).

Functional genomics approach have been widely applied to expedite commercial production of domesticated animals including chickens, cows, sheep and other animals. However, the contribution of genomic selection in fish breeding is still limited. In recent years, several genetic markers were identified that were associated with disease resistance, growth and other production traits in different aquaculture animals (Houston et al. 2009, Salem et al. 2012). Most of these approaches have focused heavily on protein coding genes to identify genetic markers for selection. In most eukaryotic genome studies, the majority of the genome is transcribed to generate non-coding RNAs (Derrien et al. 2012) that play an important gene regulatory role in the cell. Therefore, the investigation of non-coding RNAs may lead to identification of suitable genetic marker for breeding purpose.

### **AQUACULTURE PRODUCTION TRAITS**

Salmonid fish are susceptible to many bacterial and viral pathogens that cause significant loss of aquaculture production annually (Asche et al. 2009). BCWD (bacterial cold water disease) caused by *Flavobacterium psychrophilum* (*Fp*), causes significant mortalities in cultured trout and salmon every year globally (see review (Nematollahi et al. 2003). This pathogen affects very young fish that do not have a developed immune system. For example, high mortality rates and post-infection complications occur in fish infected with *F. psychrophilum*. The major hindrances in the control of BCWD include lack of effective vaccine and chemotherapeutic agents, wide geographic distribution of the pathogen (Carson and Schmidtke 1995) and the ability of the bacteria to survive in harsh environmental conditions. Recently, live attenuated vaccine provided protection under

laboratory conditions, but the commercial use of live culture is an environment safety issue (Gómez et al. 2014). Strategies of minimizing loss due to BCWD include improving immune system of the host by selective breeding (Gjedrem 2005) and/or development of effective vaccine. However, there is little progress toward this goal as pathogenesis of *F. psychrophilum* is not well studied. The molecular characterization of the host response to *Flavobacterium* infection will help to identify the genes implicated in pathogenesis and/or defense against BCWD.

In addition to disease resistance, the growth rate and muscle fillet quality traits determine the profitability of aquaculture industries. While faster growth of fish reduces the time and cost of production, muscle fillet qualities determine the customer satisfaction and market value. Muscle shear force, muscle crude fat content and fillet whiteness/color are such important muscle quality traits in aquaculture salmonids. Limited collagen and extracellular matrix results in softer flesh and gaping in fish muscle. Such softer fillet reduces the value of fish fillet for processing (Michie 2001). Similarly, flesh color in many salmonids such as salmon determines customer's willingness to pay premium price for the product (Steine, Alfnes and Rørå 2005). These muscle quality traits are interrelated (Mørkøre et al. 2001) and are determined by both genetic and non-genetic factors. Non genetic factors such as diet (Jacob et al. 1995) and sexual maturation (Salem et al. 2013) have been extensively studied in relation to fillet quality traits. However, limited effort has been made to identify the genes that determine muscle quality traits (some examples are given in following 'non-coding RNA' section). Identification of genes that affect these muscle quality traits will help develop suitable genetic markers for selection.

## **NON-CODING RNAS: LONG NON-CODING RNAS (LNCRNAS) AND MICRORNAS**

In higher eukaryotes, the vast majority of the transcribed genome gives non-protein coding RNA transcripts (Clark et al. 2013). Various classes of non-protein coding RNAs include long non-coding RNA (lncRNA), rRNA, microRNA, piRNA and small non-coding RNAs such as small interfering RNAs.

LncRNAs are longer than 200 nucleotides (Rinn and Chang 2012, Zhu and Wang 2012) which are expressed at lower levels (Al-Tobasei, Paneru and Salem 2016) and exhibit strict tissue-specific or developmental stage associated the expression pattern (Prasanth et al. 2005, Mercer et al. 2008, Cabili et al. 2011). They regulate transcription as well as post transcriptional events at least by acting as decoy, molecular scaffold or guide (Pandey et al. 2008, Kino et al. 2010, Tsai et al. 2010). Such gene regulatory role of lncRNAs significantly impacts several cellular and physiological processes including immunity (Carpenter et al. 2013), development (Fatica and Bozzoni 2014), metabolism (Kornfeld and Brüning 2014), myogenesis (Ballarino et al. 2015) and muscle atrophy (Cabianca et al. 2012) in different animals. However, there are no previous studies aimed at identification of lncRNAs associated with performance traits in salmonids. In-depth investigations of lncRNAs in association with performance traits will help to understand lncRNA-mediated regulation of important production traits in Rainbow trout.

MicroRNA is another class of non-coding RNA which plays important role in gene regulation. Mature sequences of microRNAs (~22 nts) binds 3' -UTR of mRNA by complementary base pairing that leads to downregulation of the gene by translation suppression (Olsen and Ambros 1999), target mRNA cleavage (Bagga et al. 2005) or

deadenylation (Wu, Fan and Belasco 2006). In humans, ~30 percentage of genes are regulated by microRNAs (Lewis, Burge and Bartel 2005) which suggests their important role in phenotype (Zhang, Wang and Pan 2007). MicroRNA-mediated gene regulation has versatile roles in disease, development, myogenesis and others (Wang 2013). In Nile tilapia, mir-206 regulates skeletal muscle growth by targeting insulin-like growth factor-1 (Yan et al. 2013b). MicroRNAs mir-1 and mir-133 control more than 50% of microRNA-mediated muscle gene regulations in Zebra fish (Mishima et al. 2009). Myogenic transcription factor myoD in Mandarin fish and Nile tilapia is regulated by mir-143 and mir-203b respectively (Yan et al. 2013a, Chen et al. 2014). Several microRNAs including let-7, mir-19 and mir-130 are reported to be differentially expressed during skeletal muscle development in fish (Johnston et al. 2009).

In trout, microRNAs expressed during embryonic development have been identified previously (Ramachandra et al. 2008). However, so far association of microRNAs with growth and muscle quality traits has not been investigated in Rainbow trout. Investigation of microRNAs associated with growth and muscle quality traits will help understand how post-transcriptional gene regulation determines the growth and muscle quality traits in Rainbow trout.

NCCCWA has been conducting family based phenotypic selection of Rainbow trout for nearly a decade, and since this time, these trout families have gone multiple generations of phenotypic selections for growth and disease resistance. One population selected for BCWD resistance has gone 5 generations of family based phenotypic selection. In this population, significant differences in survival rates after BCWD has been observed among families that ranges from ~29% to ~94% for susceptible and resistant line

respectively (Silverstein et al. 2009, Marancik et al. 2014). Similarly, growth selected line has gone five generations of family based phenotypic selection for growth traits. At ~13 month post-hatch, phenotypes were statistically different between high ranked and low ranked families ( $P < 0.05$ ): WBW ( $1221.6\text{g} \pm 84.3$  vs.  $502.1 \pm 28.0\text{g}$ ); muscle yield of WBW (%) ( $50.9\% \pm 1.6$  vs.  $43.3\% \pm 2.3$ ); crude-fat ( $9.2\% \pm 1.2$  vs.  $4.8\% \pm 1.3$ ); shear force force;  $539.6 \pm 12.3$  vs.  $310.0 \pm 49.2$ ); and FWI (fillet whiteness index) ( $44.7 \pm 0.8$  vs.  $41.2 \pm 0.4$ ).

In the current project, the aims are to discover protein coding and lncRNAs to annotate the genome reference and to investigate the protein coding and non-coding genes that explain the variation in disease resistance, growth and muscle quality phenotypes in these selectively bred trout populations. Identification of underlying genes that contribute to the difference in phenotypes will help identify suitable genetic markers for breeding purpose. The long term goal of this project is to identify suitable genetic markers predictive of better performance traits to improve commercial aquaculture production.



## OBJECTIVES

### GENERAL OBJECTIVES

- To annotate the Rainbow trout genome for protein coding genes and long non-coding RNAs by sequencing RNA from diverse tissues.
- To identify the role of lncRNAs and microRNAs in disease resistance, growth and muscle quality traits in Rainbow trout.

### SPECIFIC OBJECTIVES

- To sequence Rainbow trout transcriptome from 13 vital tissues to discover protein coding transcripts.
- To discover lncRNA transcripts expressed in Rainbow trout using computational and experimental approaches.
- To identify differentially expressed lncRNAs between selectively bred resistant, control and susceptible line of Rainbow trout in response to infection with *Flavobacterium psychrophilum*, and to study genomic co-localization and expression correlation between lncRNAs and immunity related protein coding genes.
- To identify association of microRNA expression and genetic variation in microRNA binding sites of target genes with growth and muscle quality traits in Rainbow trout.
- To identify lncRNAs, microRNAs and protein coding genes associated with sexual maturation associated muscle atrophy in Rainbow trout, and to elucidate functional links among three classes of genes.

## CHAPTER I

### TRANSCRIPTOME ASSEMBLY, GENE ANNOTATION AND TISSUE GENE

#### EXPRESSION ATLAS OF THE RAINBOW TROUT

Salem, M., B. Paneru, R. Al-Tobasei, F. Abdouni, G. H. Thorgaard, C. E. Rexroad & J.

Yao (2015) Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout. *PLoS One*, 10, e0121778.

#### ABSTRACT

Efforts to obtain a comprehensive genome sequence for Rainbow trout are ongoing and will be complemented by transcriptome information that will enhance genome assembly and annotation. Previously, transcriptome reference sequences were reported using data from different sources. Although the previous work added a great wealth of sequences, a complete and well-annotated transcriptome is still needed. In addition, gene expression in different tissues was not completely addressed in the previous studies. In this study, non-normalized cDNA libraries were sequenced from 13 different tissues of a single doubled haploid Rainbow trout from the same source used for the Rainbow trout genome sequence. A total of ~1.167 billion paired-end reads were de novo assembled using the Trinity RNA-Seq assembler yielding 474,524 contigs > 500 base-pairs. Of them, 287,593 had homologies to the NCBI non-redundant protein database. The longest contig of each cluster was selected as a reference, yielding 44,990 representative contigs. A total of 4,146 contigs (9.2%), including 710 full-length sequences, did not match any mRNA sequences in the current Rainbow trout genome reference. Mapping reads to the reference genome identified an additional 11,843 transcripts not annotated in the genome. A digital gene expression atlas revealed 7,678 housekeeping and 4,021 tissue-specific genes. Expression

of about 16,000–32,000 genes (35–71% of the identified genes) accounted for basic and specialized functions of each tissue. White muscle and stomach had the least complex transcriptomes, with high percentages of their total mRNA contributed by a small number of genes. Brain, testis and intestine, in contrast, had complex transcriptomes, with a large number of genes involved in their expression patterns. This study provides comprehensive de novo transcriptome information that is suitable for functional and comparative genomics studies in Rainbow trout, including annotation of the genome.

## **INTRODUCTION**

Rainbow trout (*Oncorhynchus mykiss*), a member of *Salmonidae* family, is a native species of the Pacific coasts of North America and Russia. 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 (Papanastasiou, Georgaka and Zarkadis 2007), carcinogenesis (Williams 2012), physiology (Giaquinto and Hara 2008), toxicology (Patel et al. 2006, Welsh et al. 2008), microbial pathogenesis (Speare, Arsenault and Buote 1998), and ecology (Davidson 2012).

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 database (Palti et al. 2009, Berthelot et al. 2014). 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 (Salem et al. 2012, Devisetty et al. 2014, Marancik et al. 2014,

Salgado et al. 2014, Liu et al. 2015). 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 (Salem et al. 2010). In addition, another reference transcriptome was sequenced in our laboratory representing responses to several stressors affecting the aquaculture production environments (Sánchez et al. 2011). Further, a transcriptome sequence of the anadromous steelhead (*Oncorhynchus mykiss*) was recently reported (Fox et al. 2014). While the first study aimed toward 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. More successfully, a draft of the genome sequence has been assembled from a single homozygous doubled haploid YY male from the same clonal line (Berthelot et al. 2014). 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 (Berthelot et al. 2014). To improve annotation of the under-development trout genome sequence and estimate coverage of assembly, 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 (Alkan, Sajjadian and Eichler 2011). In addition, about 50% of the genes in salmonids are duplicated (Bailey, Poulter and Stockwell 1978), which makes de novo assembly and annotation of the transcriptome difficult and complicates SNP/variant discovery (Ryynänen and Primmer 2006). 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 (Berthelot et al. 2014). This doubled haploid clonal line, which contains two identical copies of each chromosome, was previously established by chromosome set manipulation techniques (Young et al. 1996, Robison 1999) and has been used in sequencing the Rainbow trout genome (Berthelot et al. 2014). Recently, dramatic improvements in genome assembly of *Takifugu rubripes* were achieved by using doubled-haploid individuals compared to the wild types (Zhang et al. 2014).

Housekeeping genes were initially described as genes which are always expressed in the cell. Later, this concept has been refined to refer to genes with constitutive expression that maintain normal cellular functions (Butte, Dzaou and Glueck 2001). In contrast, tissue-specific genes are transcripts whose functions and expressions are favored in specific tissue/cell types (Xiao et al. 2010). 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 (Su et al. 2004, Tomancak et al. 2007, Fowlkes et al. 2008). In fish, gene expression atlases were characterized in only few model species (Kudoh et al. 2001, Henrich et al. 2005). 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. 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 (Berthelot et al. 2014).

## **MATERIALS AND METHODS**

### **Ethics statement**

The fish sacrificed 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 (Young et al. 1996, Robison 1999). First generation homozygous Rainbow trout were produced by androgenesis using gamma irradiation of eggs prior to fertilization (Young et al. 1996, Robison 1999) and then gynogenesis by blockage of first cleavage (Scheerer and Allendorf 1986, Young et al. 1996, Robison 1999). 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 (Scheerer and Allendorf 1986).

### **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 to the manufacturer's procedure as previously described (Salem et al. 2010).

### **Illumina paired-end sequencing**

Construction of RNA-Seq libraries and sequencing on an Illumina Genome Analyzer Ix 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<sup>+</sup> 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 Ix according to the manufacturer's protocols (Illumina, San Diego, CA). 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 (Grabherr et al. 2011). 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 = 2`, `min_iso_ratio = 0.05` and `glue_factor = 0.05`). Transcript reconstruction from a deBruijn graph by Butterfly was also performed at default parameters (`max_number_of_paths_per_node = 10`, `group_pairs_distance = 500`, `path_reinforcement_distance = 75`, `lenient_path_extension = 1`). Trinity contigs that were more than 500 nucleotides long were BLAST searched 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 (Iseli, Jongeneel and Bucher 1999) to search for an open reading frame (ORF), which distinguishes coding and non-coding sequences (Iseli et al. 1999, Lottaz et al. 2003). 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 (Götz et al. 2008). Blast2GO analysis provides a controlled vocabulary to describe gene product characteristics in three independent ontologies: biological process, molecular function, and cellular (Ashburner et al. 2000).

### **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 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 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 was 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 RPKM threshold of 0.5). A similar method of comparing the composition and complexity of tissue-specific transcriptomes was employed by Jongeneel and coworkers (Jongeneel et al. 2005). 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  $1.00E-$

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

Alternative transcription/splicing events were assessed using the Bowtie2, TopHat and Cufflinks software package (Langmead and Salzberg 2012, Trapnell et al. 2012). First, a Rainbow trout draft genome assembly was downloaded from <http://www.genoscope.cns.fr/trout-ggb/data/> (Berthelot et al. 2014). 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 (Min et al. 2005) 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 (Kent 2002) 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 (Palti et al. 2012, Berthelot et al. 2014) and in our previous transcriptome assembly (Salem et al. 2010). 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 1. 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 (Martin and Wang 2011).

**Table 1:** cDNA library information and summary of the high-throughput sequencing yield.

	<b>Tissue</b>	<b>Number of reads</b>
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

doi:10.1371/journal.pone.0121778.t001

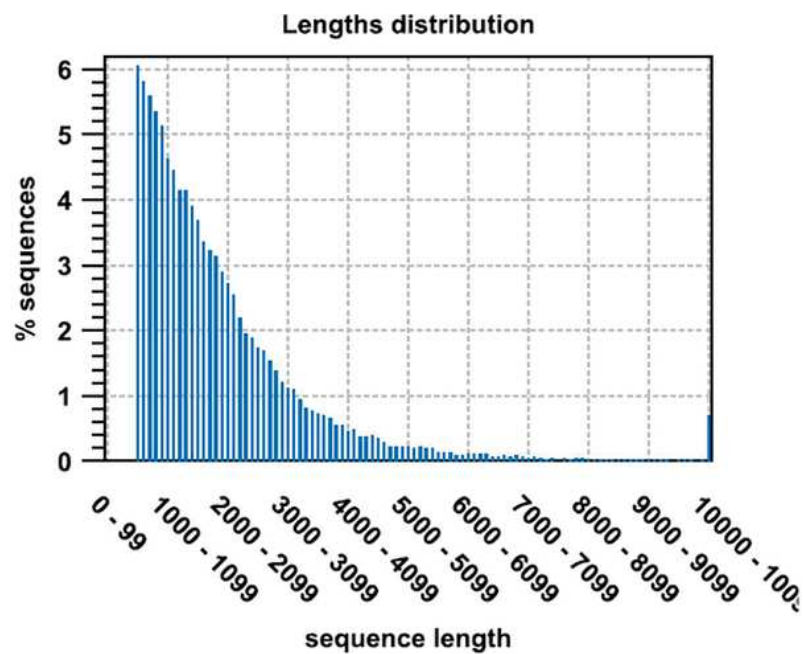
RNA-Seq data were *de novo* assembled using the Trinity assembly package which comprises combining sequence reads into larger contigs (by Inchworm), clustering contigs into a component (by Chrysalis), and producing the most plausible sets of transcripts from these groups (by Butterfly) (Grabherr et al. 2011). 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 2 and Figure 1. 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 (Grabherr et al. 2011). 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 (Lottaz et al. 2003, Grabherr et al. 2011).

**Table 2:** Assembly statistics of Illunina paired-end data.

	All contigs	Long contigs ( $\geq 500$ 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

doi:10.1371/journal.pone.0121778.t002



**Figure 1:** Distribution of contig ( $\geq 500$  nt) length of a Rainbow trout Illumina/Trinity transcriptome assembly.

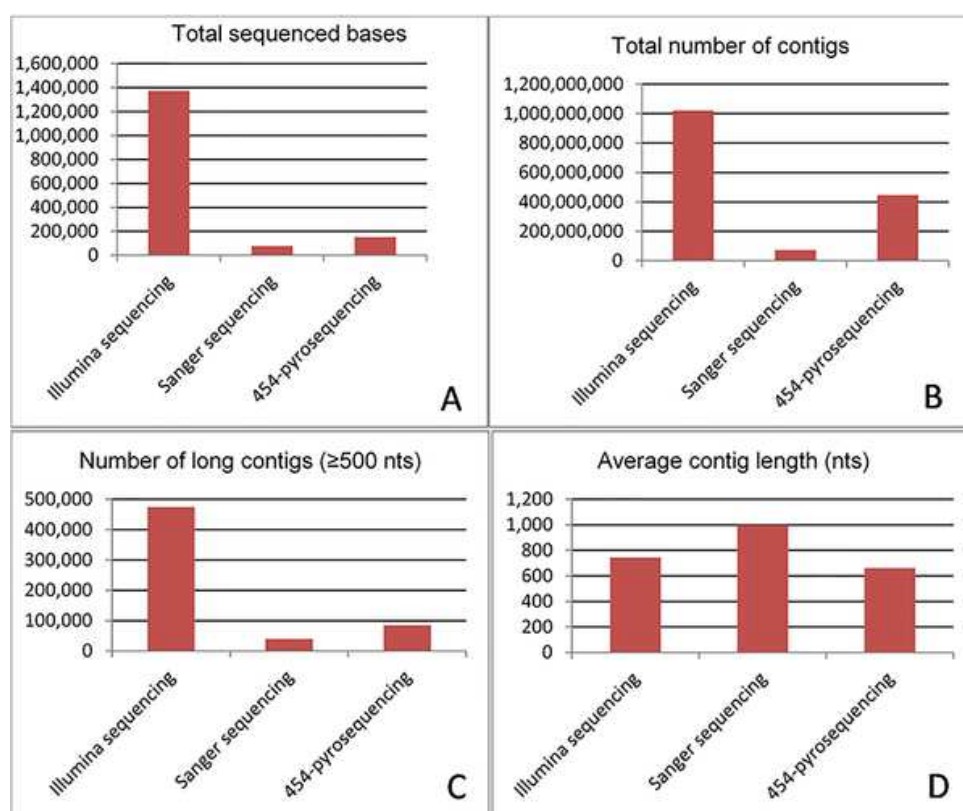
All 474,524 Trinity contigs longer than 500 nucleotides were searched against the NCBI non-redundant (NR) protein database. A total of 287,593 (60.60%) contigs had hits to the database proteins. Importantly, 92.5% (266,188) of these contigs were part of the components with more than one contig, indicating 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 (Rexroad et al. 2003). 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 (Wang et al. 2010). Likewise, three transcriptomes from Antarctic notothenioid fish revealed 38–45% significant BLASTx hits in the NR protein database (Shin et al. 2012). The unmatched contigs were used to identify a large number of non-coding RNAs (Al-Tobasei, Paneru and Salem 2016). In addition, the unmatched contigs may result from mistakes in assembly (contigs from reads with sequence errors) (Grabherr et al. 2011), lack of protein sequences of related fish in the database, or “trout-specific” diverged sequences due to the whole genome duplication (Ravi and Venkatesh 2008, Lee et al. 2011).

Previously, we utilized Sanger-based and 454-pyrosequencing approaches for transcriptomic analysis of the Rainbow trout (Salem et al. 2010). Figure 2 shows comparisons of the total number of sequenced bases, number of contigs, number of long contigs ( $\geq 500$  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 2).



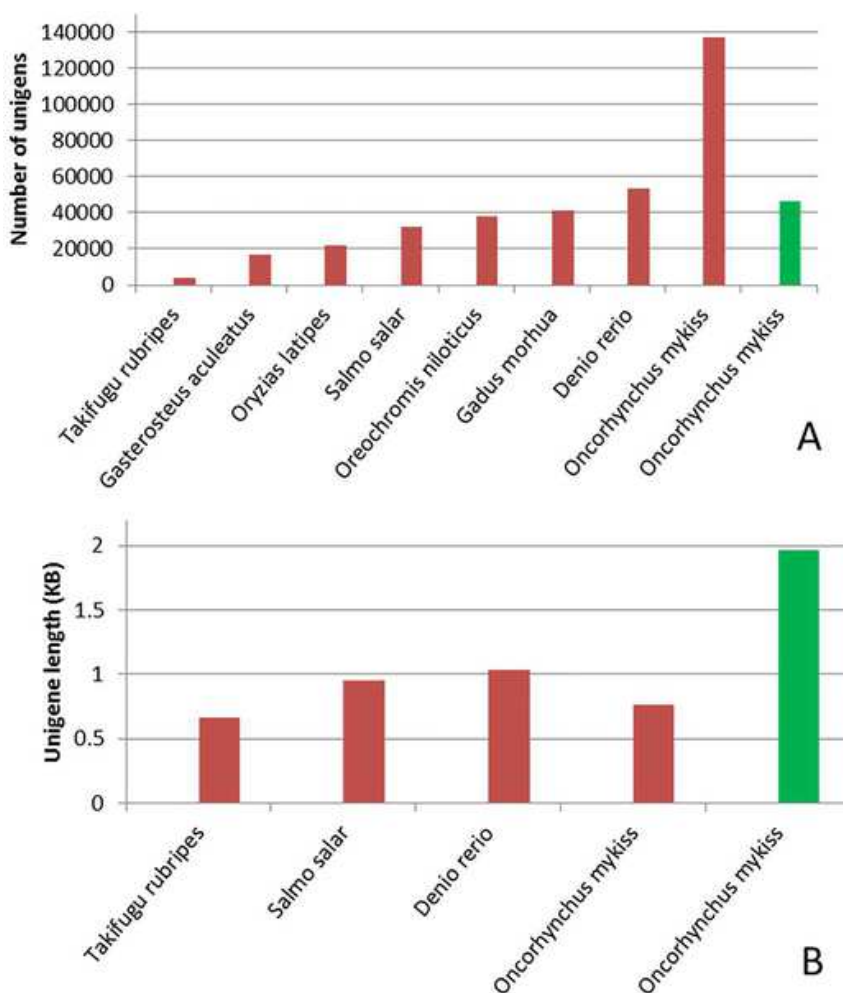
**Figure 2:** Comparison of total number of sequenced bases (A), total number of contigs (B), number of long contigs ( $\geq 500$  bp) (C), and average length of contigs (D) obtained from Illumina, Sanger-based, and 454-pyrosequencing techniques. Data on Sanger-based and 454-pyrosequencing techniques were obtained from Salem et. al (Salem et al. 2010).

### **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 hits to 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 1.64 kb and 44,990 versus 46,585 in the Trinity assembly and the Rainbow trout genome, respectively (Berthelot et al. 2014). 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 (Wang et al. 2010). 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. Coding sequences were annotated in a recent assembly of the Rainbow trout genome (Berthelot et al. 2014), 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 3). 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 3:** 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) (A). Average length of UniGenes of model fish species and Rainbow trout UniGenes that are available in the NCBI database (red bars) compared with the average length of Rainbow trout protein coding transcripts obtained from Illumina sequencing (green bar) (B). The high number and short length of Rainbow trout UniGenes suggest incomplete partial sequences.

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 (Grabherr *et al.* 2011). 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 (Zhao *et al.* 2011). 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 (Zhao *et al.* 2011).

### **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 (Iseli *et al.* 1999). 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 (Kent 2002). 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 (Koop et al. 2008). 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 3). Similar levels of homology to model fish species were reported in a catfish EST project (54% to 57%) (Wang et al. 2010) and a common carp transcriptome study (47.7% to 54.2%) (Ji et al. 2012). 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 3:** 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
<i>Takifugu rubripes</i>	16,621	34.73% of 47,856
<i>Danio rerio</i>	16,345	40.11% of 40,747
<i>Oryzias latipes</i>	11,854	48.14% of 24,619
<i>Gasterosteus aculeatus</i>	13,409	49.21% of 27,248
<i>Tetraodon nigroviridis</i>	11,617	50.24% of 23,123
<i>Oreochromis niloticus</i>	14,206	53.10% of 26,753
<i>Gadus morhua</i>	14,961	67.69% of 22,100

doi:10.1371/journal.pone.0121778.t003

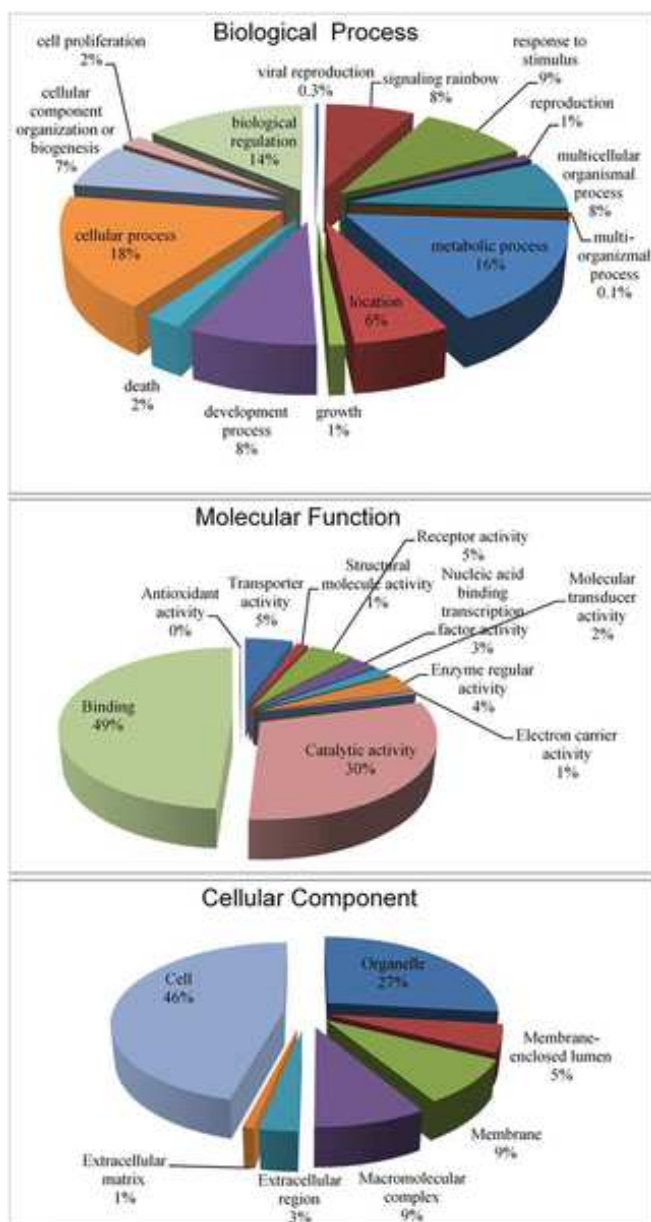
### **Functional annotation and gene ontology analyses**

Gene ontology provides organized terms to describe characteristics of gene products in three independent categories: biological processes, molecular function, and cellular components (Ashburner et al. 2000, Consortium 2008). Functional annotation of the Illumina/Trinity transcriptome contigs was performed by BLASTx search against the NCBI NR protein database using the Blast2GO suite (Götz et al. 2008). 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 4). 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 4).

Previously, we performed functional annotation of Rainbow trout transcripts sequenced using Sanger based and 454-pyrosequencing techniques (Salem et al. 2010). 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 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 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.



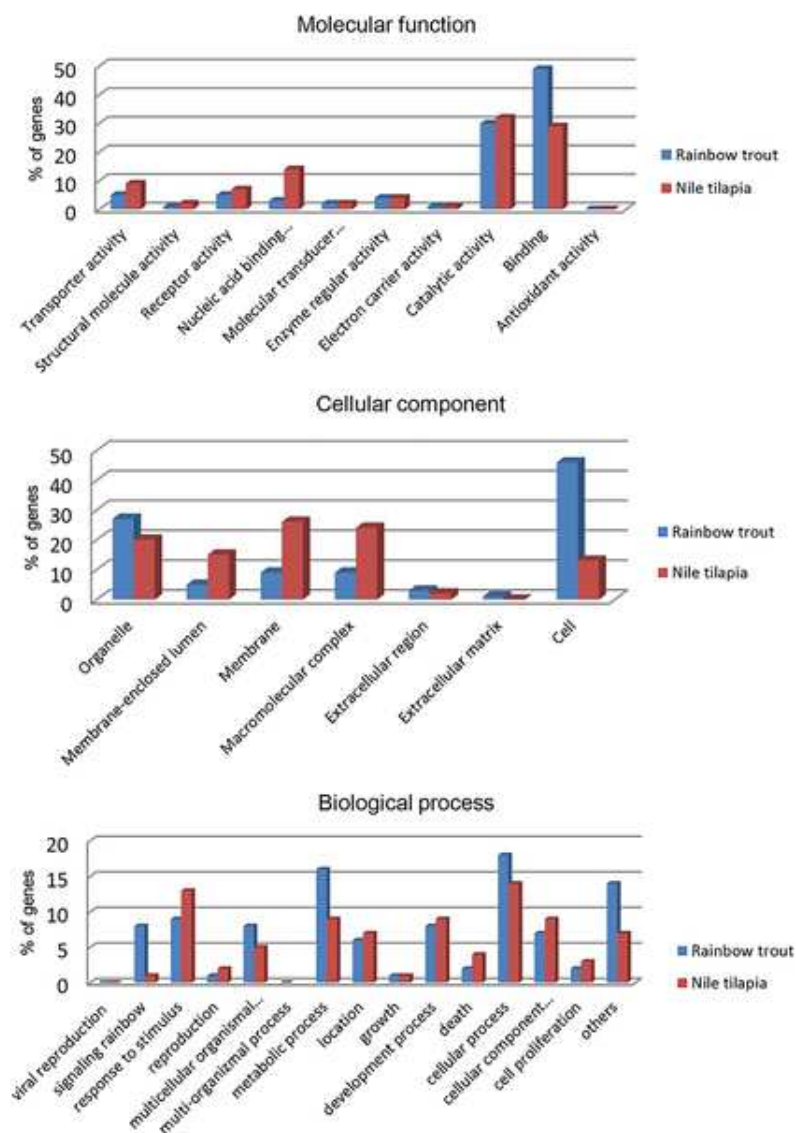
**Figure 4:** Gene Ontology (GO) assignment (2nd level GO terms) of the Rainbow trout of 13 lanes of Illumina Trinity assembly.

### Taxonomic analysis

BLASTx top-hit species distribution of the gene annotations showed the highest number of matches to Nile tilapia (*Oreochromis niloticus*) followed by Zebra fish (*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 5). 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 5). 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 5:** Gene Ontology (2<sup>nd</sup> level GO terms) comparison of Rainbow trout and Nile tilapia.

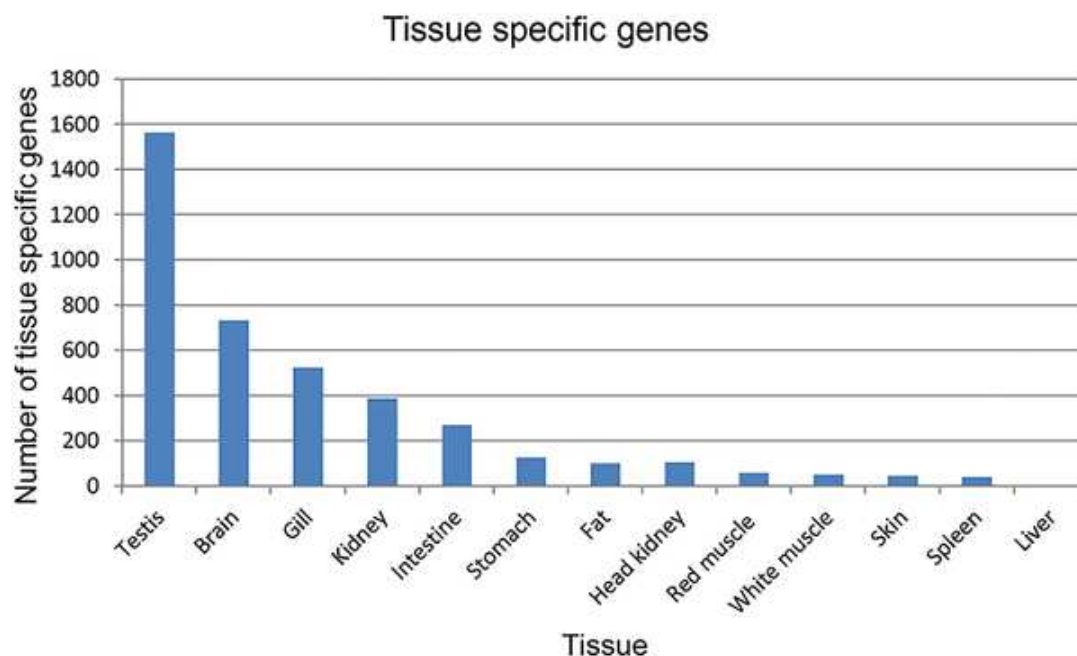
### Characterization of housekeeping and tissue-specific genes

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 (data not shown). 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 (Su et al. 2004, Jongeneel et al. 2005, Ramsköld et al. 2009). 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 6). 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 6). A similar trend of tissue specificity was observed in the human and mouse genomes (Ramsköld et al. 2009). Some of the brain specific genes were highly enriched in brain compared to other tissues (> 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 (Spooren et al. 2001), and GABA (gamma-aminobutyric acid) receptor A is the principal inhibitory neurotransmitter in the mammalian central nervous system (Lamp et al. 2001). In skin, one of the three most highly expressed protein is lily-type lectin which is a predominant protein in mucus of fish skin and provides important innate immunity (Suzuki et al. 2003, Tsutsui et al. 2003). 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 6:** Number of tissue-specific genes predicted in different 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 (Ramsköld et al. 2009). 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 (Ramsköld et al. 2009).

Taken together, these data indicate 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 4). 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 (Ramsköld et al. 2009). 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 (Bishop et al. 1974, Hastie and Bishop 1976). 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 4). 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) (Jongeneel et al. 2005).

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 4) with a small fraction of the mRNA pool contributed by the most highly expressed genes (Figure 7). 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 the mRNA pool in testis (Figure 7). Similar trends in transcriptome complexity were reported from previous studies in mammals (Jongeneel et al. 2005, Ramsköld et al. 2009) 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 (Chan et al. 2009).

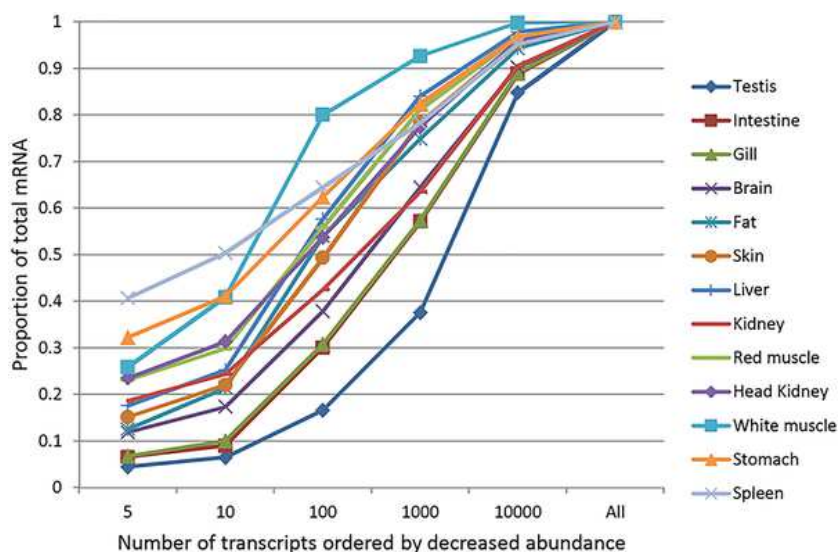


**Table 4:** Number of genes expressed in 13 Rainbow trout tissues at different RPKM threshold.

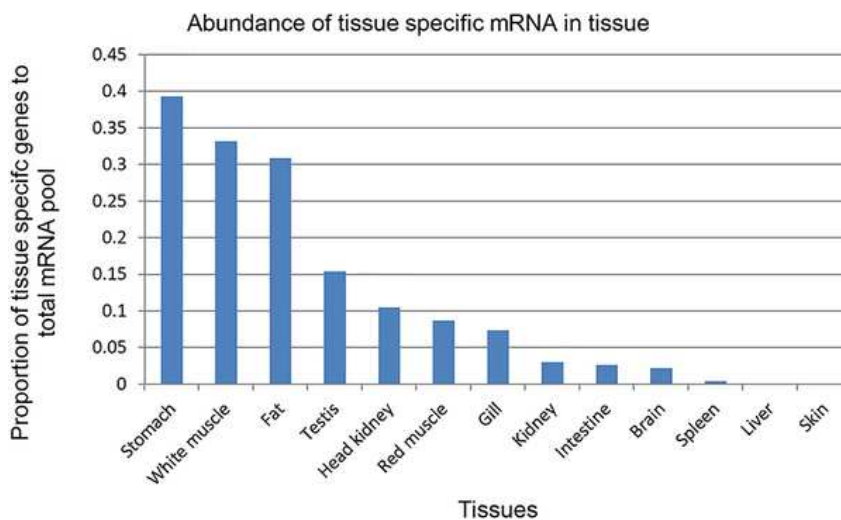
Tissue	RPKM $\geq$ 5.0		RPKM $\geq$ 1.0		RPKM $\geq$ 0.5		RPKM $\geq$ 0.1	
	Number of genes expressed	Fraction of total genes	Number of genes expressed	Fraction of total genes	Number of genes expressed	Fraction of total genes	Number of genes expressed	Fraction of total genes
White muscle	2,949	0.06	10,798	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	0.60	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	0.60	32,053	0.71	39,697	0.88
Intestine	13,655	0.30	27,018	0.60	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

doi:10.1371/journal.pone.0121778.t004

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 8). 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 8). 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 7:** Distribution of gene abundance in various tissues.



**Figure 8:** Transcript abundance of tissue-specific genes in various tissues.

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 8). 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 (data not shown). Recently, we reported a detailed expression in the spleen transcriptome in Rainbow trout (Ali et al. 2014). 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 (Berthelot et al. 2014). 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 (Berthelot et al. 2014). 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 Zebra fish 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 Zebra fish 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%) Zebra fish 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 Zebra fish 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 (Berthelot et al. 2014).

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 (Kent 2002). 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 (Trapnell et al. 2012). 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 (Min et al. 2005) 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 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 (Fox et al. 2014). 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 (Modrek and Lee 2002). However, because of the whole genome duplication event in teleost fish, many genes have paralogous duplicates (Taylor et al. 2003, Hoegg et al. 2004, Steinke et al. 2006). Indeed, gene duplication can lead to loss of alternative splicing of genes (Altschmied et al. 2002, Yu, Brenner and Venkatesh 2003) and many of the splice variants present in an ancestor are found to be expressed separately from duplicated genes in teleost fish (Xing and Lee 2006). The rate of alternative splicing was lowest (17%) in the highly duplicated genome of Zebra fish compared to the compact genome of the pufferfish (43%) (Lu et al. 2010). 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.



## REFERENCES

- Al-Tobasei, R., B. Paneru & M. Salem (2016) Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout. *PLoS One*, 11, e0148940.
- Ali, A., C. E. Rexroad, G. H. Thorgaard, J. Yao & M. Salem (2014) Characterization of the rainbow trout spleen transcriptome and identification of immune-related genes. *Front Genet*, 5, 348.
- Alkan, C., S. Sajjadian & E. E. Eichler (2011) Limitations of next-generation genome sequence assembly. *Nat Methods*, 8, 61-5.
- Altschmied, J., J. Delfgaauw, B. Wilde, J. Duschl, L. Bouneau, J. N. Volff & M. Scharl (2002) Subfunctionalization of duplicate *mitf* genes associated with differential degeneration of alternative exons in fish. *Genetics*, 161, 259-67.
- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin & G. Sherlock (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25, 25-9.
- Bailey, G. S., R. T. Poulter & P. A. Stockwell (1978) Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. *Proc Natl Acad Sci USA*, 75, 5575-9.
- Berthelot, C., F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noël, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J. M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J. N. Volff, C. Genêt, P. Wincker, O. Jaillon, H. Roest Crollius & Y. Guiguen (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*, 5, 3657.
- Bishop, J. O., J. G. Morton, M. Rosbash & M. Richardson (1974) Three abundance classes in HeLa cell messenger RNA. *Nature*, 250, 199-204.
- Butte, A. J., V. J. Dzau & S. B. Glueck (2001) Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". *Physiol Genomics*, 7, 95-6.
- Chan, E. T., G. T. Quon, G. Chua, T. Babak, M. Trochesset, R. A. Zirngibl, J. Aubin, M. J. Ratcliffe, A. Wilde, M. Brudno, Q. D. Morris & T. R. Hughes (2009) Conservation of core gene expression in vertebrate tissues. *J Biol*, 8, 33.
- Consortium, G. O. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res*, 36, D440-4.
- Davidson, W. S. (2012) Adaptation genomics: next generation sequencing reveals a shared haplotype for rapid early development in geographically and genetically distant populations of rainbow trout. *Mol Ecol*, 21, 219-22.
- Devisetty, U. K., M. F. Covington, A. V. Tat, S. Lekkala & J. N. Maloof (2014) Polymorphism identification and improved genome annotation of *Brassica rapa* through Deep RNA sequencing. *G3 (Bethesda)*, 4, 2065-78.
- Fowlkes, C. C., C. L. Hendriks, S. V. Keränen, G. H. Weber, O. Rübél, M. Y. Huang, S. Chatoor, A. H. DePace, L. Simirenko, C. Henriquez, A. Beaton, R. Weiszmann, S.

- Celniker, B. Hamann, D. W. Knowles, M. D. Biggin, M. B. Eisen & J. Malik (2008) A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm. *Cell*, 133, 364-74.
- Fox, S. E., M. R. Christie, M. Marine, H. D. Priest, T. C. Mockler & M. S. Blouin (2014) Sequencing and characterization of the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome. *Mar Genomics*, 15, 13-5.
- Giaquinto, P. C. & T. J. Hara (2008) Discrimination of bile acids by the rainbow trout olfactory system: evidence as potential pheromone. *Biol Res*, 41, 33-42.
- Götz, S., J. M. García-Gómez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talón, J. Dopazo & A. Conesa (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*, 36, 3420-35.
- Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman & A. Regev (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, 29, 644-52.
- Hastie, N. D. & J. O. Bishop (1976) The expression of three abundance classes of messenger RNA in mouse tissues. *Cell*, 9, 761-74.
- Henrich, T., M. Ramialison, B. Wittbrodt, B. Assouline, F. Bourrat, A. Berger, H. Himmelbauer, T. Sasaki, N. Shimizu, M. Westerfield, H. Kondoh & J. Wittbrodt (2005) MEPD: a resource for medaka gene expression patterns. *Bioinformatics*, 21, 3195-7.
- Hoegg, S., H. Brinkmann, J. S. Taylor & A. Meyer (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol*, 59, 190-203.
- Iseli, C., C. V. Jongeneel & P. Bucher (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*, 138-48.
- Ji, P., G. Liu, J. Xu, X. Wang, J. Li, Z. Zhao, X. Zhang, Y. Zhang, P. Xu & X. Sun (2012) Characterization of common carp transcriptome: sequencing, de novo assembly, annotation and comparative genomics. *PLoS One*, 7, e35152.
- Jongeneel, C. V., M. Delorenzi, C. Iseli, D. Zhou, C. D. Haudenschild, I. Khrebtukova, D. Kuznetsov, B. J. Stevenson, R. L. Strausberg, A. J. Simpson & T. J. Vasicek (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res*, 15, 1007-14.
- Kent, W. J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, 12, 656-64.
- Koop, B. F., K. R. von Schalburg, J. Leong, N. Walker, R. Lieph, G. A. Cooper, A. Robb, M. Beetz-Sargent, R. A. Holt, R. Moore, S. Brahmabhatt, J. Rosner, C. E. Rexroad, C. R. McGowan & W. S. Davidson (2008) A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics*, 9, 545.
- Kudoh, T., M. Tsang, N. A. Hukriede, X. Chen, M. Dedekian, C. J. Clarke, A. Kiang, S. Schultz, J. A. Epstein, R. Toyama & I. B. Dawid (2001) A gene expression screen in zebrafish embryogenesis. *Genome Res*, 11, 1979-87.

- Lamp, K., A. Humeny, Z. Nikolic, K. Imai, J. Adamski, K. Schiebel & C. M. Becker (2001) The murine GABA(B) receptor 1: cDNA cloning, tissue distribution, structure of the *Gabbr1* gene, and mapping to chromosome 17. *Cytogenet Cell Genet*, 92, 116-21.
- Langmead, B. & S. L. Salzberg (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9, 357-9.
- Lee, A. P., S. Y. Kerk, Y. Y. Tan, S. Brenner & B. Venkatesh (2011) Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Mol Biol Evol*, 28, 1205-15.
- Liu, S., R. L. Vallejo, G. Gao, Y. Palti, G. M. Weber, A. Hernandez & C. E. Rexroad (2015) Identification of single-nucleotide polymorphism markers associated with cortisol response to crowding in rainbow trout. *Mar Biotechnol (NY)*, 17, 328-37.
- Lottaz, C., C. Iseli, C. V. Jongeneel & P. Bucher (2003) Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics*, 19 Suppl 2, ii103-12.
- Lu, J., E. Peatman, W. Wang, Q. Yang, J. Abernathy, S. Wang, H. Kucuktas & Z. Liu (2010) Alternative splicing in teleost fish genomes: same-species and cross-species analysis and comparisons. *Mol Genet Genomics*, 283, 531-9.
- Marancik, D., G. Gao, B. Paneru, H. Ma, A. G. Hernandez, M. Salem, J. Yao, Y. Palti & G. D. Wiens (2014) Whole-body transcriptome of selectively bred, resistant-, control-, and susceptible-line rainbow trout following experimental challenge with *Flavobacterium psychrophilum*. *Front Genet*, 5, 453.
- Martin, J. A. & Z. Wang (2011) Next-generation transcriptome assembly. *Nat Rev Genet*, 12, 671-82.
- Min, X. J., G. Butler, R. Storms & A. Tsang (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res*, 33, W677-80.
- Modrek, B. & C. Lee (2002) A genomic view of alternative splicing. *Nat Genet*, 30, 13-9.
- Palti, Y., C. Genet, G. Gao, Y. Hu, F. M. You, M. Boussaha, C. E. Rexroad & M. C. Luo (2012) A second generation integrated map of the rainbow trout (*Oncorhynchus mykiss*) genome: analysis of conserved synteny with model fish genomes. *Mar Biotechnol (NY)*, 14, 343-57.
- Palti, Y., M. C. Luo, Y. Hu, C. Genet, F. M. You, R. L. Vallejo, G. H. Thorgaard, P. A. Wheeler & C. E. Rexroad (2009) A first generation BAC-based physical map of the rainbow trout genome. *BMC Genomics*, 10, 462.
- Papanastasiou, A. D., E. Georgaka & I. K. Zarkadis (2007) Cloning of a CD59-like gene in rainbow trout. Expression and phylogenetic analysis of two isoforms. *Mol Immunol*, 44, 1300-6.
- Patel, M., J. T. Rogers, E. F. Pane & C. M. Wood (2006) Renal responses to acute lead waterborne exposure in the freshwater rainbow trout (*Oncorhynchus mykiss*). *Aquat Toxicol*, 80, 362-71.
- Ramsköld, D., E. T. Wang, C. B. Burge & R. Sandberg (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*, 5, e1000598.
- Ravi, V. & B. Venkatesh (2008) Rapidly evolving fish genomes and teleost diversity. *Curr Opin Genet Dev*, 18, 544-50.

- Rexroad, C. E., Y. Lee, J. W. Keele, S. Karamycheva, G. Brown, B. Koop, S. A. Gahr, Y. Palti & J. Quackenbush (2003) Sequence analysis of a rainbow trout cDNA library and creation of a gene index. *Cytogenet Genome Res*, 102, 347-54.
- Robison, B., PA Thorgaard, GH (1999) Variation in development rate among clonal lines of 890 rainbow trout (*Oncorhynchus mykiss*). *Aquaculture*, 173, 131-141.
- Ryynänen, H. J. & C. R. Primmer (2006) 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*, 7, 192.
- Salem, M., C. E. Rexroad, J. Wang, G. H. Thorgaard & J. Yao (2010) Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches. *BMC Genomics*, 11, 564.
- Salem, M., R. L. Vallejo, T. D. Leeds, Y. Palti, S. Liu, A. Sabbagh, C. E. Rexroad, 3rd & J. Yao (2012) RNA-Seq identifies SNP markers for growth traits in rainbow trout. *PLoS One*, 7, e36264.
- Salgado, L. R., D. M. Koop, D. G. Pinheiro, R. Rivallan, V. Le Guen, M. F. Nicolás, L. G. de Almeida, V. R. Rocha, M. Magalhães, A. L. Gerber, A. Figueira, J. C. Cascardo, A. R. de Vasconcelos, W. A. Silva, L. L. Coutinho & D. Garcia (2014) De novo transcriptome analysis of *Hevea brasiliensis* tissues by RNA-seq and screening for molecular markers. *BMC Genomics*, 15, 236.
- Sánchez, C. C., G. M. Weber, G. Gao, B. M. Cleveland, J. Yao & C. E. Rexroad (2011) Generation of a reference transcriptome for evaluating rainbow trout responses to various stressors. *BMC Genomics*, 12, 626.
- Scheerer, P., GH & F. K. Allendorf (1986) Androgenetic rainbow trout produced from inbred and outbred sperm show similar survival. *Aquaculture*, 57, 289-298.
- Shin, S. C., S. J. Kim, J. K. Lee, D. H. Ahn, M. G. Kim, H. Lee, J. Lee, B. K. Kim & H. Park (2012) Transcriptomics and comparative analysis of three antarctic notothenioid fishes. *PLoS One*, 7, e43762.
- Speare, D., G. Arsenault & M. Buote (1998) Evaluation of Rainbow Trout as a Model for use in Studies on Pathogenesis of the Branchial Microsporidian *Loma salmonae*. *Contemp Top Lab Anim Sci*, 37, 55-58.
- Spooren, W. P., F. Gasparini, T. E. Salt & R. Kuhn (2001) Novel allosteric antagonists shed light on mglu(5) receptors and CNS disorders. *Trends Pharmacol Sci*, 22, 331-7.
- Steinke, D., W. Salzburger, I. Braasch & A. Meyer (2006) Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics*, 7, 20.
- Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker & J. B. Hogenesch (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101, 6062-7.
- Suzuki, Y., S. Tasumi, S. Tsutsui, M. Okamoto & H. Suetake (2003) Molecular diversity of skin mucus lectins in fish. *Comp Biochem Physiol B Biochem Mol Biol*, 136, 723-30.

- Taylor, J. S., I. Braasch, T. Frickey, A. Meyer & Y. Van de Peer (2003) Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res*, 13, 382-90.
- Tomancak, P., B. P. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S. E. Celniker & G. M. Rubin (2007) Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol*, 8, R145.
- Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn & L. Pachter (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7, 562-78.
- Tsutsui, S., S. Tasumi, H. Suetake & Y. Suzuki (2003) Lectins homologous to those of monocotyledonous plants in the skin mucus and intestine of pufferfish, *Fugu rubripes*. *J Biol Chem*, 278, 20882-9.
- Wang, S., E. Peatman, J. Abernathy, G. Waldbieser, E. Lindquist, P. Richardson, S. Lucas, M. Wang, P. Li, J. Thimmapuram, L. Liu, D. Vullaganti, H. Kucuktas, C. Murdock, B. C. Small, M. Wilson, H. Liu, Y. Jiang, Y. Lee, F. Chen, J. Lu, W. Wang, P. Xu, B. Somridhivej, P. Baoprasertkul, J. Quilang, Z. Sha, B. Bao, Y. Wang, Q. Wang, T. Takano, S. Nandi, S. Liu, L. Wong, L. Kaltenboeck, S. Quiniou, E. Bengten, N. Miller, J. Trant, D. Rokhsar, Z. Liu & C. G. Consortium (2010) Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. *Genome Biol*, 11, R8.
- Welsh, P. G., J. Lipton, C. A. Mebane & J. C. Marr (2008) Influence of flow-through and renewal exposures on the toxicity of copper to rainbow trout. *Ecotoxicol Environ Saf*, 69, 199-208.
- Williams, D. E. (2012) The rainbow trout liver cancer model: response to environmental chemicals and studies on promotion and chemoprevention. *Comp Biochem Physiol C Toxicol Pharmacol*, 155, 121-7.
- Xiao, S. J., C. Zhang, Q. Zou & Z. L. Ji (2010) TiSGeD: a database for tissue-specific genes. *Bioinformatics*, 26, 1273-5.
- Xing, Y. & C. Lee (2006) Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes. *Nat Rev Genet*, 7, 499-509.
- Young, W. P., P. A. Wheeler, R. D. Fields & G. H. Thorgaard (1996) DNA fingerprinting confirms isogenicity of androgenetically derived rainbow trout lines. *J Hered*, 87, 77-80.
- Yu, W. P., S. Brenner & B. Venkatesh (2003) Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in *Fugu*. *Trends Genet*, 19, 180-3.
- Zhang, H., E. Tan, Y. Suzuki, Y. Hirose, S. Kinoshita, H. Okano, J. Kudoh, A. Shimizu, K. Saito, S. Watabe & S. Asakawa (2014) Dramatic improvement in genome assembly achieved using doubled-haploid genomes. *Sci Rep*, 4, 6780.
- Zhao, Q. Y., Y. Wang, Y. M. Kong, D. Luo, X. Li & P. Hao (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, 12 Suppl 14, S2.

**CHAPTER II**  
**GENOME-WIDE DISCOVERY OF LONG NON-CODING RNA IN RAINBOW**  
**TROUT**

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

**ABSTRACT**

The ENCODE project revealed that ~70% of the human genome is transcribed. While only 1–2% of the RNAs encode for proteins, the rest are non-coding RNAs. Long non-coding RNAs (lncRNAs) form a diverse class of non-coding RNAs that are longer than 200nt. Emerging evidence indicates that lncRNAs play critical roles in various cellular processes including regulation of gene expression. LncRNAs show low levels of gene expression and sequence conservation, which make their computational identification in genomes difficult. In this study, more than two billion Illumina sequence reads were mapped to the genome reference using the TopHat and Cufflinks software. Transcripts shorter than 200nt, with more than 83–100 amino acids ORF, or with significant homologies to the NCBI nr-protein database were removed. In addition, a computational pipeline was used to filter the remaining transcripts based on a protein-coding-score test. Depending on the filtering stringency conditions, between 31,195 and 54,503 lncRNAs were identified, with only 421 matching known lncRNAs in other species. A digital gene expression atlas revealed 2,935 tissue-specific and 3,269 ubiquitously-expressed lncRNAs. This study annotates the lncRNA Rainbow trout genome and provides a valuable resource for functional genomics research in salmonids.

## 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 (Derrien et al. 2012, Clark et al. 2013). 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 (Rinn and Chang 2012, Zhu and Wang 2012). 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 (Ponting, Oliver and Reik 2009, Gibb, Brown and Lam 2011). 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 (Derrien et al. 2012).

Like protein-coding genes, lncRNAs are often transcribed by RNA polymerase II and can be post transcriptionally modified by splicing, capping and polyadenylation (Guttman et al. 2009, Beaulieu et al. 2012, Guttman and Rinn 2012, Yin et al. 2012). In contrast to protein-coding genes, a majority of lncRNA transcripts tend to have fewer exons (Derrien et al. 2012) and a shorter transcript size (average of 800 nucleotides) (Ørom et al. 2010). LncRNAs usually exhibit highly cell- and tissue-specific expression patterns and sometimes they are uniquely localized to a specific cellular compartment (Prasanth et al. 2005, Ginger et al. 2006, Mercer et al. 2008).

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 (Kino et al. 2010, Hung et al. 2011). Some lncRNAs regulate genes by acting as scaffolds to bring two or more proteins into a discrete complex (Pandey et al. 2008, Yap et al. 2010, Kotake et al. 2011). Other lncRNAs regulate different developmental and cellular processes by guiding a specific protein complexes to a specific promoter in response to certain molecular signals (Rinn et al. 2007, Huarte et al. 2010). 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* (Martianov et al. 2007, Jeon and Lee 2011). Beside transcriptional control, lncRNAs regulate many molecular processes including alternative splicing (Tripathi et al. 2010, Zong, Tripathi and Prasanth 2011), other post transcriptional processes (Yoon, Abdelmohsen and Gorospe 2013), and mRNA transport (Tripathi et al. 2012).

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 (Papanastasiou, Georgaka and Zarkadis 2007, Giaquinto and Hara 2008). 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 (Berthelot et al. 2014). 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 (Pandey et al. 2008, Yap et al. 2010, Kotake et al. 2011). With growing interest in lncRNAs-mediated gene regulation, these RNAs have been characterized, genome-wide, in limited animal and plant species in recent years (Cabili et al. 2011, Li et al. 2014). And, our knowledge of lncRNAs in fish is still very limited (Pauli et al. 2012). 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 (Salem et al. 2015). 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. (Berthelot et al. 2014). 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. (Fox et al. 2014). 4) About 89 million reads data set from redband trout (*Oncorhynchus mykiss*) by Narum et al. (Narum and Campbell 2015). 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 (Berthelot et al. 2014) using the TopHat and Cufflinks software packages (Trapnell et al. 2012). 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) (Kong et al. 2007). 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 (Wuyts et al. 2002, Chan and Lowe 2009, Quast et al. 2013, Van Peer et al. 2014). Any transcripts which showed sequence similarity with any of these classes of

RNAs with cut-off E value of  $\leq 0.0001$  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 (Bu et al. 2012, Pauli et al. 2012, Kaushik et al. 2013, Xie et al. 2014, Quek et al. 2015, RNAcentral Consortium 2015). 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 (Narum and Campbell 2015). 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. In addition, the cut-off value for the CPC (Kong et al. 2007) 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 (Salem et al. 2015).

### **Gene Clustering**

Sequencing reads from each tissue were mapped to combined reference consisting of the lncRNAs and mRNAs from the Rainbow trout genome reference (Berthelot et al. 2014). 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 (Saeed et al. 2003). The minimum correlation threshold to generate clusters was 0.97.

### **Identification of Genomic Location of lncRNAs Relative to Neighboring Protein-Coding Genes**

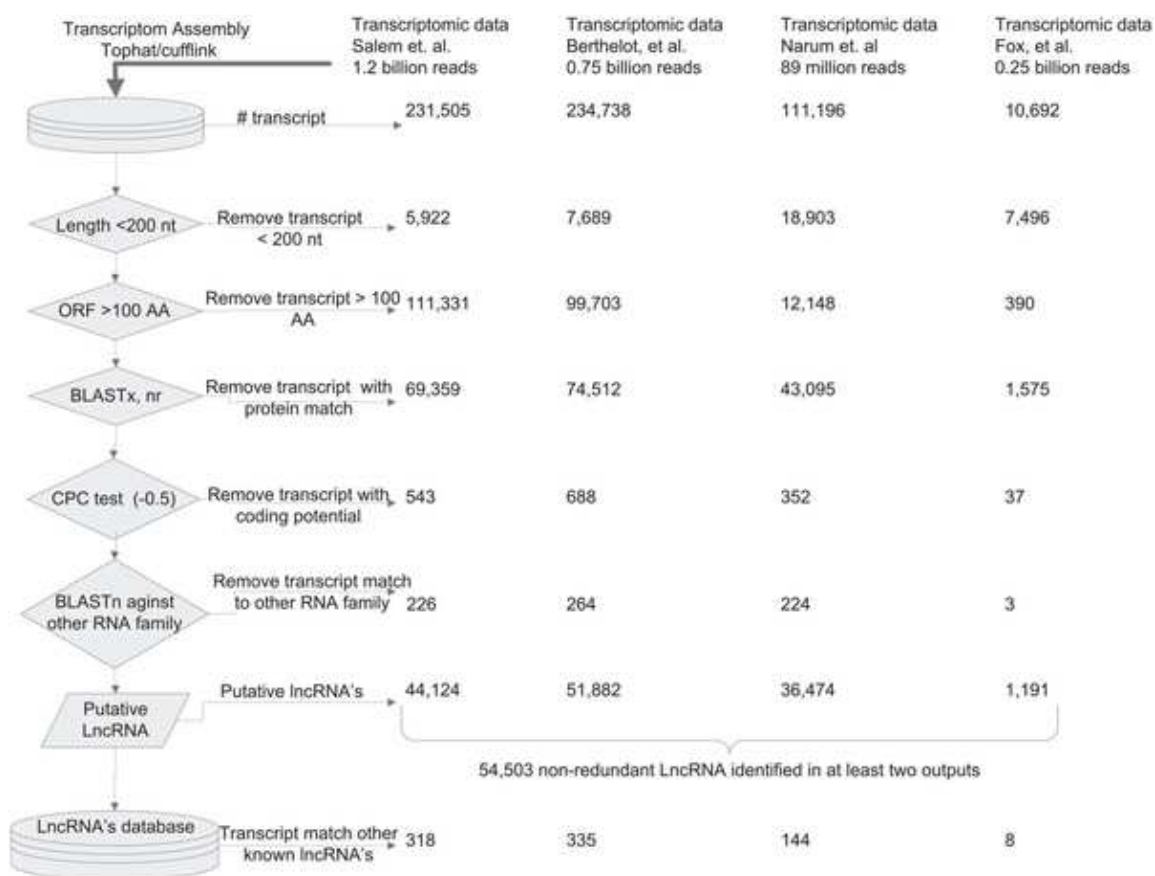
lncRNAs were classified based on their intersection or relative location to protein-coding genes using in-house Perl scripts using the Rainbow trout genome data (downloaded from <http://www.genoscope.cns.fr/trout/data/>).

## 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 (Salem et al. 2015). 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 (Langmead and Salzberg 2012, Trapnell et al. 2012), 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 (Figure 1). 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  $\leq 0.0001$ . 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 (Kong et al. 2007). 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 (Wuyts et al. 2002, Chan and Lowe 2009, Quast et al. 2013, Van Peer et al. 2014). After application of the above filtration steps, we found 44,124 putative lncRNAs from our sequence data set. These lncRNAs exhibited little or no evidence of coding potential or belonging to other non-coding classes of RNA.



**Figure 1:** Bioinformatics pipeline used in prediction of Rainbow trout lncRNAs.

Because some of the lncRNAs are thought to be due to expression noise (Louro, Smirnova and Verjovski-Almeida 2009), 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 (Figure 1). Those three sources were sequence data used by Berthelot et al. (Berthelot et al. 2014) in annotating the Rainbow trout genome, a data set used by Fox et al. (Fox et al. 2014) in assembling the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome and a data set from redband trout (*Oncorhynchus mykiss*) that was reported by Narum et al. (Narum and Campbell 2015). 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. Figure 1 illustrates the bioinformatics pipeline used in prediction of lncRNAs in all four datasets, and Table 1 reports the number of putative lncRNAs predicted in each dataset. FASTA and GTF annotation of files are available at: <http://www.animalgenome.org/repository/pub/MTSU2015.1014/>.

To look for evolutionarily conserved lncRNAs in Rainbow trout, all putative lncRNA transcripts (54,503) were searched against several noncoding-RNA databases (E

$\leq 0.0001$ ) (Bu et al. 2012, Pauli et al. 2012, Kaushik et al. 2013, Xie et al. 2014, Quek et al. 2015, RNAcentral Consortium 2015). Of those 54,503 lncRNAs, only 421 had sequence homology to lncRNAs from other species. This low evolutionary conservation of lncRNAs is in agreement with previous reports (Derrien et al. 2012).

**Table 1:** Number of lncRNA predicted in at least 2 of the 4 datasets and final numbers after merging and removal of redundant sequences.

Source	LncRNAs common between two data sources				Putative non-redundant lncRNA from each sources after combining all four sources	
	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
					<b>Total</b>	<b>54,503</b>

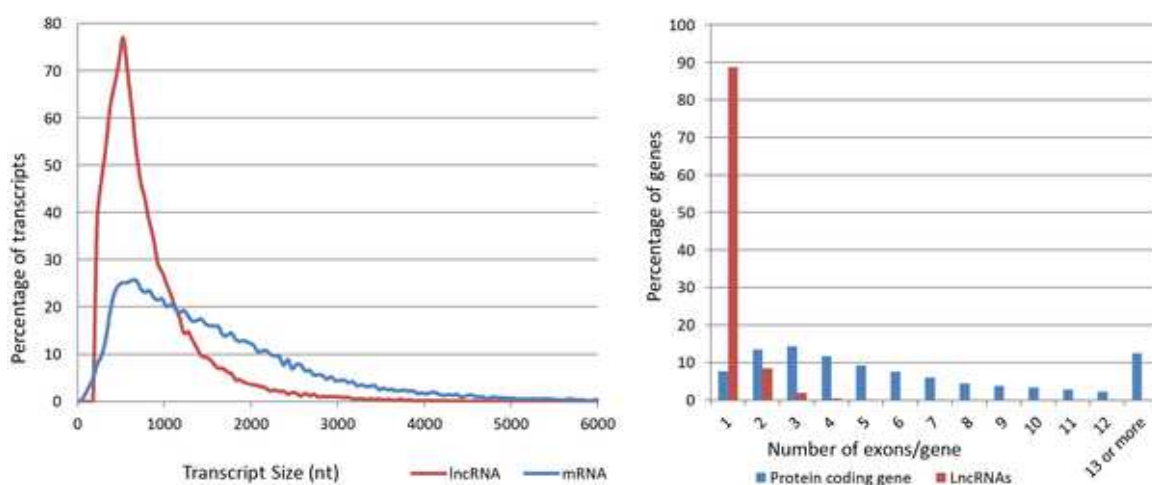
doi:10.1371/journal.pone.0148940.t001

## 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 (Pauli et al. 2012, Li et al. 2014). Consistent with previous reports, our study indicates that trout lncRNAs were shorter (0.821 kb) than protein-coding genes (1.636 kb) (Figure 2). 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. Figure 2 and Table 2 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 (Ravasi et al. 2006, Li et al. 2014). Conversely, some human studies showed a remarkable two-exon prevalence in the majority of lncRNAs (Derrien et al. 2012). 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 2:** Distribution of sequence length and number of exons in LncRNAs compared to protein-coding transcripts in Rainbow trout.

**Table 2:** Number of exons and average length of lncRNAs in different data sets.

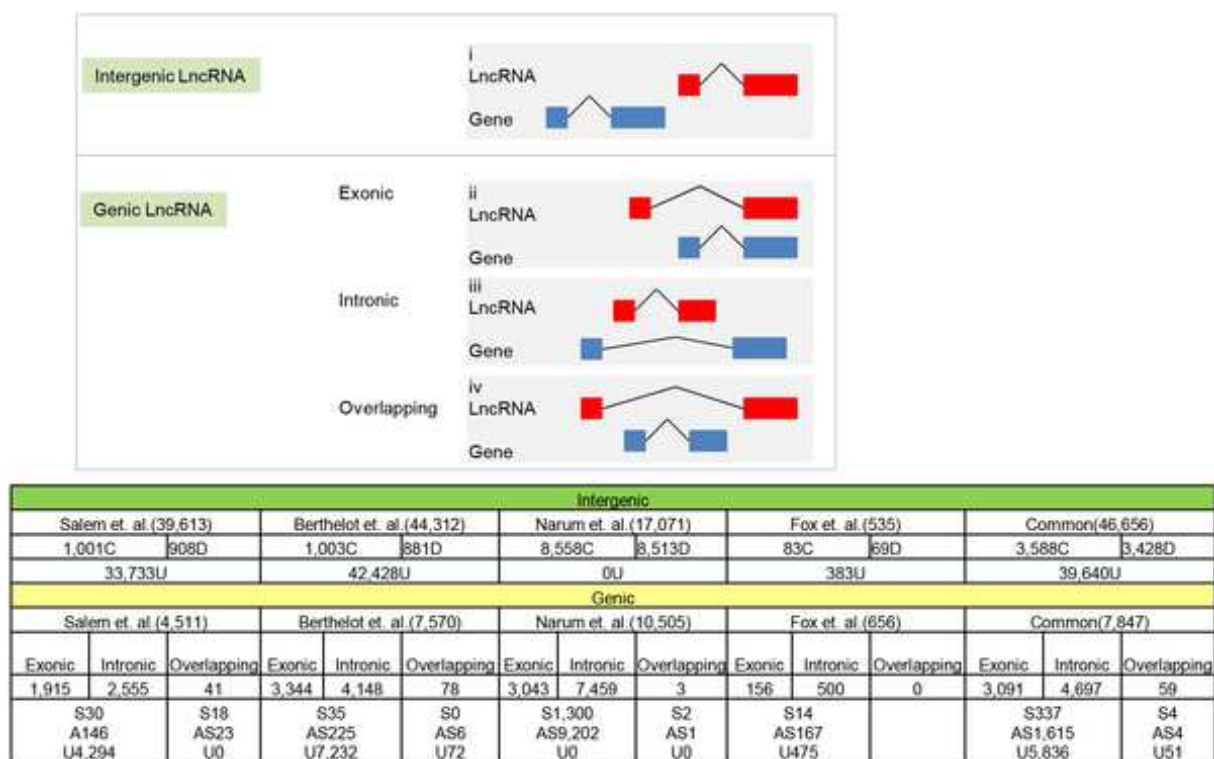
# of exon	Salem et al.		Berthelot et al.		Narum et al.		Fox et al.		Common	
	LncRNA %	Average length	LncRNA %	Average length	LncRNA %	Average length	LncRNA %	Average length	LncRNA %	Average length
1	86.14	790	88.52	682	96.62	453	98.24	353	88.84	796
2	10.63	888	8.71	846	2.79	462	1.34	377	8.49	1007
3	2.37	973	2.07	893	0.43	480	0.42	359	1.91	1044
4	0.51	1090	0.47	1030	0.1	475	0	0	0.46	1225
5	0.15	1284	0.11	1217	0.02	792	0	0	0.13	1390
6	0.08	1289	0.04	1157	0.02	514	0	0	0.07	1206
7	0.05	1379	0.03	1076	0.01	477	0	0	0.03	1183
8	0.03	1322	0.01	1227	0	631	0	0	0.02	1364
9	0.01	1217	0.01	1394	0.01	620	0	0	0.01	1302
10	0.02	1167	0.01	1199	0	0	0	0	0.01	1181

LncRNAs are classified, based on their intersection with protein-coding genes, as genic and intergenic (Derrien et al. 2012). Some of the lncRNAs are located in transcriptionally-active regions and influence expression of neighboring genes (Mercer, Dinger and Mattick 2009, Ponting et al. 2009). Therefore, the genomic position of lncRNAs relative to protein-coding genes can possibly provide important clues about lncRNA-mediated regulation of protein-coding genes (Villegas and Zaphiropoulos 2015). Our data indicate that 7,847 (14.4%) of the lncRNAs intersected with protein-coding gene and thus are called genic (Figure 3). 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. Figure 3 shows 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 (Derrien et al. 2012).

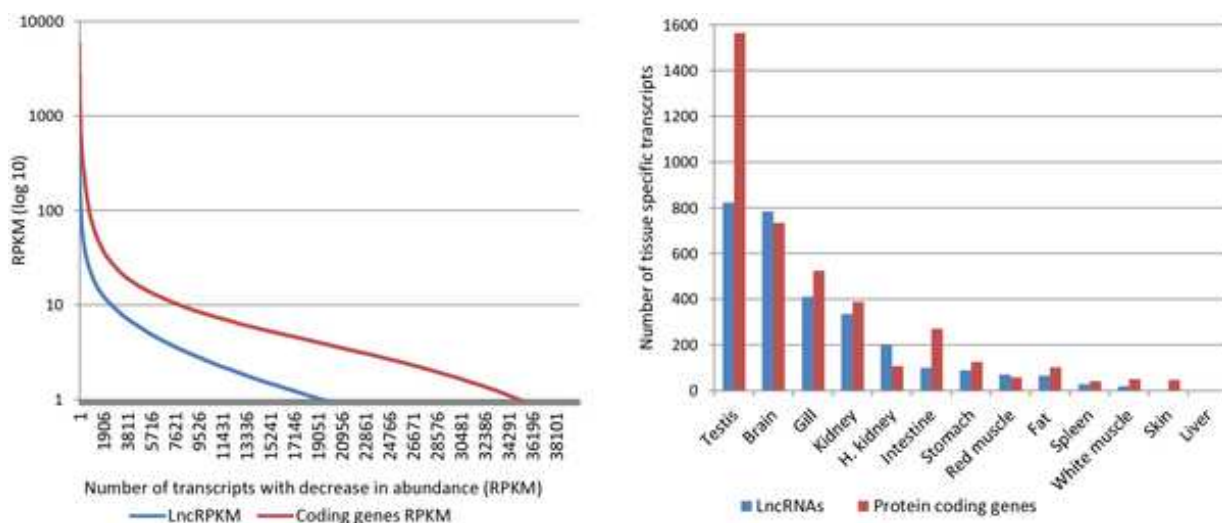
### **Expression of lncRNA in Different Tissues**

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 (Figure 4). Similar trends, showing lower lncRNAs expression in all human tissues compared to mRNAs, were reported (Derrien et al. 2012).



**Figure 3:** Classification of lncRNAs based on their intersection with protein-coding genes and number of lncRNAs in each class. Letters C, D, S, AS and U indicate number of convergent, divergent, sense, anti-sense and transcripts with unknown directionality, respectively.



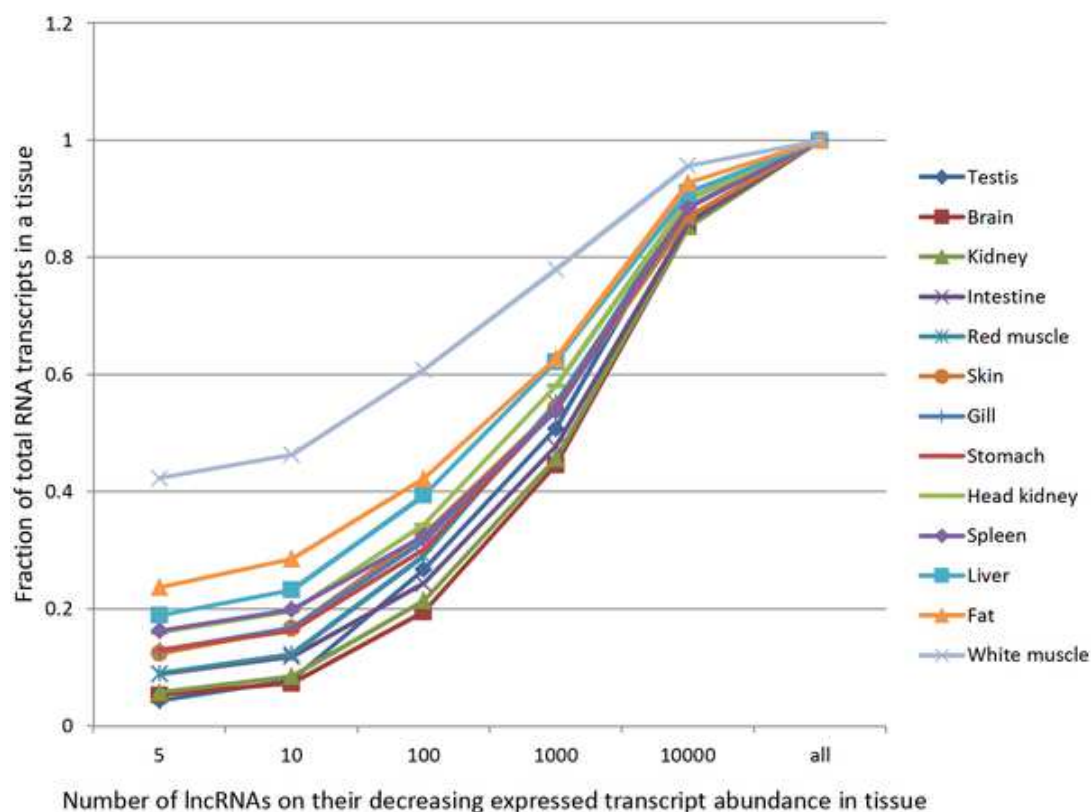
**Figure 4:** RPKM comparison of protein-coding genes and lncRNAs (left). Number of tissue-specific lncRNAs and protein-coding genes in various tissues (right).

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 (Cabili et al. 2011, Li et al. 2014). 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. On the other hand, 2,935 tissue-specific lncRNAs (5.4%) were identified from 13 tissues. 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 (Salem et al. 2015) (Figure 4). Previously, we reported 17.1%

and 8.9%, respectively, for housekeeping and tissue-specific protein-coding genes (Salem et al. 2015). 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 (Figure 4). 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 (Salem et al. 2015). Similar to the protein-coding genes, expression patterns of tissue-specific lncRNAs can be explained in terms of tissue complexity (Salem et al. 2015).

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 (Salem et al. 2015). 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 (Figure 5). 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 5:** Distribution of lncRNA expression in various tissues. Proportion of the transcriptome contributed by the most abundant lncRNAs is plotted in various tissues.

### 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) (Saeed et al. 2003). 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 ( $R^2 > 0.97$ ) (Appendix A). 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 correlation 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) (Dogan et al. 2011), RNA editing (adenosine deaminase) (Bass 2002), transcriptional control (histone deacetylase 3) (Wen et al. 2000), splicing, and development. These findings nicely correlate with previously characterized molecular functions of lncRNAs in different species (Tripathi et al. 2010, Yap et al. 2010, Zong et al. 2011). 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 54,503 putative lncRNAs were identified using filtration steps with traditional cutoff values (Pauli et al. 2012, Zhang et al. 2014). 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 (Davidson 2012); 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. Second, we decreased the cut-off value for the CPC (Kong et al. 2007) 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. 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, 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.

## REFERENCES

- Bass, B. L. (2002) RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem*, 71, 817-46.
- Beaulieu, Y. B., C. L. Kleinman, A. M. Landry-Voyer, J. Majewski & F. Bachand (2012) Polyadenylation-dependent control of long noncoding RNA expression by the poly(A)-binding protein nuclear 1. *PLoS Genet*, 8, e1003078.
- Berthelot, C., F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noel, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J. M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J. N. Volff, C. Genet, P. Wincker, O. Jaillon, H. Roest Crolius & Y. Guiguen (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*, 5, 3657.
- Bu, D., K. Yu, S. Sun, C. Xie, G. Skogerbø, R. Miao, H. Xiao, Q. Liao, H. Luo, G. Zhao, H. Zhao, Z. Liu, C. Liu, R. Chen & Y. Zhao (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res*, 40, D210-5.
- Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev & J. L. Rinn (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 25, 1915-27.
- Chan, P. P. & T. M. Lowe (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res*, 37, D93-7.
- Clark, M. B., A. Choudhary, M. A. Smith, R. J. Taft & J. S. Mattick (2013) The dark matter rises: the expanding world of regulatory RNAs. *Essays Biochem*, 54, 1-16.
- Davidson, W. S. (2012) Adaptation genomics: next generation sequencing reveals a shared haplotype for rapid early development in geographically and genetically distant populations of rainbow trout. *Mol Ecol*, 21, 219-22.
- Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow & R. Guigó (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, 22, 1775-89.
- Dougan, S. K., C. C. Hu, M. E. Paquet, M. B. Greenblatt, J. Kim, B. N. Lilley, N. Watson & H. L. Ploegh (2011) Derlin-2-deficient mice reveal an essential role for protein dislocation in chondrocytes. *Mol Cell Biol*, 31, 1145-59.
- Fox, S. E., M. R. Christie, M. Marine, H. D. Priest, T. C. Mockler & M. S. Blouin (2014) Sequencing and characterization of the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome. *Mar Genomics*, 15, 13-5.
- Giaquinto, P. C. & T. J. Hara (2008) Discrimination of bile acids by the rainbow trout olfactory system: evidence as potential pheromone. *Biol Res*, 41, 33-42.
- Gibb, E. A., C. J. Brown & W. L. Lam (2011) The functional role of long non-coding RNA in human carcinomas. *Mol Cancer*, 10, 38.

- Ginger, M. R., A. N. Shore, A. Contreras, M. Rijnkels, J. Miller, M. F. Gonzalez-Rimbau & J. M. Rosen (2006) A noncoding RNA is a potential marker of cell fate during mammary gland development. *Proc Natl Acad Sci U S A*, 103, 5781-6.
- Guttman, M., I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn & E. S. Lander (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458, 223-7.
- Guttman, M. & J. L. Rinn (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, 482, 339-46.
- Huarte, M., M. Guttman, D. Feldser, M. Garber, M. J. Koziol, D. Kenzelmann-Broz, A. M. Khalil, O. Zuk, I. Amit, M. Rabani, L. D. Attardi, A. Regev, E. S. Lander, T. Jacks & J. L. Rinn (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142, 409-19.
- Hung, T., Y. Wang, M. F. Lin, A. K. Koegel, Y. Kotake, G. D. Grant, H. M. Horlings, N. Shah, C. Umbricht, P. Wang, B. Kong, A. Langerød, A. L. Børresen-Dale, S. K. Kim, M. van de Vijver, S. Sukumar, M. L. Whitfield, M. Kellis, Y. Xiong, D. J. Wong & H. Y. Chang (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet*, 43, 621-9.
- Jeon, Y. & J. T. Lee (2011) YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*, 146, 119-33.
- Kaushik, K., V. E. Leonard, S. Kv, M. K. Lalwani, S. Jalali, A. Patowary, A. Joshi, V. Scaria & S. Sivasubbu (2013) Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish. *PLoS One*, 8, e83616.
- Kino, T., D. E. Hurt, T. Ichijo, N. Nader & G. P. Chrousos (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal*, 3, ra8.
- Kong, L., Y. Zhang, Z. Q. Ye, X. Q. Liu, S. Q. Zhao, L. Wei & G. Gao (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*, 35, W345-9.
- Kotake, Y., T. Nakagawa, K. Kitagawa, S. Suzuki, N. Liu, M. Kitagawa & Y. Xiong (2011) Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene*, 30, 1956-62.
- Langmead, B. & S. L. Salzberg (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9, 357-9.
- Li, L., S. R. Eichten, R. Shimizu, K. Petsch, C. T. Yeh, W. Wu, A. M. Chettoor, S. A. Givan, R. A. Cole, J. E. Fowler, M. M. Evans, M. J. Scanlon, J. Yu, P. S. Schnable, M. C. Timmermans, N. M. Springer & G. J. Muehlbauer (2014) Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol*, 15, R40.
- Louro, R., A. S. Smirnova & S. Verjovski-Almeida (2009) Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics*, 93, 291-8.
- Martianov, I., A. Ramadass, A. Serra Barros, N. Chow & A. Akoulitchev (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature*, 445, 666-70.

- Mercer, T. R., M. E. Dinger & J. S. Mattick (2009) Long non-coding RNAs: insights into functions. *Nat Rev Genet*, 10, 155-9.
- Mercer, T. R., M. E. Dinger, S. M. Sunkin, M. F. Mehler & J. S. Mattick (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A*, 105, 716-21.
- Narum, S. R. & N. R. Campbell (2015) Transcriptomic response to heat stress among ecologically divergent populations of redband trout. *BMC Genomics*, 16, 103.
- Ørom, U. A., T. Derrien, M. Beringer, K. Gumireddy, A. Gardini, G. Bussotti, F. Lai, M. Zytynicki, C. Notredame, Q. Huang, R. Guigo & R. Shiekhattar (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143, 46-58.
- Pandey, R. R., T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, T. Nagano, D. Mancini-Dinardo & C. Kanduri (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell*, 32, 232-46.
- Papanastasiou, A. D., E. Georgaka & I. K. Zarkadis (2007) Cloning of a CD59-like gene in rainbow trout. Expression and phylogenetic analysis of two isoforms. *Mol Immunol*, 44, 1300-6.
- Pauli, A., E. Valen, M. F. Lin, M. Garber, N. L. Vastenhouw, J. Z. Levin, L. Fan, A. Sandelin, J. L. Rinn, A. Regev & A. F. Schier (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*, 22, 577-91.
- Ponting, C. P., P. L. Oliver & W. Reik (2009) Evolution and functions of long noncoding RNAs. *Cell*, 136, 629-41.
- Prasanth, K. V., S. G. Prasanth, Z. Xuan, S. Hearn, S. M. Freier, C. F. Bennett, M. Q. Zhang & D. L. Spector (2005) Regulating gene expression through RNA nuclear retention. *Cell*, 123, 249-63.
- Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies & F. O. Glöckner (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, 41, D590-6.
- Quek, X. C., D. W. Thomson, J. L. Maag, N. Bartonicek, B. Signal, M. B. Clark, B. S. Gloss & M. E. Dinger (2015) lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res*, 43, D168-73.
- Ravasi, T., H. Suzuki, K. C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M. C. Frith, M. M. Gongora, S. M. Grimmond, D. A. Hume, Y. Hayashizaki & J. S. Mattick (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res*, 16, 11-9.
- Rinn, J. L. & H. Y. Chang (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem*, 81, 145-66.
- Rinn, J. L., M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal & H. Y. Chang (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129, 1311-23.
- RNAcentral Consortium (2015) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res*, 43, D123-9.

- Saeed, A. I., V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush & J. Quackenbush (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34, 374-8.
- Salem, M., B. Paneru, R. Al-Tobasei, F. Abdouni, G. H. Thorgaard, C. E. Rexroad & j. Yao (2015) Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout. *PLoS ONE*.
- Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn & L. Pachter (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7, 562-78.
- Tripathi, V., J. D. Ellis, Z. Shen, D. Y. Song, Q. Pan, A. T. Watt, S. M. Freier, C. F. Bennett, A. Sharma, P. A. Bubulya, B. J. Blencowe, S. G. Prasanth & K. V. Prasanth (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell*, 39, 925-38.
- Tripathi, V., D. Y. Song, X. Zong, S. P. Shevtsov, S. Hearn, X. D. Fu, M. Dunder & K. V. Prasanth (2012) SRSF1 regulates the assembly of pre-mRNA processing factors in nuclear speckles. *Mol Biol Cell*, 23, 3694-706.
- Van Peer, G., S. Lefever, J. Anckaert, A. Beckers, A. Rihani, A. Van Goethem, P. J. Volders, F. Zeka, M. Ongenaert, P. Mestdagh & J. Vandesompele (2014) miRBase Tracker: keeping track of microRNA annotation changes. *Database (Oxford)*, 2014.
- Villegas, V. E. & P. G. Zaphiropoulos (2015) Neighboring gene regulation by antisense long non-coding RNAs. *Int J Mol Sci*, 16, 3251-66.
- Wen, Y. D., V. Perissi, L. M. Staszewski, W. M. Yang, A. Krones, C. K. Glass, M. G. Rosenfeld & E. Seto (2000) The histone deacetylase-3 complex contains nuclear receptor corepressors. *Proc Natl Acad Sci U S A*, 97, 7202-7.
- Wuyts, J., Y. Van de Peer, T. Winkelmans & R. De Wachter (2002) The European database on small subunit ribosomal RNA. *Nucleic Acids Res*, 30, 183-5.
- Xie, C., J. Yuan, H. Li, M. Li, G. Zhao, D. Bu, W. Zhu, W. Wu, R. Chen & Y. Zhao (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res*, 42, D98-103.
- Yap, K. L., S. Li, A. M. Muñoz-Cabello, S. Raguz, L. Zeng, S. Mujtaba, J. Gil, M. J. Walsh & M. M. Zhou (2010) Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol Cell*, 38, 662-74.
- Yin, Q. F., L. Yang, Y. Zhang, J. F. Xiang, Y. W. Wu, G. G. Carmichael & L. L. Chen (2012) Long noncoding RNAs with snoRNA ends. *Mol Cell*, 48, 219-30.
- Yoon, J. H., K. Abdelmohsen & M. Gorospe (2013) Posttranscriptional gene regulation by long noncoding RNA. *J Mol Biol*, 425, 3723-30.
- Zhang, K., K. Huang, Y. Luo & S. Li (2014) Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data. *BMC Genomics*, 15, 845.

- Zhu, Q. H. & M. B. Wang (2012) Molecular Functions of Long Non-Coding RNAs in Plants. *Genes (Basel)*, 3, 176-90.
- Zong, X., V. Tripathi & K. V. Prasanth (2011) RNA splicing control: yet another gene regulatory role for long nuclear noncoding RNAs. *RNA Biol*, 8, 968-77.

## **APPENDICES**

## APPENDIX A

**CLUSTER OF LNCRNA AND PROTEIN CODING GENES BASED ON THEIR  
EXPRESSION VALUES ACROSS 13 TISSUES. CLUSTERS WERE  
GENERATED AT THRESHOLD OF  $R^2 > 0.97$**

Expression line	Feature ID	Expression line	Feature ID
	GSONMT00007811001		Omy200106644
	GSONMT00067984001		GSONMT00058062001
	GSONMT00060849001	Expression line	Feature ID
	GSONMT00038863001		Omy300072481
	Omy100084431		GSONMT00066016001
	GSONMT00018324001	Expression line	Feature ID
Expression line	Feature ID		Omy200109475
	Omy200109475		GSONMT00041727001
	GSONMT00041727001		GSONMT00023776001
	GSONMT00023776001		GSONMT00042121001
	GSONMT00042121001		Omy100032261
	Omy100032261	Expression line	Feature ID
Expression line	Feature ID		GSONMT00004576001
	GSONMT00022941001		Omy100085482
	GSONMT00038477001	Expression line	Feature ID
	GSONMT00053908001		GSONMT00007875001
	GSONMT00078882001		Omy100095867
	Omy100001615	Expression line	Feature ID
Expression line	Feature ID		GSONMT00021515001
	Omy200064145		Omy100034895
	Omy100138726	Expression line	Feature ID
	GSONMT00082784001		GSONMT00027216001
Expression line	Feature ID		Omy100174349
	Omy300017627	Expression line	Feature ID
	Omy300017628		GSONMT00038305001
	GSONMT00047316001		Omy100136317
Expression line	Feature ID	Expression line	Feature ID
	Omy300066671		Omy200101305
	GSONMT00021822001		GSONMT00007048001
	GSONMT0000999001		



**CHAPTER III**

**DIFFERENTIAL EXPRESSION OF LONG NON-CODING RNAs IN THREE  
GENETIC LINES OF RAINBOW TROUT IN RESPONSE TO INFECTION  
WITH *FLAVOBACTERIUM PSYCHROPHILUM***

Paneru, B., R. Al-Tobasei, Y. Palti, G. D. Wiens & M. Salem (2016) Differential expression of long non-coding RNAs in three genetic lines of rainbow trout in response to infection with *Flavobacterium psychrophilum*. *Sci Rep*, 6, 36032.

**ABSTRACT**

Bacterial cold-water disease caused by *Flavobacterium psychrophilum* is one of the major causes of mortality of salmonids. Three genetic lines of Rainbow trout designated as ARS-Fp-R (resistant), ARS-Fp-C (control) and ARS-Fp-S (susceptible) have significant differences in survival rate following *F. psychrophilum* infection. Previous study identified transcriptome differences of immune-relevant protein-coding genes at basal and post infection levels among these genetic lines. Using RNA-Seq approach, we quantified differentially expressed (DE) long non-coding RNAs (lncRNAs) in response to *F. psychrophilum* challenge in these genetic lines. Pairwise comparison between genetic lines and different infection statuses identified 556 DE lncRNAs. A positive correlation existed between the number of the differentially regulated lncRNAs and that of the protein-coding genes. Several lncRNAs showed strong positive and negative expression correlation with their overlapped, neighboring and distant immune related protein-coding genes including complement components, cytokines, chemokines and several signaling molecules involved in immunity. The correlated expressions and genome-wide co-localization suggested that some lncRNAs may be involved in regulating immune-relevant protein-coding genes. This

study provides the first evidence of lncRNA-mediated regulation of the anti-bacterial immune response in a commercially important aquaculture species and will likely help developing new genetic markers for Rainbow trout disease resistance.

## **INTRODUCTION**

World aquaculture industries suffer considerable economic losses annually because of infectious diseases (Asche et al. 2009). *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 (Nematollahi et al. 2003)). Infection of Rainbow trout with *Fp* results in mortality of up to 30% and several complications in the survivors (Kent et al. 1989). Originally, the pathogen was considered to be endemic to North America but in recent years it has been reported from almost every continent (Carson and Schmidtke 1995). Multiple routes of transmission (Brown, Cox and Levine 1997), wide geographical distribution, the ability of pathogen to cope with harsh survival condition (Brown et al. 1997), 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 (Gómez et al. 2014)).

Harnessing the host's immune system by selective breeding is a strategy being pursued to improve farmed fish health (Gjedrem 2005). 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 (Wiens et al. 2013). Previously, we performed global expression analysis of protein-coding genes in these genetic lines upon *Fp* challenge (Marancik et al. 2014). 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 (Cabili et al. 2011). lncRNAs regulate diverse cellular processes, including disease, immunity, development and cell proliferation (Peng et al. 2010). In mammals, lncRNAs regulate various immune responses including the interferon response, inflammatory processes, and other aspects of innate and adaptive immune responses (Carpenter et al. 2013, Hu et al. 2013, Kambara et al. 2014). TLR signaling and inflammatory responses increase the expression of lncRNA-Cox2 that regulates both activation and repression of innate response genes (Carpenter et al. 2013). lncRNA NeST controls susceptibility to Theiler's virus and Salmonella infection through epigenetic regulation of the interferon-gamma locus (Collier et al. 2012, Gomez et al. 2013). 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 (Peng et al. 2010, Boltana et al. 2016). 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 (Al-Tobasei, Paneru and Salem 2016). 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 (Wiens et al. 2013) and we have previously reported the challenge experiment utilized in this study (Marancik et al. 2014). 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  $\mu$ l of chilled PBS or PBS alone, and survival was monitored daily for 21 days (Marancik et al. 2014). 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 (Marancik et al. 2014). 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) at the University of Illinois at Urbana-Champaign. RNA-Seq reads were downloaded from the NCBI Short Read Archive accession number 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 (Al-Tobasei et al. 2016). 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 (Salem et al. 2015). To control the false discovery due to multiple testing, p-values were FDR-corrected. LncRNA was considered significant at a fold-change cutoff value of  $\pm 2$  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 (Marancik et al. 2014). 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 (Marancik et al. 2014). Briefly, β-actin (Accession: [AJ438158](#)) was used as endogenous reference to normalize each target lncRNA. β-actin expression levels demonstrated in RNA-Seq data were similar in PBS and *F. psychrophilum*-injected fish. qPCR data were quantified using delta delta Ct ( $\Delta\Delta Ct$ ) methods (Schmittgen and Livak 2008). Ct-values of β-actin were subtracted from Ct-values of the target gene to calculate the normalized value ( $\Delta Ct$ ) of the target lncRNA in both the calibrator samples (PBS-injected) and test samples (*Fp*-injected). The  $\Delta Ct$  value of the calibrator sample was subtracted from the  $\Delta Ct$  value of the test sample to get the  $\Delta\Delta Ct$  value. Fold change in

gene expression in the test sample relative to the calibrator sample was calculated by the formula  $2^{-\Delta\Delta C_t}$  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 (Al-Tobasei et al. 2016), and mRNAs that were identified in the Rainbow trout genome (Berthelot et al. 2014). 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 (Bolstad et al. 2003). 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 (Al-Tobasei et al. 2016). Briefly, sequencing reads from each genetic line (resistance, control and



susceptible) were aligned to a Rainbow trout reference genome using TopHat. 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 selectively expressed lncRNAs in 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 (Marancik et al. 2014). 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 (Appendix A). 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  $\geq 0.5$ . Out of 31,195 reference lncRNAs, only 933 were not expressed in any dataset (RPKM  $\geq 0.5$ ). 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 (Al-Tobasei et al. 2016). 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 933 lncRNAs, 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 the comparisons (FDR-P-value  $< 0.05$ ) (Table 1). Of these, 556 were unique lncRNA showing differential expression in at least one comparison. In our previous study using the same genetic lines, about 2,600 DE immune-related and other protein-coding genes were identified in response to *Fp* infection (Marancik et al. 2014). 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 1). In general, within each pair-wise comparison, fewer differentially regulated lncRNA were identified as

compared to DE protein coding transcripts (Table 1). This may, in part, be, due to the overall lower expression level of lncRNA as compared to protein-coding genes (Derrien et al. 2012). 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 1). 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 5<sup>th</sup> 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 (Appendix B).

Recently, we reported tissue specificity of lncRNAs in Rainbow trout (Al-Tobasei et al. 2016). 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 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 (Madetoja, Nyman and Wiklund 2000).

### **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 (Carpenter et al. 2013, Hu et al. 2013, Kambara et al. 2014). 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  $\pm$  2 and FDR-corrected  $p$  value  $<$  0.05).

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 5<sup>th</sup> day of infection, 12 lncRNAs were significantly upregulated ( $>$  2-fold) in all three genetic lines (FDR-corrected  $p$ - value  $<$  0.05) (Table 2, 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 1:** Comparison of differentially expressed lncRNA and protein coding genes in response to Fp infection.

Comparison	Day, genetic line and infection status	No. differentially expressed protein-coding genes	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

Among DE lncRNAs, 6 lncRNAs showed fold changes > 100-fold following *Fp* challenge (Table 2, 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.

**Table 2:** 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 less than 0.05.

LncRNAs upregulated in all three genetic lines (> 2 fold) upon infection and their expression correlation with protein coding genes				
LncRNA	Fold change (Fp/PBS)			Correlation with (R <sup>2</sup> )
	Resistant line	Control line	Susceptible line	
Omy200117486	24.36	41.6	91.18	Interferon-induced guanylate-binding protein 1 (0.82)
Omy100128008	14.95	22.23	46.4	Complement protein component C7-1 (c7-1) (0.82)
Omy200138656	24.3	11.63	28.75	Complement C5 (0.66)
Omy100149048	7.93	5.95	14.15	Unknown
Omy200107378	6.38	11.22	11.22	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (0.92)
Omy200165911	3.68	4.5	9.98	Unknown
Omy100052789	5.51	5.13	8.65	Unknown
Omy200107535	3.71	6.16	8.12	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (0.92)
Omy300025398	4.37	5.15	4.86	Unknown
Omy300085997	3.68	3.16	4.02	Unknown
Omy200206941	3.16	2.33	3.44	Lysozyme C II precursor (0.83)
Omy300043066	3	3.74	3.03	Properdin (0.82) and complement factor b-like (0.89)
LncRNAs showing drastic (> 100) fold change upon infection in one particular genetic line and associated gene in genome				
LncRNAs	Fold change	Comparison	Classification of LncRNA	Associated
				Gene(s) (R <sup>2</sup> )
Upregulated upon infection				
Omy200018785	136.06	D1_S_FP vs D1_S_PBS	Intergenic	
Omy200132807	121.83	D5_S_FP vs D5_S_PBS	Intergenic	
Omy100037031	105.28	D5_C_FP vs D5_C_PBS	Intergenic	
Downregulated upon infection				
Omy200194608	-168.77	D1_S_FP vs D1_S_PBS	Genic, antisense	GSONMG00062425001 sich73- protein (0.27)
Omy200226560	-121.9	D1_R_Fp vs D1_R_PBS	Genic, antisense	GSONMG00065518001 (fatty-acyl reductase-1) (0.36)
Omy100064313	-108.56	D1_R_Fp vs D1_R_PBS	Intergenic	

## **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 (Al-Tobasei et al. 2016). Classification of all 556 DE lncRNA is given in appendix C. Lack of lncRNAs sequence conservation across species (Derrien et al. 2012) makes their annotation difficult. In addition, currently there are not 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 (Ørom et al. 2010, Tian, Sun and Lee 2010). 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. Mapping those 92 genes to KEGG pathways showed a total of 304 hits belonging to different pathways, of them about 11% (33/304) hits were involved in different microbial pathogenesis including hepatitis B, tuberculosis, shigellosis and other bacterial, viral and protozoal infections (Moriya et al. 2007). Similarly, 13.2 % of the hits (40/304) were directly involved in components of the immune system like Tumor necrosis factor (TNF) signaling, platelet activation, leukocyte trans-

endothelial migration, complement and coagulation, phagocytosis, intestinal immune network for IgA production, Natural killer cell mediated cytotoxicity, chemokine signaling and immune cell activation including T and B cell signaling. Out of 92 overlapped genes with DE lncRNA, only 36 genes had hits to KEGG pathways, of them 8 different genes were responsible for immunity pathways, 5 genes were responsible for microbial diseases, 6 genes were common in both pathways.

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. The DE lncRNAs and their overlapping protein-coding genes with a strong expression correlation are listed in Table 3. 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 (NE et al. 2014).

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 ( $R^2 = 0.80$ ) (Figure 1, 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. *GPBI* gene transcript also shows correlated expression with lncRNA in human (Barriocanal et al. 2014). Previous reports suggested that *gbp1* is one of the differentially regulated immune response genes against microbial pathogens in salmon and trout (Jeffries et al. 2014, Marancik et al. 2014).

**Table 3:** Correlation between expression patterns of lncRNAs and their overlapping protein-coding genes ( $R^2 > 0.70$ ).

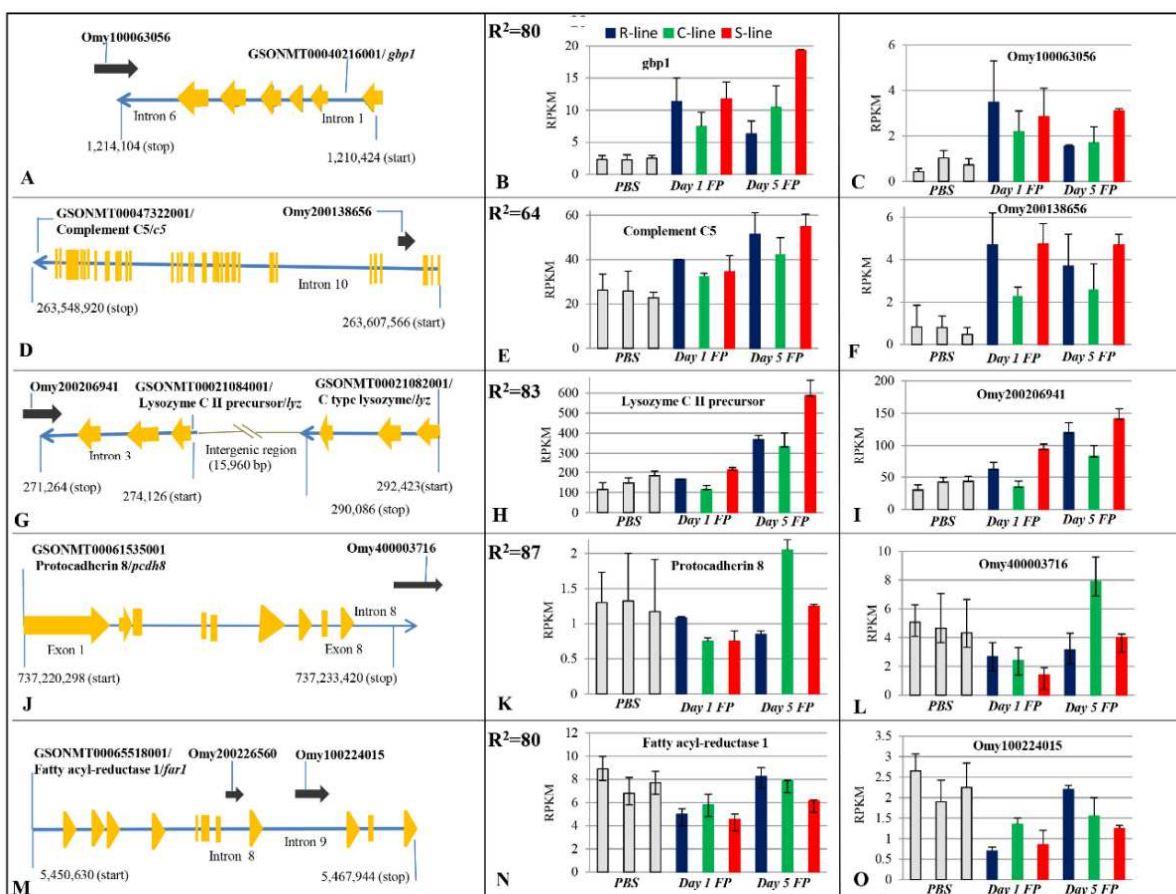
LncRNA	Neighboring protein-coding genes	LncRNA type	Correlation ( $R^2$ )	Neighboring gene name
Omy100063056	GSONMT00040216001	Intronic	Positive (0.73)	Interferon-induced guanylate-binding protein 1-like
		Antisense		
Omy200083892	GSONMT00050654001	Intronic	Positive (0.84)	Tumor necrosis factor receptor superfamily member 9-like
		Antisense		
Omy200080884	GSONMT00034829001	Exonic	Positive (0.93)	Response gene to complement 32 protein (rgc32)
		Antisense		
Omy200206941	GSONMT00021084001	Intronic	Positive (0.83)	Lysozyme C II precursor
		Unknown		
Omy200107012	GSONMT00019341001	Intronic	Positive (0.89)	Stromal interaction molecule 2-like
		Unknown		
Omy100228715	GSONMT00079494001	Exonic	Positive (0.83)	Unnamed protein product/transcobalamin-1 like
		Unknown		
Omy400008156	GSONMT00041383001	Intronic	Positive (0.87)	Reticulon-2 like
		Unknown		
Omy300038945	GSONMT00049537001	Intronic	Positive (0.81)	Cytochrome P450 7B1
		Antisense		
Omy400006181	GSONMT00049631001	Intronic	Positive (0.77)	Collagen alpha-1(IX) chain-like
		Unknown		
Omy400003716	GSONMT00061535001	Intronic	Positive (0.87)	Protocadherin 8 (pcdh8)
		Sense		
Omy100224015	GSONMT00065518001	Intronic	Positive (0.71)	Fatty acyl-CoA reductase 1 (facr1)
		Antisense		
Omy200181316	GSONMT00071779001	Exonic	Positive (0.82)	Muscular LMNA-interacting protein
		Unknown		
Omy100171980	GSONMT00073108001	Exonic	Positive (0.82)	Immunoglobulin-like and fibronectin type III domain-containing protein 1
		Unknown		

LncRNA Omy200128656 was located in intron 11 of complement C5 (*c5*) (GSONMT00047322001) gene in antisense orientation and their expression was positively

correlated ( $R^2 = 0.64$ ) (Figure 1, 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. Like in Rainbow trout, human C5 also shows positively correlated expression pattern with a lncRNA called C5T1lncRNA located in 3' UTR of the gene (Messemaeker et al. 2016). LncRNA Omy200206941 was partially overlapped with intron 4 of lysozyme CII precursor (*lyz*) (GSONMT00021084001) gene in antisense orientation and the expression was positively correlated ( $R^2 = 0.83$ ) (Figure 1, 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 ( $R^2 = 0.88$ ). All these three transcripts showed upregulation on day 5 post-challenge, in accordance with lysozyme's role in salmonid immunity (Saurabh and Sahoo 2008). Consistent with this correlated expression, regulation of lysozyme expression by lncRNA has been reported in other species (Lefevre et al. 2008). LncRNA Omy400003716 partially overlapped with intron 8 of protocadherin 8 (*pcdh8*) (GSONMT00061535001) in sense orientation and the expression was highly positively correlated ( $R^2=0.87$ ) (Figure 1, 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 ( $R^2$ ) of 0.36 and 0.80 respectively (Figure 1, 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 (Appendix D).



**Figure 1:** 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.

***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 (Moriya et al. 2007). Out of 464 DE intergenic lncRNAs, 371 had protein-coding genes within 50 kb distance in the genome. KEGG search of these neighboring protein-coding genes identified total of 1345 hits belonging to different pathways (Moriya et al. 2007). Pathway analysis showed that about 17% (227/1345) of the hits had known functions related to host defense system or microbial pathogenesis. Total of 290 genes neighboring to DE lncRNAs had hits to KEGG pathways, of them 51 different genes were responsible for immunity pathways, 49 genes were responsible for microbial infection processes and 23 genes were common in both of these pathways.

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. 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 4). 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 (Hu et al. 2013, NE et al. 2014).

**Table 4:** Correlation between expression patterns of lncRNAs and their intergenic neighboring protein-coding genes (within < 50 kb and  $R^2 > 0.70$ ).

LncRNA	Neighboring protein-coding genes (ID)	Distance from LncRNA (KB)	Direction relative to LncRNA	Expression correlation type ( $R^2$ )	Neighboring gene name
Omy200174653	GSONMT00031633001	5	Unknown/Intergenic	Positive (0.92)	Complement factor D-like
Omy300084989	GSONMT00013116001	2.6	Antisense/Intergenic	Positive (0.71)	Suppressor of cytokine signaling
Omy300074800	GSONMT00003195001	1.1	Unknown/Intergenic	Positive (0.79)	Ninjurin-1
Omy200206941	GSONMT00021082001	18.6	Unknown/Intergenic	Positive (0.88)	C type lysozyme
Omy200073559	GSONMT00017721001	3.5	Unknown/Intergenic	Positive (0.77)	Ceramide-1-phosphate transfer protein-like
Omy200061208	GSONMT00041695001	0.3	Unknown/Intergenic	Positive (0.90)	Coiled-coil transcriptional coactivator b
Omy200112536	GSONMT00001821001	18.2	Unknown/Intergenic	Positive (0.88)	Serum albumin 1
Omy300087476	GSONMT00010387001	0.9	Unknown/Intergenic	Positive (0.83)	Neutral amino acid transporter B(0)
Omy200075445	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 (Martianov et al. 2007, Schmitz et al. 2010) 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 5). 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 5).

***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 2). Similarly, expression of Omy100124197 was strongly correlated with 8 different proteins including matrix metallo-proteinase (Astacin) (GSONMT00014156001), elastase-1

(GSONMT00002714001), natectin (GSONMT00024075001), phospholipase-A2 (GSONMT00073599001), and syncollin (GSONMT00034810001) ( $R^2 > 0.98$ ) (Figure 2). Role of these correlated proteins in the immune system has already been characterized in different species (Belaouaj et al. 1998, Belaouaj, Kim and Shapiro 2000, Belaouaj 2002, Parks, Wilson and López-Boado 2004, Bach et al. 2006, Lopes-Ferreira et al. 2011, Saraiva et al. 2011). Several studies have also reported correlated expression of several immune related protein-coding genes with a single lncRNA (Carpenter et al. 2013).

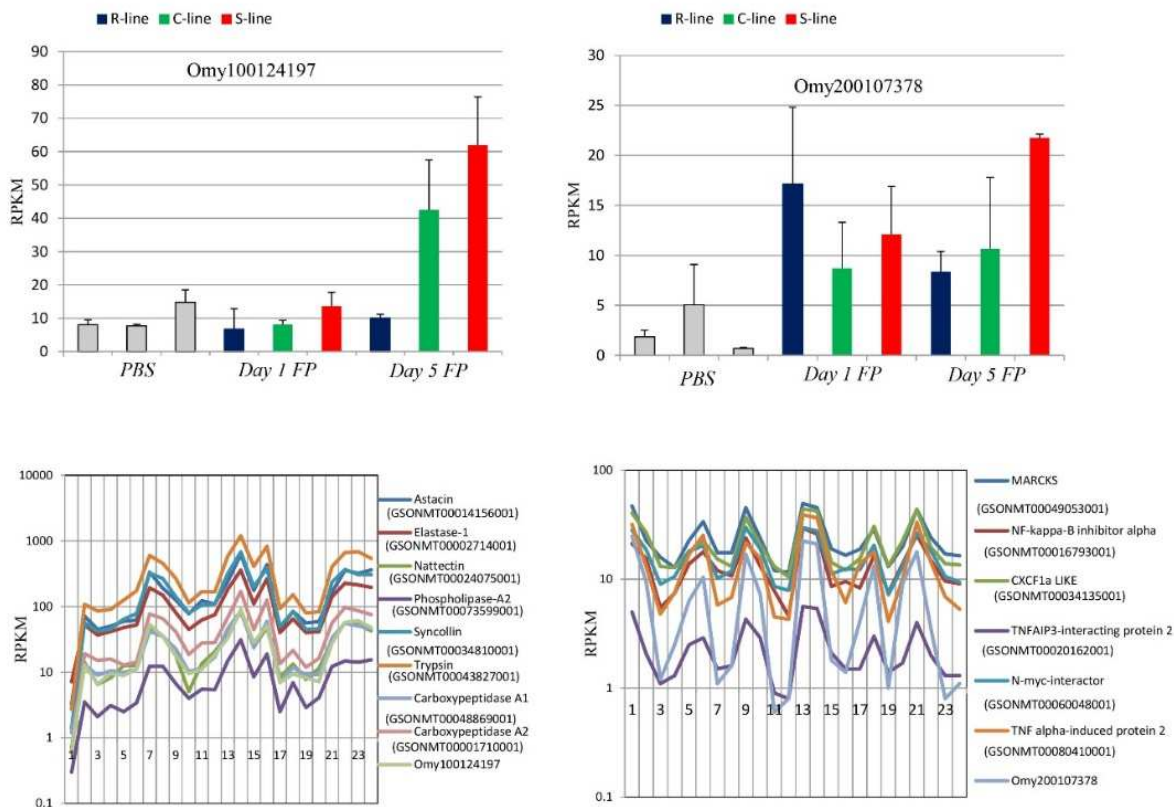
**Table 5:** Correlation between expression patterns of lncRNAs and some distantly located (> 50 kb or different chromosome) immune-relevant protein-coding genes.

LncRNA	Protein-coding genes (ID)	Expression correlation type ( $R^2$ )	Name of coding gene
Omy100104455	GSONMT00024124001	Positive (0.96)	Chemokine CK1
Omy200174653	GSONMT00051250001	Positive (0.92)	C-C motif chemokine 19 precursor
Omy300084989	GSONMT00062775001	Positive (0.83)	C4b-binding protein alpha chain precursor
Omy300041057	GSONMT00042009001	Positive (0.80)	Caspase-8
Omy300043066	GSONMT00001792001	Positive (0.82)	Properdin
Omy300043066	GSONMT00027840001	Positive (0.89)	Complement factor b-like
Omy200100893	GSONMT00051250001	Positive (0.92)	C-C motif chemokine 19 precursor
Omy200107378	GSONMT00016681001	Positive (0.92)	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2
Omy200107535	GSONMT00016681001	Positive (0.92)	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2
Omy200117486	GSONMT00005714001	Positive (0.82)	Interferon-induced guanylate-binding protein 1
Omy100066751	GSONMT00080410001	Positive (0.84)	Tumor necrosis factor, alpha-induced protein 2 (tnfaip2)
Omy100128008	GSONMT00070499001	Positive (0.82)	Complement protein component C7-1 (c7-1)
Omy100063056	GSONMT00075049001	Positive (0.85)	Tumor necrosis factor, alpha-induced protein 3 (tnfaip3)
Omy200053140	GSONMT00071335001	Negative (0.84)	NF-kappa-B inhibitor alpha

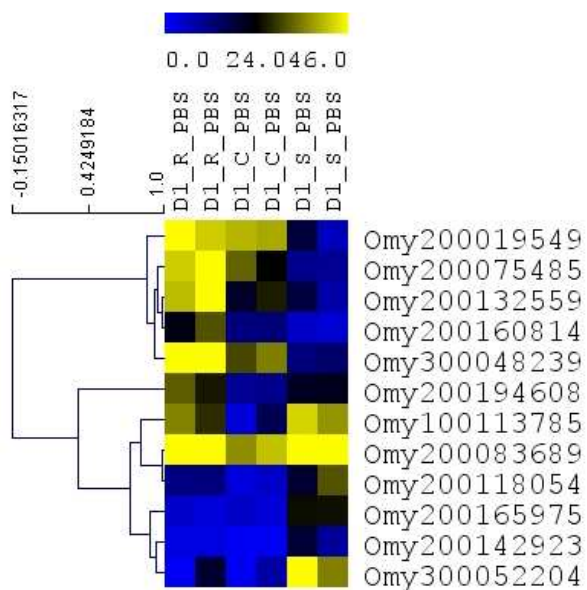
### **LncRNAs expression of naïve fish in different genetic lines**

Three genetic lines of Rainbow trout used in this study had significant differences in infection susceptibility to *Fp* as a result of selective breeding (Marancik et al. 2014). To investigate the potential basic 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. In our previous study, we identified differences in transcriptome abundance of protein-coding genes among naïve genetic lines (Marancik et al. 2014). The numbers of DE lncRNAs were smaller but consistent with the numbers of DE protein-coding genes among different naïve genetic lines (Table 1). 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 3). 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 (Liu et al. 2012, Kumar et al. 2013).





**Figure 2:** 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.

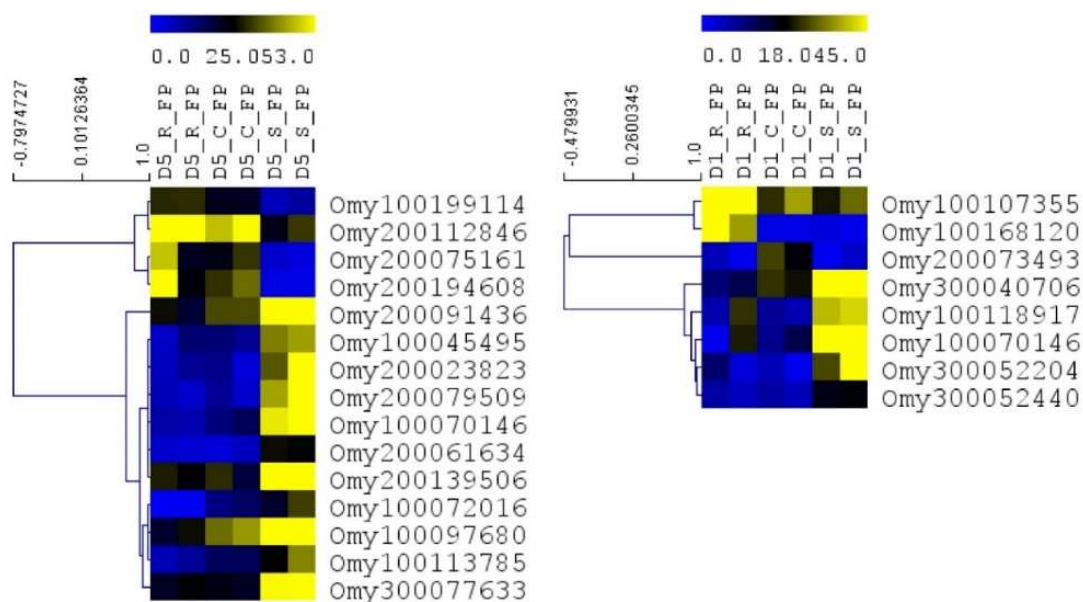


**Figure 3:** Comparison of transcriptome abundance of selected lncRNAs among naïve fish in all genetic lines. Genes are hierarchically clustered based on their expression pattern. D1 indicates day 1 post challenge and PBS indicates PBS injection. C, R and S represent control, resistant and susceptible 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 1<sup>st</sup> and 5<sup>th</sup> days of infection (Table 1). On 5<sup>th</sup> 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 1<sup>st</sup> day of infection, these numbers were 15, 12 and 13 respectively. Similarly, on the 1<sup>st</sup> 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 (Pawar et al. 2016). Figure 4 shows abundance of selected hierarchically clustered lncRNAs among genetic lines after infection with *Fp*. On the 5<sup>th</sup> 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 4).

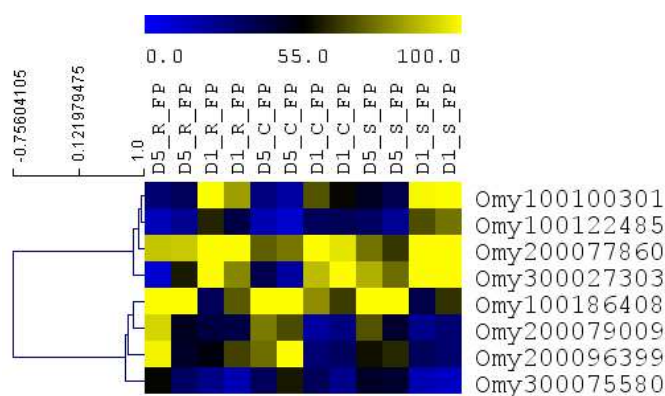


**Figure 4:** Comparison of transcriptome abundance of selected lncRNAs among genetic lines after infection with *Fp*. Genes are hierarchically clustered based on their expression pattern. D1 and D5 indicate day 1 and day 5 of sampling after injection. *Fp* indicates *Fp* injection. C, R and S represent 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 (Table 1). 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 (Boltana et

al. 2016). Figure 5 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 4). 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 5). NFKB2 is a transcription factor required to maintain normal level of antigen specific antibody production in response to antigen challenge (Caamaño et al. 1998). 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 2). 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 (Scaria and Pasha 2012).



**Figure 5:** 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 (Boltana et al. 2016). 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) (Appendix E). 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. 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 5) and was one of the 12

lncRNAs that were upregulated during infection in all three genetic lines relative to their PBS injected fish (Table 2). 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 (Al-Tobasei et al. 2016). 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 (Appendix F). However, none of these proteins was overlapped with lncRNA or had previously described role in immune system. 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. 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 (Appendix G). Similarly, three lncRNA (dis\_S\_00030301, dis\_S\_00043616 and dis\_S\_00083595) were susceptible-line specific. 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 (Appendix G). These findings may suggest 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 (Peng et al. 2010, Carpenter et al. 2013, Gomez et al. 2013, Hu et al. 2013, Kambara et al. 2014, Wang et al. 2014, Xia et al. 2014, Boltana et al. 2016). Manipulation of individual lncRNAs is sufficient to change the expression of hundreds of immune response genes (Carpenter et al. 2013), and variation in expression of other lncRNA's alter host susceptibility to different microbial pathogens (Gomez et al. 2013). 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 (Nematollahi et al. 2003). 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 (Derrien et al. 2012). 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 (Cabili et al. 2011). 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.

## REFERENCES

- Al-Tobasei, R., B. Paneru & M. Salem (2016) Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout. *PLoS One*, 11, e0148940.
- Asche, F., H. Håvard, T. Ragnar & T. Sigbjørn (2009) The salmon disease crisis in Chile. *Marine Resource Economics*, 24, 405-411.
- Bach, J. P., H. Borta, W. Ackermann, F. Faust, O. Borchers & M. Schrader (2006) The secretory granule protein syncollin localizes to HL-60 cells and neutrophils. *J Histochem Cytochem*, 54, 877-88.
- Barriocanal, M., E. Carnero, V. Segura & P. Fortes (2014) Long Non-Coding RNA BST2/BISPR is Induced by IFN and Regulates the Expression of the Antiviral Factor Tetherin. *Front Immunol*, 5, 655.
- Belaouaj, A. (2002) Neutrophil elastase-mediated killing of bacteria: lessons from targeted mutagenesis. *Microbes Infect*, 4, 1259-64.
- Belaouaj, A., K. S. Kim & S. D. Shapiro (2000) Degradation of outer membrane protein A in Escherichia coli killing by neutrophil elastase. *Science*, 289, 1185-8.
- Belaouaj, A., R. McCarthy, M. Baumann, Z. Gao, T. J. Ley, S. N. Abraham & S. D. Shapiro (1998) Mice lacking neutrophil elastase reveal impaired host defense against gram negative bacterial sepsis. *Nat Med*, 4, 615-8.
- Berthelot, C., F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noel, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J. M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Bousaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J. N. Volff, C. Genet, P. Wincker, O. Jaillon, H. Roest Crollius & Y. Guiguen (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*, 5, 3657.
- Bolstad, B. M., R. A. Irizarry, M. Astrand & T. P. Speed (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19, 185-93.
- Boltana, S., D. Valenzuela-Miranda, A. Aguilar, S. Mackenzie & C. Gallardo-Escarate (2016) Long noncoding RNAs (lncRNAs) dynamics evidence immunomodulation during ISAV-Infected Atlantic salmon (*Salmo salar*). *Sci Rep*, 6, 22698.
- Brown, L., W. Cox & R. Levine (1997) Evidence that the causal agent of bacterial cold-water disease *Flavobacterium psychrophilum* is transmitted within salmonid eggs. *Diseases of Aquatic Organisms*, 29, 213-218.
- Caamaño, J. H., C. A. Rizzo, S. K. Durham, D. S. Barton, C. Raventós-Suárez, C. M. Snapper & R. Bravo (1998) Nuclear factor (NF)-kappa B2 (p100/p52) is required for normal splenic microarchitecture and B cell-mediated immune responses. *J Exp Med*, 187, 185-96.
- Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev & J. L. Rinn (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 25, 1915-27.
- Carpenter, S., D. Aiello, M. K. Atianand, E. P. Ricci, P. Gandhi, L. L. Hall, M. Byron, B. Monks, M. Henry-Bezy, J. B. Lawrence, L. A. O'Neill, M. J. Moore, D. R. Caffrey

- & K. A. Fitzgerald (2013) A long noncoding RNA mediates both activation and repression of immune response genes. *Science*, 341, 789-92.
- Carson, L. & J. Schmidtke (1995) Characteristics of *Flexibacter psychrophilus* isolated from Atlantic salmon in Australia. *Diseases of Aquatic Organisms*, 21, 157-161.
- Collier, S. P., P. L. Collins, C. L. Williams, M. R. Boothby & T. M. Aune (2012) Cutting edge: influence of Tmevpg1, a long intergenic noncoding RNA, on the expression of Ifng by Th1 cells. *J Immunol*, 189, 2084-8.
- Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhata, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow & R. Guigó (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, 22, 1775-89.
- Gjedrem, T. (2005) *Selection and breeding programs in aquaculture*. Dordrecht: Springer.
- Gómez, E., J. Méndez, D. Cascales & J. A. Guijarro (2014) *Flavobacterium psychrophilum* vaccine development: a difficult task. *Microb Biotechnol*, 7, 414-23.
- Gomez, J. A., O. L. Wapinski, Y. W. Yang, J. F. Bureau, S. Gopinath, D. M. Monack, H. Y. Chang, M. Brahic & K. Kirkegaard (2013) The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. *Cell*, 152, 743-54.
- Hu, G., Q. Tang, S. Sharma, F. Yu, T. M. Escobar, S. A. Muljo, J. Zhu & K. Zhao (2013) Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat Immunol*, 14, 1190-8.
- Jeffries, K. M., S. G. Hinch, M. K. Gale, T. D. Clark, A. G. Lotto, M. T. Casselman, S. Li, E. L. Rechisky, A. D. Porter, D. W. Welch & K. M. Miller (2014) Immune response genes and pathogen presence predict migration survival in wild salmon smolts. *Mol Ecol*, 23, 5803-15.
- Kambara, H., F. Niazi, L. Kostadinova, D. K. Moonka, C. T. Siegel, A. B. Post, E. Carnero, M. Barriocanal, P. Fortes, D. D. Anthony & S. Valadkhan (2014) Negative regulation of the interferon response by an interferon-induced long non-coding RNA. *Nucleic Acids Res*, 42, 10668-80.
- Kent, L., J. Groff, J. Morrison, W. Yasutake & R. Holt (1989) Spiral swimming behavior due to cranial and vertebral lesions associated with *Cytophaga psychrophila* infections in salmonid fishes. *Diseases of Aquatic Organisms*, 6, 11-16.
- Kumar, V., H. J. Westra, J. Karjalainen, D. V. Zhernakova, T. Esko, B. Hrdlickova, R. Almeida, A. Zhernakova, E. Reinmaa, U. Vosa, M. H. Hofker, R. S. Fehrmann, J. Fu, S. Withoff, A. Metspalu, L. Franke & C. Wijmenga (2013) Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet*, 9, e1003201.
- Lefevre, P., J. Witham, C. E. Lacroix, P. N. Cockerill & C. Bonifer (2008) The LPS-induced transcriptional upregulation of the chicken lysozyme locus involves CTCF eviction and noncoding RNA transcription. *Mol Cell*, 32, 129-39.

- Liu, Y., S. Pan, L. Liu, X. Zhai, J. Liu, J. Wen, Y. Zhang, J. Chen, H. Shen & Z. Hu (2012) A genetic variant in long non-coding RNA HULC contributes to risk of HBV-related hepatocellular carcinoma in a Chinese population. *PLoS One*, 7, e35145.
- Lopes-Ferreira, M., G. S. Magalhães, J. H. Fernandez, I. e. L. Junqueira-de-Azevedo, P. Le Ho, C. Lima, R. H. Valente & A. M. Moura-da-Silva (2011) Structural and biological characterization of Nattectin, a new C-type lectin from the venomous fish *Thalassophryne nattereri*. *Biochimie*, 93, 971-80.
- Madetoja, J., P. Nyman & T. Wiklund (2000) *Flavobacterium psychrophilum*, invasion into and shedding by rainbow trout *Oncorhynchus mykiss*. *Diseases of Aquatic Organisms*, 43, 27-38.
- Marancik, D., G. Gao, B. Paneru, H. Ma, A. G. Hernandez, M. Salem, J. Yao, Y. Palti & G. D. Wiens (2014) Whole-body transcriptome of selectively bred, resistant-, control-, and susceptible-line rainbow trout following experimental challenge with *Flavobacterium psychrophilum*. *Front Genet*, 5, 453.
- Martianov, I., A. Ramadass, A. Serra Barros, N. Chow & A. Akoulitchev (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature*, 445, 666-70.
- Messemaker, T. C., M. Frank-Bertoncelj, R. B. Marques, A. Adriaans, A. M. Bakker, N. Daha, S. Gay, T. W. Huizinga, R. E. Toes, H. M. Mikkers & F. Kurreeman (2016) A novel long non-coding RNA in the rheumatoid arthritis risk locus TRAF1-C5 influences C5 mRNA levels. *Genes Immun*, 17, 85-92.
- Moriya, Y., M. Itoh, S. Okuda, A. C. Yoshizawa & M. Kanehisa (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*, 35, W182-5.
- NE, I. I., J. A. Heward, B. Roux, E. Tsitsiou, P. S. Fenwick, L. Lenzi, I. Goodhead, C. Hertz-Fowler, A. Heger, N. Hall, L. E. Donnelly, D. Sims & M. A. Lindsay (2014) Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. *Nat Commun*, 5, 3979.
- Nematollahi, A., A. Decostere, F. Pasmans & F. Haesebrouck (2003) *Flavobacterium psychrophilum* infections in salmonid fish. *J Fish Dis*, 26, 563-74.
- Ørom, U. A., T. Derrien, M. Beringer, K. Gumireddy, A. Gardini, G. Bussotti, F. Lai, M. Zytnicki, C. Notredame, Q. Huang, R. Guigo & R. Shiekhattar (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143, 46-58.
- Parks, W. C., C. L. Wilson & Y. S. López-Boado (2004) Matrix metalloproteinases as modulators of inflammation and innate immunity. *Nat Rev Immunol*, 4, 617-29.
- Pawar, K., C. Hanisch, S. E. Palma Vera, R. Einspanier & S. Sharbati (2016) Down regulated lncRNA MEG3 eliminates mycobacteria in macrophages via autophagy. *Sci Rep*, 6, 19416.
- Peng, X., L. Gralinski, C. D. Armour, M. T. Ferris, M. J. Thomas, S. Prohl, B. G. Bradel-Tretheway, M. J. Korth, J. C. Castle, M. C. Biery, H. K. Bouzek, D. R. Haynor, M. B. Frieman, M. Heise, C. K. Raymond, R. S. Baric & M. G. Katze (2010) Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling. *MBio*, 1.

- Salem, M., B. Paneru, R. Al-Tobasei, F. Abdouni, G. H. Thorgaard, C. E. Rexroad & J. Yao (2015) Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout. *PLoS One*, 10, e0121778.
- Saraiva, T. C., L. Z. Grund, E. N. Komegae, A. D. Ramos, K. Conceição, N. M. Orii, M. Lopes-Ferreira & C. Lima (2011) Nattectin a fish C-type lectin drives Th1 responses in vivo: licenses macrophages to differentiate into cells exhibiting typical DC function. *Int Immunopharmacol*, 11, 1546-56.
- Saurabh, S. & P. K. Sahoo (2008) Lysozyme: an important defence molecule of fish innate immune system. *Aquaculture Research*, 39, 223-239.
- Scaria, V. & A. Pasha (2012) Long Non-Coding RNAs in Infection Biology. *Front Genet*, 3, 308.
- Schmittgen, T. D. & K. J. Livak (2008) Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc*, 3, 1101-8.
- Schmitz, K. M., C. Mayer, A. Postepska & I. Grummt (2010) Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev*, 24, 2264-9.
- Tian, D., S. Sun & J. T. Lee (2010) The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell*, 143, 390-403.
- Wang, P., Y. Xue, Y. Han, L. Lin, C. Wu, S. Xu, Z. Jiang, J. Xu, Q. Liu & X. Cao (2014) The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. *Science*, 344, 310-3.
- Wiens, G. D., L. Scott E, W. Timothy J, E. Jason P, R. Caird E & L. Timothy D (2013) 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*, 388, 128-136.
- Xia, F., F. Dong, Y. Yang, A. Huang, S. Chen, D. Sun, S. Xiong & J. Zhang (2014) Dynamic transcription of long non-coding RNA genes during CD4<sup>+</sup> T cell development and activation. *PLoS One*, 9, e101588.

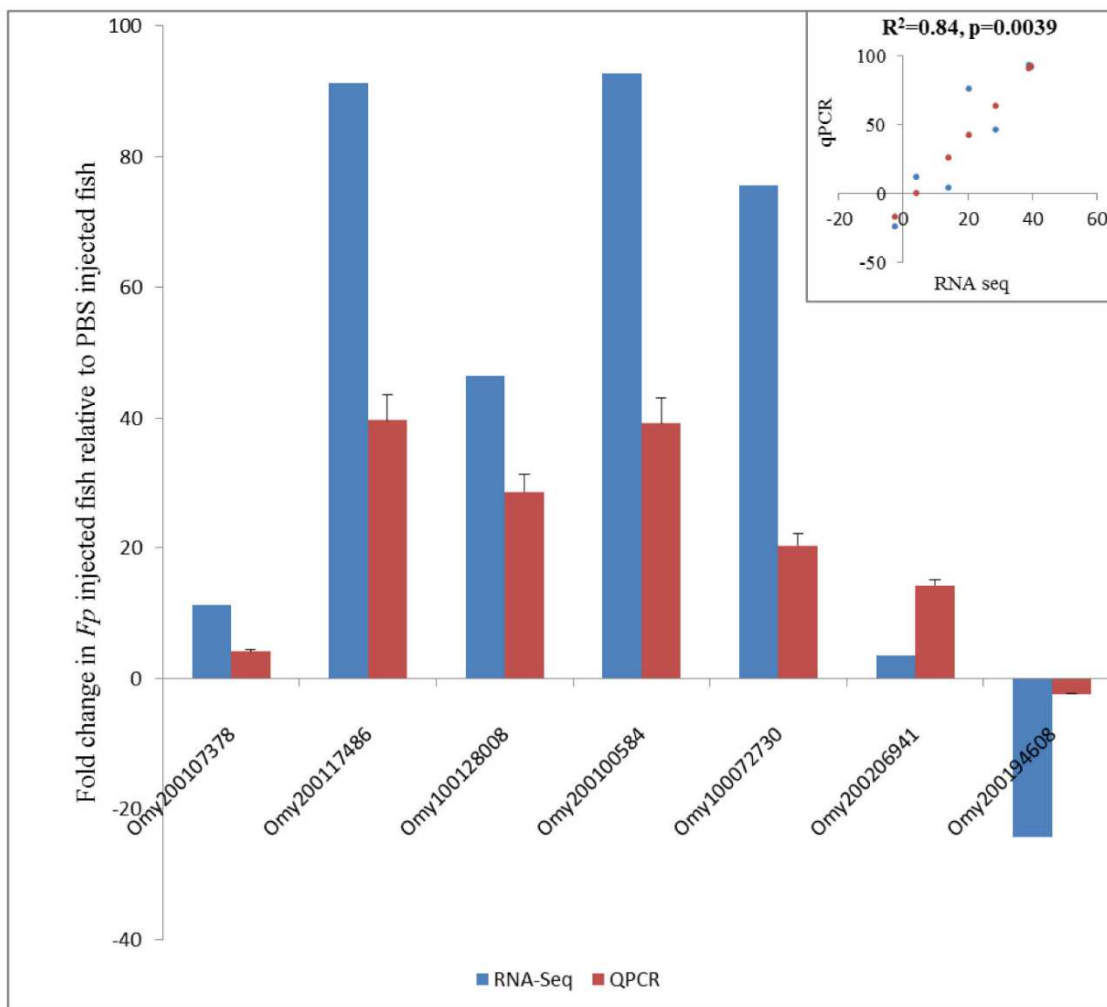
**APPENDICES**

## APPENDIX A

**SUMMARY STATISTICS OF 24 RNA SEQ LIBRARIES WITH DIFFERENT  
GENETIC LINES, TIME, INFECTION STATUS AND TANK REPLICATE**

Genetic line	Day	Infection, Tank	Biosample Accession No	Total reads	Mapped reads	Percentage reads mapped (%)	Uniquely mapped reads	Total number of expressed lncRNAs (RPKM $\geq 5.0$ )	Total number of expressed lncRNAs (RPKM $\geq 1.0$ )	Total number of expressed lncRNAs (RPKM $\geq 0.50$ )	% of expressed genes (RPKM $\geq 0.50$ )
ARS-Fp-R	1	Fp, Tk25	SAMN03014722	20,061,852	1,761,205	8.8	1,663,742	19,540	26,371	27,561	88.4
		Fp, Tk26	SAMN03014723	20,226,280	1,672,737	8.3	1,580,463	17,868	24,824	26,156	83.8
	5	Fp, Tk25	SAMN03014726	21,409,329	1,769,707	8.3	1,672,380	18,718	25,562	26,834	86
		Fp, Tk26	SAMN03014727	23,958,681	2,038,391	8.5	1,929,750	19,728	26,518	27,885	89.4
	1	PBS, Tk27	SAMN03014724	22,129,914	1,795,272	8.1	1,697,092	19,482	26,241	27,491	88.1
		PBS, Tk28	SAMN03014725	23,909,110	1,904,430	8	1,801,598	19,978	26,747	28,018	89.8
	5	PBS, Tk27	SAMN03014728	24,361,298	2,009,877	8.3	1,899,213	19,311	26,127	27,559	88.3
		PBS, Tk28	SAMN03014729	23,318,224	1,917,703	8.2	1,809,859	19,399	26,284	27,636	88.6
ARS-Fp-C	1	Fp, Tk33	SAMN03014738	20,940,097	1,734,138	8.3	1,642,815	20,027	26,680	27,806	89.1
		Fp, Tk34	SAMN03014739	19,151,755	1,534,009	8	1,450,878	18,174	25,184	26,340	84.4
	5	Fp, Tk33	SAMN03014742	21,117,398	1,713,769	8.1	1,618,904	18,796	25,550	26,767	85.8
		Fp, Tk34	SAMN03014743	21,763,498	1,800,967	8.3	1,699,043	19,159	26,005	27,256	87.4
	1	PBS, Tk35	SAMN03014740	20,314,994	1,655,407	8.2	1,567,306	17,357	24,396	25,710	82.4
		PBS, Tk36	SAMN03014741	20,372,060	1,643,044	8.1	1,554,591	19,555	26,351	27,437	88
	5	PBS, Tk35	SAMN03014744	24,480,995	2,043,423	8.4	1,929,929	18,728	25,650	27,153	87
		PBS, Tk36	SAMN03014745	20,535,582	1,653,545	8.1	1,561,351	18,880	25,776	26,941	86.4
ARS-Fp-S	1	Fp, Tk29	SAMN03014730	17,389,164	1,392,049	8	1,317,455	19,201	25,958	26,878	86.2
		Fp, Tk30	SAMN03014731	21,095,859	1,737,084	8.2	1,644,537	19,005	25,820	27,089	86.8
	5	Fp, Tk29	SAMN03014734	19,446,430	1,556,366	8	1,473,769	19,378	26,118	27,172	87.1
		Fp, Tk30	SAMN03014735	23,427,268	1,887,381	8.1	1,784,521	19,520	26,237	27,526	88.2
	1	PBS, Tk31	SAMN03014732	20,622,290	1,631,915	7.9	1,545,298	18,861	25,810	26,975	86.5
		PBS, Tk32	SAMN03014733	24,139,642	1,961,550	8.1	1,856,305	19,849	26,643	27,910	89.5
	5	PBS, Tk31	SAMN03014736	22,371,928	1,820,511	8.1	1,718,566	19,576	26,350	27,607	88.5
		PBS, Tk32	SAMN03014737	22,338,190	1,772,639	7.9	1,676,640	19,054	25,976	27,257	87.4
Average				#####	#####	8.2	#####	19,131.00	25,965.80	27,206.80	87.2

## APPENDIX B

FOLD CHANGE COMPARISON OF SELECTED DIFFERENTIALLY  
EXPRESSED GENES BY RNA-SEQ AND REAL TIME PCR

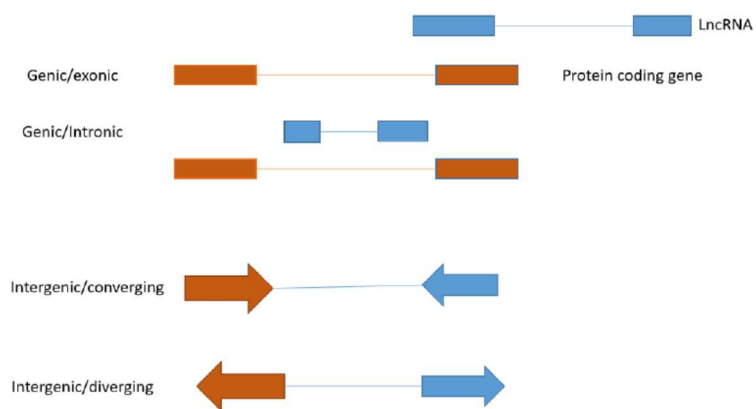


## APPENDIX C

**CLASSIFICATION OF DE LNCRNA IN RESPONSE TO FP CHALLENGE  
BASED ON THEIR INTERSECTION WITH PROTEIN CODING GENES AND  
NUMBER OF LNCRNAs IN EACH CLASS**

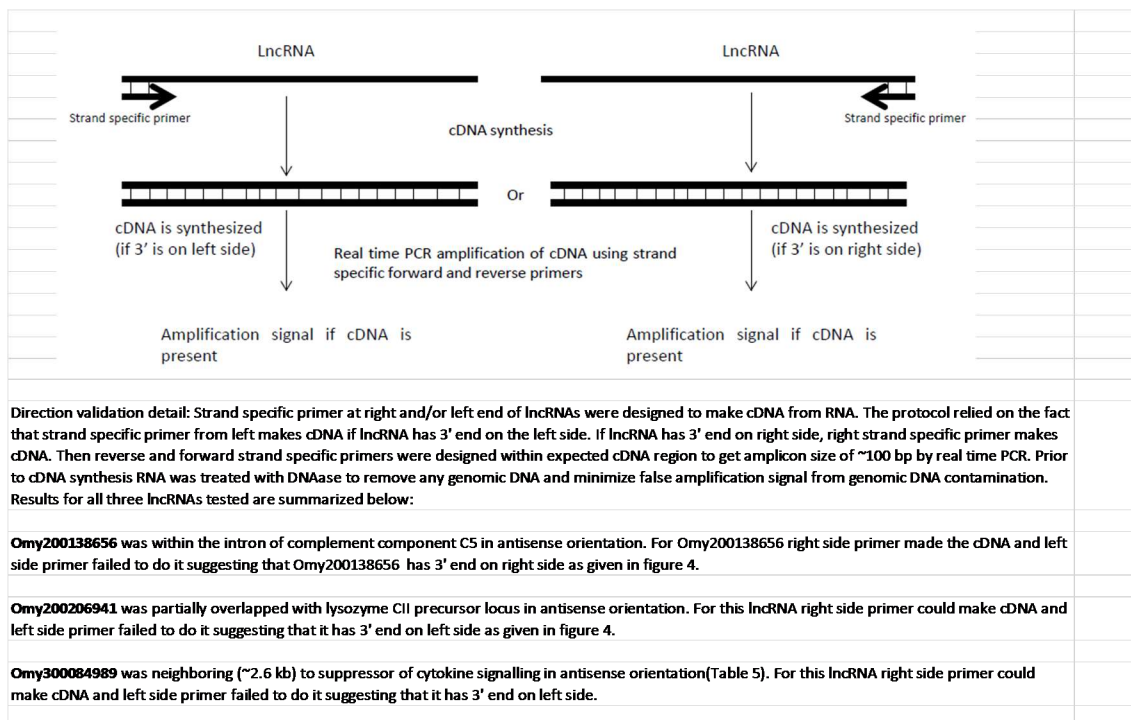
**Classification of differentially expressed lncRNAs**

Genic (92)						Intergenic (464)		
Exonic (28)			Intronic (64)			Converging	Diverging	Unknown
Sense	Antisense	Unknown	Sense	Antisense	Unknown			
0	6	22	4	24	34	58	62	344



## APPENDIX D

### STRAND SPECIFIC PCR METHOD USED IN VALIDATION OF STRAND ORIENTATION OF SOME OF LNCRNA TRANSCRIPT RELATIVE TO THEIR PROTEIN CODING LOCI COUNTERPARTS



## APPENDIX E

## DIFFERENTIALLY EXPRESSED (DE) LNCRNAs CONSERVED IN ATLANTIC

## SALMON

Trout LncRNA	Conserved Salmon lncRNA	Query (Trout lncRNA) cover by alignment (%)	Sequence identity (%)	E value
Omy100076105	Ss_lncRNA_4661	98.56	97.44	4.00E-138
Omy100094622	Ss_lncRNA_210	97.48	90.52	4.00E-64
Omy300077633	Ss_lncRNA_5079	96.77	91.61	0
Omy100065769	Ss_lncRNA_4505	90.04	91.83	1.00E-148
Omy100133784	Ss_lncRNA_2292	87.14	90.71	1.00E-60
Omy200084095	Ss_lncRNA_2682	84.12	92.41	0
Omy300043379	Ss_lncRNA_1263	81.32	92.79	2.00E-81
Omy100089327	Ss_lncRNA_1531	72.03	94.16	5.00E-161
Omy300045859	Ss_lncRNA_622	71.16	95.42	1.00E-66
Omy100070292	Ss_lncRNA_329	67.31	93.6	0
Omy100125126	Ss_lncRNA_5400	66.03	92.56	4.00E-121
Omy200084059	Ss_lncRNA_4410	60.22	94.12	1.00E-92
Omy200115330	Ss_lncRNA_5535	60.19	93.85	2.00E-50
Omy200119000	Ss_lncRNA_3481	59.74	90.87	0
Omy300029193	Ss_lncRNA_4515	58.82	92.31	1.00E-45
Omy100111613	Ss_lncRNA_5503	57.25	90	5.00E-99
Omy300052204	Ss_lncRNA_791	56.40	92.2	8.00E-50
Omy200107755	Ss_lncRNA_4416	55.78	90.19	9.00E-120
Omy300045687	Ss_lncRNA_3330	55.66	93.79	9.00E-69
Omy200181424	Ss_lncRNA_3473	54.55	93.56	9.00E-107
Omy200136262	Ss_lncRNA_1883	51.23	92.92	8.00E-142
Omy200129846	Ss_lncRNA_282	50.30	92.09	8.00E-95
Omy300043066	Ss_lncRNA_1643	50.07	91.62	1.00E-122

## APPENDIX F

### NOVEL LNCRNAs SPECIFIC TO RESISTANT OR SUSCEPTIBLE LINES, AND THEIR EXPRESSION CORRELATION WITH PROTEIN CODING GENE

LncRNA	Protein coding gene ID	Protein coding gene name	Expression correlation (R <sup>2</sup> )	Type of correlation
dis_R_00008399	GSONMT00054615001	potassium-transporting atpase alpha chain 1	0.75	Positive
dis_S_00062412	GSONMT00076273001	protein fam65a	0.75	Positive
dis_S_00053564	GSONMT00071789001	polyunsaturated fatty acid elongase	0.75	Positive
dis_S_00057495	GSONMT00005124001	low quality protein: chloride channel protein -kb-like	0.74	Positive
dis_S_00012584	GSONMT00071637001	sushi domain-containing protein 1-like	0.73	Negative
dis_R_00069415	GSONMT00078206001	lipoma hmgic fusion partner	0.71	Positive
dis_R_00074624	GSONMT00060250001	slow myosin heavy chain 1	0.70	Positive
dis_S_00075739	GSONMT00015921001	heat shock 70 kda protein 4l	0.70	Positive
dis_S_00032921	GSONMT00050798001	transmembrane protein 68	0.70	Negative

## APPENDIX G

**NOVEL LNCRNAs SPECIFIC TO RESISTANT AND SUSCEPTIBLE LINES  
AND THEIR RELATIVE EXPRESSION BETWEEN VARIOUS COMPARISONS**

<b>R line specific LncRNAs</b>		
<b>Feature ID</b>	<b>Fold Change (R line/S line)</b>	<b>FDR p-value correction</b>
dis_R_00048342	∞	0.01
dis_R_00050098	∞	5.61E-04
<b>S line specific LncRNAs</b>		
<b>Feature ID</b>	<b>Fold Change (S line/R line)</b>	<b>FDR p-value correction</b>
dis_S_00030301	∞	5.61E-04
dis_S_00043616	∞	0.02
dis_S_00083595	∞	7.99E-07
<b><i>Flavobacterium</i> infection associated LncRNAs</b>		
<b>Feature ID</b>	<b>Fold Change (Fp/PBS)</b>	<b>FDR p-value correction</b>
dis_R_00037228	∞	3.44E-09
dis_R_00037932	∞	3.46E-08
dis_R_00067703	∞	1.46E-30
dis_S_00002625	∞	2.55E-06
dis_R_00016845	-6546.59	4.78E-06
dis_S_00045735	∞	4.78E-06
dis_S_00034377	-5749.38	4.61E-04
dis_S_00078190	24.12	6.77E-13
dis_R_00057690	9.68	4.78E-06
dis_R_00035864	6.91	0.01
dis_S_00030209	4.09	0.01
dis_S_00063498	3.95	0.02
dis_R_00013962	2.84	0.03
dis_S_00010304	3.23	0.03
dis_R_00021365	10,253.19	5.08E-09
dis_R_00048869	-5295.19	0.01
dis_R_00036526	8,913.61	0.04
dis_S_00046889	-6,231.21	7.22E-05
dis_R_00052785	3.7	1.32E-03
dis_S_00014939	-8,831.02	0.05
<b>Infection time associated LncRNAs</b>		
<b>Feature ID</b>	<b>Fold Change (Day1/Day5)</b>	<b>FDR p-value correction</b>
dis_R_00052409	-7,336.42	4.00E-05
dis_R_00048869	-6422.15	3.03E-04
dis_R_00086335	∞	3.03E-04

## CHAPTER IV

### MICRORNA EXPRESSION AND GENETIC POLYMORPHISM ASSOCIATION WITH GROWTH AND MUSCLE QUALITY TRAITS IN RAINBOW TROUT

#### ABSTRACT

The roles of microRNA expression and genetic variation in microRNA-binding sites of target genes on growth and muscle quality traits are poorly characterized. In the present study, we used RNA-Seq approaches to investigate their importance on 5 growth and muscle quality traits: muscle yield of whole body weight (WBW), WBW, muscle whiteness, muscle shear force and crude-fat content. Phenotypic data were collected from 98 families (~500 fish) (~5 fish/family) from a growth-selected line. Muscle microRNAs and mRNAs were sequenced from 22 families showing divergent phenotypes. Ninety microRNAs showed differential expression between families with divergent phenotypes, and their expression was strongly associated with variation in phenotypes. A total of 204 single nucleotide polymorphisms (SNPs) present in 3' UTR of target genes either destroyed or created novel illegitimate microRNA target sites; of them, 72 SNPs explained significant variation in the aforementioned 5 muscle traits. Most SNPs were present in microRNA-binding sites of genes involved in energy metabolism and muscle structure. These findings suggest that variation in microRNA expression and/or sequence variation in microRNA binding sites in target genes play an important role in mediating differences in fish growth and muscle quality phenotypes.

#### INTRODUCTION

MicroRNAs are important post-transcriptional regulators of genes. In humans, about 30 percentage of genes are regulated by microRNAs (Lewis, Burge and Bartel 2005),

which suggests an important role of microRNAs (Zhang, Wang and Pan 2007). There is evidence that a single microRNA can regulate hundreds of genes whereas the same gene can be regulated by multiple microRNAs (Krek et al. 2005). These findings suggest that microRNAs play a crucial role in living organisms by their function in dynamic gene regulatory networks. During gene regulation, a mature sequence of microRNA (~22 nts) binds 3'-UTR of mRNA. The seed region in mature microRNA sequence, usually extending from 2-7 nts at 5' end, that binds to microRNA recognition element seed site (MRESS) in target gene, plays a vital role in determining specificity of microRNA-mRNA binding (Lewis et al. 2005). This 'microRNA-target mRNA' binding leads to downregulation of the gene by various mechanisms such as translation suppression (Olsen and Ambros 1999), mRNA cleavage (Bagga et al. 2005) and loss of mRNA poly A tail (Wu, Fan and Belasco 2006). Therefore, any mutation that either destroys or creates a novel illegitimate MRESS in target genes have important functional consequences in the phenotype (Clop et al. 2006, Georges et al. 2006).

Throughout their life, fish require the dynamic regulation of muscle mass as muscle proteins are mobilized in response to the energetic demands of exercise, starvation and gonadal maturation (Salem et al. 2006, Salem et al. 2013). MicroRNA mediated downregulation of genes plays critical role in embryonic myogenesis as well as post-embryonic skeletal muscle growth. Mir-1, mir-133 and mir-206 are myogenic microRNAs that control skeletal muscle growth by directly or indirectly regulating genes involved in myogenesis or muscle atrophy in mammals (see review (Wang 2013)). The loss of mir-206 function in Nile tilapia leads to accelerated muscle growth as insulin-like growth factor-1 (an important positive regulator of muscle growth) is directly targeted and repressed by

mir-206 (Yan et al. 2013b). In Zebra fish, mir-1 and mir-133 are responsible for more than half of the microRNA-mediated gene regulation in muscle including sarcomere assembly (Mishima et al. 2009). In addition to the myogenic microRNAs, non-muscle specific microRNAs also regulate different aspects of myogenesis and muscle development in fish. As an example, mir-143 and mir-203b target myoD, a member of myogenic regulatory factors (MRFs) which plays crucial role in myogenic differentiation, in Mandarin fish and Nile tilapia, respectively (Yan et al. 2013a, Chen et al. 2014). In addition, a novel Zebra fish microRNA, mir-In300, abolishes the promoter activity of myogenic protein 5 (*myf5*) via targeting dickkopf-3 gene (*dkk3*) that is required for *myf5* promoter activity (Hsu et al. 2010). Similarly, mir-214 controls hedgehog signaling mediated specification of muscle cell by negatively regulating suppressor of fused (*sufu*) mRNA in Zebra fish (Flynt et al. 2007). Further, Let-7, mir-19 and mir-130 show regulated expression during transition from muscular hyperplasia to hypertrophy in fish suggesting regulatory role in muscle development (Johnston et al. 2009). Despite the mounting evidence in previous studies that show an increasing role of microRNAs in gene regulation function, there is still need for complete microRNAome expression and genetic variant profiles in response to variation in muscle growth and muscle quality traits such as muscle tenderness, whiteness and crude fat content to understand genetic basis of these important production traits in fish.

Family based genomic selection introduces genetic variation for particular phenotypes in the population, which makes selectively-bred population a suitable model for ‘genotype-phenotype’ association analysis. United States Department of Agriculture, Agriculture Research Services (USDA, ARS), National Center for Cool and Cold Water Aquaculture (NCCCWA) started a family based growth selection program in 2002 for



Rainbow trout; the population has undergone 5 generations of selection so far (Leeds et al. 2016). The selection has achieved differences in growth performance and fillet quality traits among fish families in the population. The objectives of this study were 1) to investigate the association of microRNA expression with muscle growth and fillet quality traits in the USDA growth selected population and; 2) to investigate the effects of SNPs in microRNA binding sites on growth and muscle quality traits. In this study, using high throughput deep small RNA sequencing approach (RNAseq), we identified differentially expressed (DE) microRNAs between fish families showing contrasting phenotype for WBW (whole body weight), muscle % of WBW, crude fat content, shear-force and FWI (fillet whiteness index). We performed ‘phenotype-microRNA expression’ association in a larger fish set to investigate functional relevance of microRNAs expression to the phenotype. SNPs capable of creating or disrupting microRNA binding sites in protein coding target genes were identified and their functional consequence on growth and muscle quality phenotype was evaluated by SNPs in a large fish population.

## **MATERIALS AND METHODS**

### **Fish population and muscle sampling**

Phenotypic data and muscle samples were collected from 98 families (~500 fish) from USDA/NCCCWA growth selected trout line from each harvest year 2010 (Salem et al. 2012). Briefly, fish were reared until ~13 month post-hatch as previously described (Leeds et al. 2016). Single-sire x single-dam mating was used to produce full sib families and eggs were reared in spring water. Water temperature was adjusted from 7-13°C to synchronize the hatching time. Each family was reared in a separate 200 L tank at ~ 600 alevins/tank density. Random culling of fish was performed every month to maintain stock

density of  $< 50 \text{ kg/m}^3$ . At the age of 5 months, each fish was given a unique identification PIT (passive integrated transponder) tag, and were reared in a 1000 L commercial tanks. Commercial fishmeal-based diet (16% fat, 42% protein; Ziegler Bros Inc., Gardners, PA) was fed using an automatic feeder. The amount of fishmeal was gradually reduced from 2.5% of body weight to 0.5% of body weight as fish aged. Whole Body Weight (WBW) of all fish belonging to 98 families was estimated, and families were ranked based on their WBW measurements. Second or third ranked fish from each family was selected for muscle sampling so that WBW of sampled fish is adjusted around median of each family. Selected fish were randomly assigned to one of the 5 harvest group (~100 fish/harvest group) and each harvest groups were sampled in 5 consecutive weeks. Fish were anesthetized in 100 mg/L tricaine methanesulfonate and weighed, slaughtered and eviscerated. Muscle samples were separated from dorsal musculature and were stored in liquid nitrogen until processing. Muscle quality phenotypes were estimated at the West Virginia University (WVU) meat processing laboratory. Muscle yield was calculated as a percentage of WBW and the proximate analysis of muscle fillet was performed using a previously published protocol (Manor et al. 2015). Fillet whiteness were calculated in terms of  $a^*$  (redness),  $L^*$  (lightness), and  $b^*$  (yellowness) values and FWI (fillet whiteness index) was calculated using formula =  $100 - [(100 - L)^2 + a^2 + b^2]^{1/2}$  (Institute 1991). Muscle crude fat content was measured using Soxhlet solvent extractor with petroleum ether. For muscle peak shear force measurement, texture was analyzed using a five blade Allo-Kramer.

### **Library construction and sequencing**

White muscle sample was isolated from five individuals belonging to each family and total RNA was isolated using Trizol protocol (Invitrogen, Carlsbad, CA) as described previously (Salem et al. 2015, Tobasei et al. 2016). To ensure the total RNA from 5 individual fish from each family was pooled and sequenced on Illumina's HiSeq platform (Illumina Inc, CA, USA).

### **Data processing and prediction of trout microRNA**

Sequencing adapter 5' GCCTTGGCACCCGAGAATTCCA3' was trimmed and reads were annotated using miRBase microRNA reference (release 21) in CLC Bio small RNA analysis tool. The read alignment was run at default settings (i.e. mismatch  $\leq 2$ , additional/missing upstream/downstream bases  $\leq 2$ ). MicroRNAs with mismatches and/or additional/missing upstream or downstream nucleotides were considered variants of the same microRNA. The read count from all variants of the same microRNAs were summed and were used as expression value for that particular microRNA (default method of merging expression values in CLC genomics workbench small RNA analysis toolkit).

### **Identification of DE microRNAs**

DE microRNAs were identified using EDGE test in CLC genomics workbench using expression values from the above step. The fold change in gene expression between two groups was considered significant if FDR-p  $< 0.05$  and fold change  $< -2$  or  $> 2$ -fold.

### **Real time PCR validation of DE microRNAs and 'microRNA expression-phenotype variation' correlation**

Same RNA samples from high and low ranked families used for sequencing were used to validate RNA-Seq differential expression. cDNA was synthesized using miScript

II RT kit (Qiagen, Valencia, CA, USA) and microRNA was quantified in Bio-Rad CFX96™ Real Time System (Bio-Rad, Hercules, CA) using miScript<sup>R</sup> SYBR<sup>R</sup> green (Qiagen, Valencia, CA, USA). A non-coding RNA U6 was used as a endogenous control for normalization, and fold change was calculated by  $\Delta\Delta\text{Ct}$  method (Schmittgen and Livak 2008), as described previously (Marancik et al. 2014, Paneru et al. 2016).

The correlation between microRNA expression and phenotype was studied in 90 random individual fish from USDA/NCCCWA's 2010-harvest fish population (described above).  $\Delta\text{Ct}$  value of each microRNA in all 90 fish was estimated. Pearson correlation and simple linear regression were used to determine whether microRNA expression (by estimating  $\Delta\text{Ct}$ ) and phenotypic traits were associated.

### **Bioinformatics prediction of DE microRNA target**

For the prediction of microRNA targets, 3'-UTR of trout mRNA were retrieved from the genome reference (Berthelot et al. 2014). Due to difference in the sensitivity and specificity of different target prediction algorithms, targets were predicted using 3 tools: miRanda, PITA and TargetSpy in small RNA analysis server sRNAtoolbox (Rueda et al. 2015). If the same target site is predicted by all 3 tools, it was considered a potential microRNA target. For PITA, prediction parameter chosen were: seed length 6-8 nucleotides (nts), no G: U wobble allowed in seed of size 6 nts, one G:U wobble allowed in seed of size 7-8 nts, no mismatches allowed in seed of size 6 and 7 nts, one mismatch allowed in seed of size 8 nts, and no loop is allowed in microRNA or target for any seed size. Miranda parameters chosen for target prediction were a score threshold of 150, gap-open penalty -4.0 and gap-extend penalty 9.0. For TargetSpy, the score threshold was

chosen minimum of 0.99. For all tools, minimum energy threshold was chosen as 15 Kcal/mole.

### **Gene enrichment analysis of microRNA targets**

Gene enrichment analysis (GEA) was performed by using DAVID (Huang, Sherman and Lempicki 2009b, Huang, Sherman and Lempicki 2009a) (FDR- $p < 0.05$ ) and overrepresented pathways were visualized by EnrichmentMap (Merico et al. 2010) in cytoscape (Lopes et al. 2010) ( $P < 0.005$  and FDR- $q < 0.05$ ). Overlap between gene sets was computed according to an overlap coefficient; analysis was set to the default recommended value of 0.5.

### **MicroRNA-target gene co-expression and identification of cis regulatory promoter motifs**

Small RNA and mRNA sequencing reads from 22 families were used to estimate the correlation of microRNA-target genes. For target genes, mRNA sequencing reads were mapped to the trout mRNA reference and transcript per million (TPM) was calculated for each mRNA. For microRNA, small RNA sequencing reads were mapped to mature microRNA sequence from miRBase release 21 (June 2014) and total count was calculated for each microRNA. TPM (for mRNA) and total count (for microRNA) were normalized and used to estimate a gene expression correlation.

Transcription factors (TFs) binding motif were searched in the 500 upstream promoter sequences of DE microRNAs and correlated target genes using Alggen Promo TF motif search tool (Messeguer et al. 2002, Farré et al. 2003). Search parameters used were ‘only teleost transcription factors’ and ‘only teleost motif sites.’ The maximum

dissimilarity rate between putative and consensus TF binding site was set 5%, and RE equality/query (expectation of finding motif in random sequence) was set  $< 0.05$ .

### **Prediction of single nucleotide polymorphism (SNP)**

The same RNA samples were used to sequence small RNAs were used to sequence mRNA. Briefly, sequencing libraries were prepared using Illumina's Truseq RNA library preparation kit and sequenced on Illumina's HiSeq platform (Illumina Inc, CA, USA). Two different softwares, GATK (McKenna et al. 2010) and SAMTool (Li et al. 2009), were used to identify SNPs. Using the SAMtool approach, STAR alignment (Dobin et al. 2013) was used to align the sequencing reads from each family to the trout genome. SAMtools was used to determine variant genotypes. Popoolation2 package (version 1.201) was used to calculate allele frequencies (Li et al. 2009, Kofler, Pandey and Schlötterer 2011). SNPs were considered as putative trait-associated if the allele frequency (allele A/allele B) ratio between high and low ranked families was  $\geq 2.0$  or  $\leq 0.5$ . For the GATK approach, reads were aligned to the reference genome using STAR alignment. Finally, the base quality score recalibrations were performed on the data result from Picard tools. Haplotype Caller was used to determine the genotypic variants, followed by filtration of SNPs using strict thresholds; QD (Qual by depth) 2.0, Fisher Strand (FS) 60.0, RMS MQ (mapping quality) 40.0 and MAF  $> 0.05$ .

### **SNP genotyping and SNP-phenotype association analysis**

SNP genotyping was performed as a part of development of 50K SNP chip for Rainbow trout (full SNP chip study will be published elsewhere). Putative SNPs were genotyped in 1,920 individuals from USDA, NCCCWA growth selected lines harvest year 2010, 2012 and 2013 using Affymatrix SNP array protocol (Geneseek Inc., Lincoln, NE,

USA). Briefly, genomic DNA from each fish was extracted from fin clips and was amplified by PCR. DNA samples were chemically fragmented and biotinylated. Biotinylated DNA samples were hybridized to DNA probes in the SNP arrays and genotypes were determined based on ‘probe DNA: sample DNA’ hybridization.

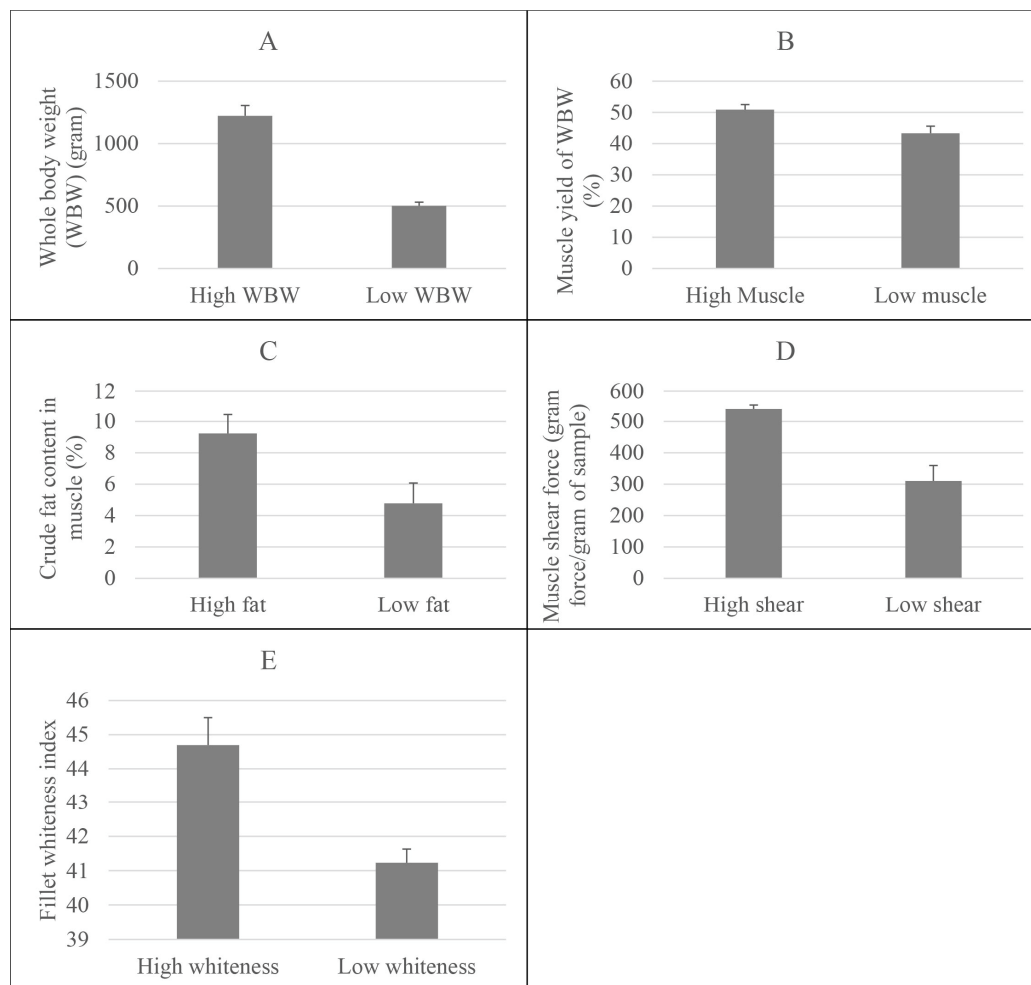
For the SNP-phenotype association study, only 249 SNPs present in 3' UTR microRNA binding sites and only 786 individual fish with available phenotype measurements of interest were considered. SNP-phenotype association were performed by linear model, logistic and quantitative trait analysis methods using the PLINK tool (Purcell et al. 2007).

## **RESULTS AND DISCUSSION**

### **Muscle trait phenotypes and Small RNA sequencing**

For ‘microRNA expression-phenotype’ association analyses, phenotypic data were collected from 500 fish belonging to 100 families (5 fish/family) of a growth-selected line at the USDA/NCCCWA Rainbow trout breeding program (harvest year 2010) (Salem et al. 2012, Leeds et al. 2016). Association of microRNA expression were analyzed for five important growth and muscle quality traits: WBW, muscle-yield of wbw, muscle crude-fat content, shear force and FWI (fillet whiteness index). The four highest-ranked and four lowest-ranked families for each phenotype were used for RNAseq. At 13 months after post-hatch, phenotypes were statistically different between these high ranked and low ranked families ( $P < 0.01$ ): whole body measurement ( $1221.6\text{g} \pm 84.25$  vs.  $502.1 \pm 28.0\text{g}$ ); muscle yield of whole body weight ( $50.9\% \pm 1.6$  vs.  $43.3\% \pm 2.3$ ); crude-fat ( $9.2\% \pm 1.2$  vs.  $4.8\% \pm 1.3$ ); muscle shear force;  $539.6 \pm 12.3$  vs.  $310.01 \pm 49.2$ ); and FWI ( $44.7 \pm 0.8$  vs.  $41.2$

$\pm 0.4$ ) (Figure 1). These phenotypic differences were achieved after four generations of selection for growth parameters (or characteristics).



**Figure 1:** Phenotypic difference for WBW and 4 muscle quality traits (muscle yield, crude-fat content, shear force and FWI) of top 4 high ranked and 4 low ranked families (5 fish/family) of selectively-bred trout at ca. 13 months post-hatch. A: whole body weight (wbw), B: muscle yield, C: muscle crude fat content, D: muscle shear force and E: muscle whiteness index.



Fish from 22 families were sequenced which included four top-ranked and four bottom-ranked families for each of the five traits (Table 1). High throughput small RNA sequencing resulted in mean sequencing depth of 14.5 million reads per sample. After trimming of sequencing adaptors, the average length of reads was 22 nts, a typical length average for mature microRNAs. After filtration and adapter trimming, the average number of reads in each sample was 11.9 million. On average, about 0.2 million potential microRNA transcripts were detected from trimmed reads in each sample. Of these potential microRNA transcripts, about 17.5% had sequence homology with mature microRNAs in miRBase database. From these annotated microRNAs in each sample, different variants of the same microRNA were collectively counted as a single microRNA, which resulted into an average of 1,154 different unique microRNA per sample (Table 1).

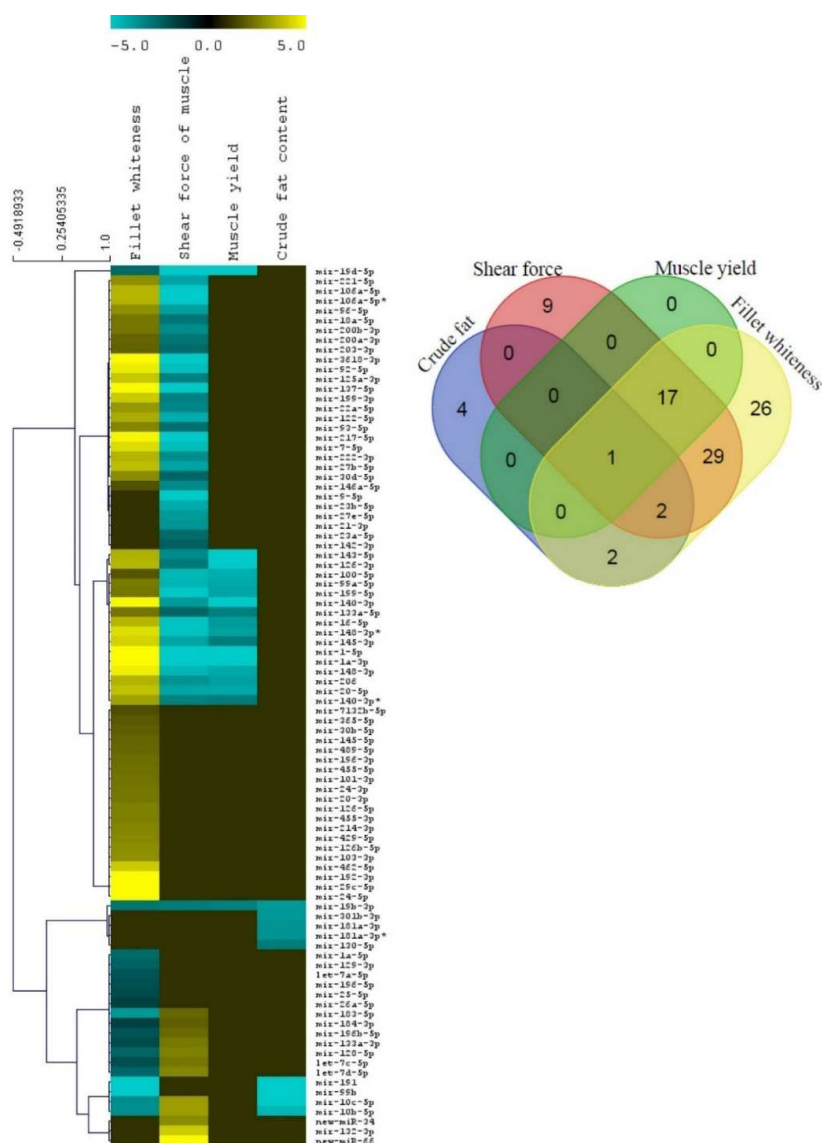
**Table 1:** Small RNA sequencing and annotation statistics of 22 samples used in the study.

Table shows name of fish families used for different traits, number of raw reads, number of reads after trimming/filtration, total number of putative microRNAs detected in each sample and number of putative microRNAs annotated to the miRBase microRNA reference. Last column represents the number of non-redundant microRNAs annotated to MiRBase microRNA reference in each sample. Note: <sup>a</sup>High WBW, <sup>b</sup>Low WBW, <sup>c</sup>High muscle yield, <sup>d</sup>Low muscle yield, <sup>e</sup>High crude fat, <sup>f</sup>Low crude fat, <sup>g</sup>High shear, <sup>h</sup>Low shear, <sup>i</sup>High fillet whiteness, <sup>j</sup>Low fillet whiteness and families used for more than one muscle traits are indicated with corresponding multiple superscripts.

Fish family	Read count and annotation statistics		Putative microRNA count and annotation statistics		
	Number of raw reads	Number of reads after trimming and filtration	Total putative microRNAs detected	Putative microRNAs annotated to mirbase Count (Percentage)	Number of non-redundant annotated microRNAs
195 <sup>i</sup>	11,168,118	9,881,880	150,902	35,355 (23.4%)	1,221
262 <sup>chi</sup>	14,841,753	12,357,559	318,231	42,577 (13.4%)	1,335
277 <sup>e</sup>	11,012,983	8,522,071	160,763	31,794 (19.8%)	1,188
366 <sup>fb</sup>	10,352,049	7,551,995	149,061	31,321 (21.0%)	1,212
390 <sup>cia</sup>	13,682,126	12,138,640	219,575	42,283 (19.3%)	1,277
399 <sup>hb</sup>	10,313,768	7,804,634	168,925	31,952 (18.9%)	1,194
405 <sup>i</sup>	10,309,353	7,979,759	145,673	31,256 (21.5%)	1,173
556 <sup>dhib</sup>	9,954,340	8,203,301	146,394	32,682 (22.3%)	1,205
565 <sup>d</sup>	10,384,499	7,709,062	189,110	30,388 (16.1%)	1,236
580 <sup>e</sup>	7,490,428	4,261,515	147,037	24,980 (17.0%)	1,115
595 <sup>g</sup>	12,040,061	8,372,165	319,839	35,112 (11.0%)	1,209
191 <sup>ij</sup>	21,246,629	18,766,660	247,811	30,175 (12.2%)	1,092
193 <sup>gia</sup>	25,361,890	19,439,379	141,126	35,160 (24.9%)	1,121
201 <sup>cja</sup>	12,676,918	11,296,840	138,647	21,561 (15.6%)	989
357 <sup>c</sup>	19,472,380	14,507,911	251,792	29,545 (11.7%)	1,054
408 <sup>g</sup>	13,845,919	12,152,718	94,556	21,053 (22.3%)	948
51 <sup>hi</sup>	20,964,015	17,185,620	241,354	37,891 (15.7%)	1,246
559 <sup>dib</sup>	13,714,117	11,099,365	136,184	18,551 (13.6%)	966
593 <sup>d</sup>	23,474,784	19,593,896	573,692	42,111 (7.3%)	1,261
597 <sup>c</sup>	22,048,522	20,079,565	255,680	33,440 (13.1%)	1,160
65 <sup>c</sup>	15,949,419	13,873,696	103,005	22,097 (21.5%)	998
194 <sup>a</sup>	9,115,039	7,974,575	138,450	31,695 (22.9%)	1,189
<b>Average</b>	<b>14,519,050</b>	<b>11,852,400</b>	<b>201,719</b>	<b>31,499 (17.5%)</b>	<b>1,154</b>

### **Differentially expressed microRNAs between high and low ranked families for growth and muscle traits**

In order to identify microRNAs associated with growth and muscle quality traits, we profiled DE microRNAs between two groups of fish families showing differences in for each of the five phenotypes: WBW, muscle yield (%) of WBW, crude-fat content, shear force and FWI. None of the microRNAs was DE in association with WBW. However, 90 different microRNAs were DE in the remaining four muscle traits (Figure 2). Two of the DE microRNAs were novel microRNAs recently reported recently in Rainbow trout new-miR-66 and new-miR-34 (Juanchich et al. 2016). A total of 18, 9, 56 and 77 microRNAs were DE between high and low ranked families for muscle % of WBW, crude fat (%) content, shear force and FWI, respectively. The majority of the DE microRNAs were shared among several phenotypic traits (Figure 2). For example, mir-19b was DE in all four muscle traits. In addition, all 18 DE microRNAs in muscle yield families were also DE in shear force and fillet whiteness families. Similarly, out of 56 DE microRNAs in shear force families, 49 were also DE in association with fillet whiteness. This observation suggests common mechanisms including coordinated expression of microRNAs in determining the studied muscle quality traits and is supported by the inter-related nature of these muscle traits in fish (Mørkøre et al. 2001).



**Figure 2:** Heat map of fold change of differentially expressed (DE) microRNAs between high vs low ranked families of various traits (left) and Venn diagram showing shared DE microRNAs between different traits (right). In the heat map, dark green and yellow colors indicate downregulation and upregulation respectively in high ranked families. The dark color indicates no differential expression of microRNAs. Note that the value of color limit (-5 to 5) does not reflect true fold change as values of fold change were transformed ( $\log_{10} 2$ ) and color scale was adjusted to make the heat map more visible.

Interestingly, the direction of change of the shared DE microRNAs were consistently either positively or negatively correlated between traits. For example, all the 18 DE microRNA that were downregulated in association with increased muscle yield were also downregulated in families showing high shear force ( $R = 0.73$ ) (Figure 2). In contrast, most of these downregulated microRNAs in families with high muscle yield (16 out of 18) were upregulated in the high-whiteness families ( $R = -0.74$ ). In addition, 47 out of 49 of the shared DE microRNAs between the shear force and fillet whiteness groups showed opposite pattern of differential expression ( $R = -0.93$ ). Similarly, all 5 shared DE microRNAs between crude fat and whiteness groups were downregulated in high ranked families of both traits ( $R = 0.93$ ). Out of 3 DE shared microRNAs between shear and fat group, 2 microRNAs showed opposite fold change pattern between the traits. In accordance with this observation, growth and muscle quality phenotypes of 500 fish population used in this study showed correlation between traits (Appendix A). WBW showed positive correlation with muscle yield ( $R = 0.56$ ,  $P < 0.0001$ ) and crude fat content ( $R = 0.57$ ,  $P < 0.0001$ ). Similarly, muscle yield showed very weak but significant positive correlation with crude fat ( $R = 0.25$ ,  $P < 0.0001$ ) and shear force ( $R = 0.17$ ,  $P = 0.0003$ ), and negative correlation with whiteness ( $R = -0.15$ ,  $P = 0.0009$ ). Crude fat content and whiteness had weak but positive correlation. This finding suggests that correlation among phenotypic traits could be, at least partially, explained by variation in expression level of DE microRNAs. In consistent with this observation, a recent report in salmon has indicated that crude fat content is negatively correlated with shear force of muscle and positively correlated with  $L^*$  (lightness) and  $b^*$  (yellowness) of raw salmon fillet (Mørkøre et al. 2001).

### **MicroRNAs associated with growth and muscle quality traits**

To estimate the contribution of microRNAs to the variation in phenotypes, 12 highly and commonly (among traits) DE microRNAs were selected for ‘microRNA expression-phenotype’ analyses using multiple regression. Expression level of microRNA was qPCR-quantified and correlated to phenotypes in 90 randomly selected fish from the same population. Linear regression analyses showed that of the 12 analyzed DE microRNAs, 10 correlated with WBW, all 12 correlated with muscle yield, 10 correlated with crude fat content, 5 correlated with shear force and 6 correlated with fillet whiteness index (cut off:  $R > 0.22$  or  $< -0.22$ ,  $p\text{-value} < 0.05$ ) (Table 2). Correlation (R) between expression of 12 microRNAs was 0.31, 0.42, 0.22, 0.13 and 0.26 with WBW, muscle % of WBW, crude fat (%) content, shear force and FWI, respectively.

We also analyzed the association of microRNA expression with other commercially important aquaculture traits and observed significant correlation. As an example, 11 microRNAs correlated with the percentage of trim losses to WBW, 11 correlated with fillet moisture, 3 correlated with belly flap thickness, 2 correlated with caudal body fat thickness, 2 correlated with fillet thickness and 1 correlated with percentage cook yield from fillet (Table 2). These traits were measured in the same fish of this study, but were not analyzed by RNA-Seq.

Expression of myogenic microRNA, mir-1a-3p, significantly correlated with WBW, muscle yield, and crude fat content with correlation (R) of 0.41, 0.43 and 0.27 respectively (Table 2). Similarly, other 2 myogenic microRNAs (mir-206 and mir-133a-5p) as well as mir-126-3p, mir-19b-3p, mir-148-3p, mir-20-5p, mir-143-3p, mir-99b, mir-10c-5p and mir-181a significantly correlated with muscle growth and other traits (Table

2). Significant correlation between microRNA expression and meat qualities including fat deposit, meat color, protein content and drip loss has been observed in other domesticated animal (Ponsuksili et al. 2013). These findings suggest that DE microRNAs identified in this study may be used in developing genetic markers to improve muscle growth and traits in Rainbow trout. Functional studies including microRNA gene editing may be used to validate role of these microRNA in determining variations in muscle growth and quality traits.

**Table 2:** Correlation between microRNA expression level and phenotypic variation. Correlation was calculated from phenotypic measurements and microRNA quantification (by real time PCR) in random 90 individual fish from the same NCCCWA's growth selected trout population used for small RNA sequencing. Note that negative value (-) of correlation coefficient ( $R$ ) indicates a negative correlation. Correlation was considered significant at  $P < 0.05$ .

DE microRNAs	Correlation with whole body wight	Correlation with muscle yield	Correlation with crude fat content	Correlation with shear force	Correlation with fillet whiteness
mir-1a-3p	-0.406	-0.434	-0.274	NA	NA
mir-126-3p	NA	-0.261	NA	-0.255	0.276
mir-19b-3p	-0.268	-0.415	-0.300	NA	NA
mir-148-3p*	-0.366	-0.387	-0.265	-0.300	0.295
mir-20-5p	-0.321	-0.428	-0.285	-0.270	0.297
mir-206	-0.290	-0.374	NA	-0.268	0.290
mir-133a-5p	-0.332	-0.397	-0.371	-0.265	0.255
mir-143-5p	-0.270	-0.361	-0.270	NA	0.247
mir-99b	-0.237	-0.332	-0.349	NA	NA
mir-10c-5p	-0.316	-0.286	-0.251	NA	NA
mir-10b-5p	NA	-0.259	-0.285	NA	NA
mir-181a-3p	-0.279	-0.316	-0.257	NA	NA

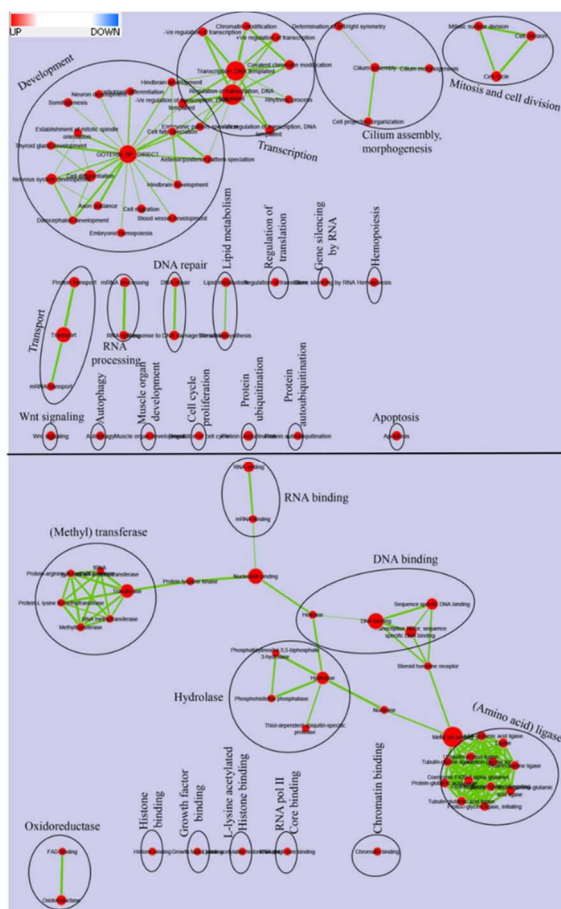
### **Targets of DE microRNAs and their functional annotation**

A total of 6837 different protein-coding genes were identified as potential targets with high confidence for the 90 DE microRNAs. In order to investigate the functional significance of the predicted target genes, we performed gene enrichment analysis (GEA) using DAVID (Huang et al. 2009b). Genes involved in multicellular organism development (4.2× fold), transcription (3.7× fold) and mitosis/cell division (4.0× fold) were highly overrepresented among predicted targets of microRNAs (Figure 3). Other pathways significantly enriched in the target genes included RNA processing, DNA repair, gene silencing by RNA, protein ubiquitination, lipid metabolism, muscle organ development, regulation of growth, cell proliferation and apoptosis. Similarly, among the signal transduction pathways, the Wnt signaling pathway was overrepresented within the target gene list, which is involved in muscle growth (Steelman et al. 2006), myogenesis (Cossu and Borello 1999, Brack et al. 2008) and regulation of growth control genes (Armstrong and Esser 2005).

In the molecular function pathways category, the majority of the overrepresented gene sets had functions related to epigenetic gene regulation such as methyltransferase activity, histone binding and chromatin binding activity (Figure 3). Methyltransferases are important in epigenetic gene-regulation and are known to regulate skeletal muscle growth by regulating expression of myoD transcription factor (Ling et al. 2012). Another enriched gene-set included various ligases that modify proteins through post translation by adding specific amino acids (e.g. protein-glycine ligases, protein-glutamic acid ligase and tubulin-glycine ligase). These results suggest that DE microRNAs may control growth and muscle



quality phenotypes via post-transcriptional regulation of genes involved in development, muscle growth, epigenetic gene regulation and protein modification.



**Figure 3:** Enrichment map and enriched gene pathways of predicted microRNA targets classified into the biological process (top) and molecular function (bottom) categories. In the enrichment map, enriched gene-sets represent nodes, which are related/ connected by their GO relation(s) (edges). Red node color represents enriched gene-set. Color intensity of the node represents significance of enrichment; node size represents number of genes in the gene-set (proportional relation) and edge thickness represents the degree of overlap between gene-sets (proportional relation).

## **MicroRNAs associated with each muscle phenotype and their relevant protein-coding gene targets**

### ***Whole body weight (WBW)***

No microRNA was DE between the high- and low-ranked families of WBW phenotype. However, microRNA expression-phenotype regression analysis performed on 90 randomly selected individual fish showed that expression of 10 DE microRNAs significantly correlated with WBW. These microRNAs included muscle specific myogenic microRNAs that were DE in response to muscle yield and/or muscle crude fat content. Future mechanistic studies involving gene knockout or dose-dependent response of individual microRNA on WBW may determine suitable genetic markers for selection.

### ***Muscle yield***

There was a total of 18 microRNAs downregulated in the high muscle yield families compared to their low muscle yield counterparts (Figure 2). Of them, the expression of 12 microRNAs was validated by real time PCR and showed consistency between qPCR and RNA-seq approaches (Appendix B). This includes mir-206, mir-1 and mir-133 (all 3 muscle specific microRNAs) (Wang 2013). In addition, all 9 DE microRNAs, that were chosen for phenotype-genotype regression analysis, negatively correlated with muscle yield ( $P < 0.05$ ). Correlation (R) between microRNA expression and variation in muscle yield ranged from -0.25 to -0.43, with the most significant microRNA being mir-1a-3p (Table 2). Although expressions of mir-10c-5p and mir-181a-3p were not statistically significant in the RNASeq between high and low ranked muscle yield groups, 'phenotype-microRNA expression' regression analysis on 90 fish, showed that their expression was significantly and negatively correlated with muscle yield (Table 2). To the best of our

knowledge, 8 out of 18 DE microRNAs, were not reported before as associated with muscle growth. These microRNAs are mir-19d, mir-100-5p, mir-99a-5p, mir-148-3p, mir-199-5p, mir-148-3p\*, mir-19b-3p and mir-140. The remaining 10 DE microRNAs are known to directly or indirectly regulate skeletal muscle development in different species by negatively regulating genes involved in myogenesis (Chen et al. 2014, Luo et al. 2015), muscle-growth-related signal transduction pathways (Wang 2013, Yuan et al. 2013) and cell cycle (Liu et al. 2008). As an example, two DE microRNA, mir-1 and mir-133, are responsible for more than half of the microRNA-mediated gene regulation in muscle of Zebra fish (Mishima et al. 2009). Similarly, mir-143 regulates expression of myoD, a major myogenic transcription factor, in skeletal muscle of fish (Chen et al. 2014). Mir-140 regulates expression of myomaker required for myoblast fusion in chicken (Luo et al. 2015). These findings suggest that differential gene expression has identified majority of microRNAs previously known to regulate myogenesis as well as several additional microRNAs potentially implicated in muscle growth.

A total of 1743 protein-coding genes were predicted as potential targets of the 18 downregulated microRNAs in high muscle yield group. Interestingly, among the target genes, 10 significantly enriched biological pathways were directly involved in muscle growth and myogenesis (FDR-p < 0.05) (Appendix C). Some enriched muscle growth related pathways included muscle organ development (4.4× fold), skeletal muscle tissue development (4.5× fold) and muscle cell differentiation (3.1× fold). Other biological pathways including chromatin modification (5.4× fold), transcription (3.2× fold), cell cycle (4.7× fold), and multicellular organism development (4.6× fold). These findings suggest

that DE microRNAs may regulate muscle yield by regulating the genes directly involved in skeletal muscle development, transcription, cell cycle, and/or protein degradation.

### ***Crude fat content in muscle***

A total of 9 microRNAs were significantly downregulated in the high-fat fish families compared to low-fat families (Figure 2). Seven out of 9 DE microRNAs are well documented as associated with adipogenesis and/or body fat deposition in different species (Jin et al. 2010, Qin et al. 2010). The remaining 2 DE microRNAs (mir-10c-5p and mir-19b-3p) were not reported before to the best of our knowledge. A total of 494 protein coding genes were predicted as potential target of the 9 DE microRNAs. Functional annotation of these target genes, exhibited that four biological pathways were significantly overrepresented; these are the multicellular organism development (4.7× fold), cell differentiation (4.1× fold), transcription (3.3× fold) and regulation of transcription (3.1× fold) (FDR-p < 0.05) (Appendix D). Mir-10c-5p and mir-19b-3p, which were not reported previously to be associated with body fat, targeted several genes involved in adipocyte differentiation (e. g. suppressor of cytokine signaling 6 and fragile site-associated protein isoform x4), fat storage (e. g. perilipin-2 isoform x1), lipid metabolism (e. g. peroxisomal fatty acyl-coA oxidase-1, lysophosphatidylcholine acyltransferase 2) and lipid transport (e. g. apolipoprotein b-100, microsomal triglyceride transfer protein and star-related lipid transfer protein-4).

Three of the nine microRNAs (mir-10c-5p, mir-10b-5p and mir-181a-3p), were qPCR analyzed for association with variation in muscle fat content in 90 individual fish. The regression analysis showed that these microRNAs significantly correlated with muscle crude fat content (Table 2). In addition to these 3 DE microRNAs, 7 non-DE microRNAs

that were associated with muscle yield were also significantly correlated with the muscle crude fat content (FDR-p <0.05), (Table 2). Those microRNAs are mir-1a-3p, mir-19b-3p, mir-148-3p\*, mir-20-5p, mir-133a-5p, mir-143-5p and mir-99b. These findings suggest that DE microRNAs may play crucial role in post-transcriptional regulation of genes associated with crude fat content in muscle.

### ***Muscle shear force***

A total of 56 microRNAs were DE between the high and low shear force groups, of them 46 microRNA were downregulated and 10 microRNAs were upregulated in fish with high shear force (Figure 2 and Supplementary dataset 1B). Correlation between expression level and variation in shear force of muscle was studied for 10 DE microRNAs in 90 fish, of them 5 microRNAs (mir-126-3p, mir-148-3p\*, mir-20-5p, mir-206 and mir-133a-5p) were significantly correlated with muscle shear force (FDR-p < 0.05) (Table 2). To the best of my knowledge, the association of microRNAs with muscle firmness was not reported in salmonids before. The connective tissue proteins especially collagen play important role in determining muscle shear force. In order to get insights into functional significance of DE microRNAs in regulation of muscle firmness, we annotated genes potentially targeted by DE microRNAs. Interestingly, out of 56 DE microRNA, 31 microRNAs targeted 42 different genes coding for collagen proteins or regulators of collagen biosynthesis, metabolism or structure (Appendix E). In addition, 31 DE microRNAs targeted at least 61 genes coding for extracellular matrix proteins and their regulators other than collagens and collagen regulators (data not shown). It was observed that the abundance of collagen and connective tissue in extracellular matrix determines stiffness and gaping in fish fillet (Johnston 1999). These DE microRNAs, which target

collagens and other extracellular matrix connective tissues and their regulators, may be used to develop suitable genetic markers to improve muscle firmness in Rainbow trout.

### ***Fillet whiteness index***

Seventy-seven microRNAs were DE, 58 upregulated and 19 downregulated, in fish families with high fillet whiteness index compared to their counterpart families with low whiteness index (Figure 2). From DE microRNA, correlation between microRNA expression level and variation in phenotype was performed for 11 microRNAs, of them, variation in expression level of 6 microRNAs significantly correlated with variation in fillet whiteness index (Table 2). In order to investigate biological/molecular process potentially involved in fillet whiteness trait, we performed gene enrichment analysis of the genes targeted by DE microRNAs. In biological function category, 30 biological pathways were significantly overrepresented which included transcription, transcription regulation, multicellular organism development, protein ubiquitination and Wnt signaling (Appendix F). We observed that most of the DE microRNA in response to variation in muscle yield, crude fat content and shear force were also DE in response to variation in fillet whiteness. These data suggest that fillet whiteness may be, at least partially, impacted by the mechanism that regulate the other three muscle quality traits. Fillet color parameters were correlated to muscle fat content in Atlantic salmon (Einen 1998). Perhaps this is among the first genome-wide studies aiming at exploring the genetic/molecular basis of fillet whiteness in salmonids.

### **MicroRNA-target gene co-expression and transcriptional regulation**

MicroRNAs and their target genes are usually co-expressed and co-regulated by common transcription factors (Wang, Li and Hu 2011). In order to investigate co-

expression of DE microRNAs and their target genes, I calculated Pearson correlation between each DE microRNA and their target genes based on their expression values in 30 RNA-Seq samples (see methods section). Several DE microRNAs showed strong positive expression correlation with their target genes suggesting their coordinated expression. Next, we investigated whether this co-expression was regulated via common transcription factors (TF). For this purpose, we scanned promoter sequences of the strongly correlated ( $R > 0.84$ ) DE microRNAs and their target genes for TF binding cis regulatory motifs. Out of 90 DE microRNAs, 15 microRNAs had strong positive expression correlation ( $R > 0.84$ ) with 194 different target genes, and all of these correlated microRNA-target gene pairs shared common TF binding motifs in their promoters. Selected microRNA-target gene pairs and common TF binding motifs in their promoters are given in Table 3. Some of those TFs are known to be heavily involved in muscle development (e. g. myoD, c-Fos, c-Jun, MAZ, NF-AT1, Smad3, Elk1, PEA3, NFI/CTF and NFY), development (e. g. HOXD9, HOXD10 and COE1), metabolism (e. g. HNF-3) and adipogenesis (e. g. C/EBPbeta). These findings suggest that myogenic TFs regulate expression of muscle important microRNAs and their target genes. Wang and coworkers suggested that both miRNAs and TFs must stay active to simultaneously regulate their target genes (Wang et al. 2011).

**Table 3:** Cis-regulatory transcription factor binding motifs that exist in promoter sequences of differentially expressed (DE) microRNAs and their positively correlated target genes.

MicroRN A	Target gene	MicroRNA-target gene expression correlation (R)	Common cis regulatory promoter motifs in microRNA and target gene
mir-92-5p	GSONMT00020058001: homeobox protein orthopedia b-like	0.95	HNF-3beta, AML1, TFIID, HNF-3
mir-26a-5p	GSONMT00040924001: kelch repeat and btb domain-containing protein 11-like	0.92	AR, C/EBPbeta, c-Jun, Fli-1, IRF-2, NF-AT1, NF-AT2, NF-AT3, NF-muNR, POU1F1, PR B
let-7a-5p	GSONMT00042722001: inhibitor of growth protein 5-like	0.92	AP-3, AR, C/EBPbeta, CBF(2), CP2, GATA-1, HNF-3alpha, HNF-3beta, HOXA4, NF-Y, PR B, Smad3, SXR:RXR-alpha, TBP
mir-222-3p	GSONMT00046577001: synembryn-b-like isoform x2	0.91	C/EBPgamma, c-Jun, NF-Y, RAR-gamma, Smad3, WT1
mir-10c-5p	GSONMT00022788001: transcription elongation factor SPT6-like	0.90	HNF-3alpha, Elk-1, HNF-3beta
let-7c-5p	GSONMT00013587001: collectrin-like isoform	0.89	AR, Crx, HNF-3alpha, HNF-3beta, HOXD10, HOXD9, Pax-2.1, PR B, TFIID
mir-130-5p	GSONMT00036000001: uncharacterized protein	0.89	CP2, HNF-3alpha, HNF-3beta, PR B, RAR-gamma
mir-27b-5p	GSONMT00073742001: serine threonine-protein phosphatase 6 regulatory ankyrin repeat subunit c	0.88	AhR, AR, c-Myb, NF1/CTF, POU1F1b, POU1F1c, PR A, PR B
mir-29c-5p	GSONMT00037095001: protein NLRC3-like	0.87	AP-3, ENKTF-1, HNF-3beta, NF-Y, PPAR-alpha:RXR-alpha, PR B
mir-221-5p	GSONMT00080720001: lathosterol oxidase	0.87	AR, C/EBPbeta, c-Jun, HNF-3beta, p53, PEA3, PR B
mir-132-3p	GSONMT00033564001: dihydropyrimidinase	0.85	AP-3, CBF(2), CP2, Elk-1, ER-alpha, ER-beta, MyoD, NF-1, NF-Y, PU.1, SF-1, T3R-beta
mir-30d-5p	GSONMT00002523001: solute carrier family 12 member 9-like	0.85	AR, PR B, CP2
mir-462-5p	GSONMT00075887001: ATP-dependent 6-phosphofructokinase, muscle type-like	0.84	HNF-3alpha, MyoD, NF-AT3, NF-AT2, PR B
mir-133a-3p	GSONMT00002111001: TPA_inf tachykinin 4	0.83	c-Fos, c-Jun, COE1, GATA-1, HNF-3, HNF-3beta, IRF-1, MAZ, NF-AT1, Pbx1b, PEA3, PR B

### Genetic polymorphism in microRNA target sites

In order to explore the phenotype-associated gene polymorphic on microRNA and microRNA binding sites in target genes, predicted target genes of all Rainbow trout microRNAs that were reported previously (Juanchich et al. 2016). SNPs were identified from the mature microRNA sequence and mRNA sequencing of the same fish families used in this study (data under review for publication elsewhere). A total of 249 SNPs existing in microRNA recognition element seed site (MRESS) of target genes showed allelic imbalance ( $> 2.0$  and  $< 0.5$ ) between high and low ranked families for all five traits; WBW, muscle % of WBW, crude-fat content, shear force or FWI. Out of 249 SNPs, 240



SNPs either destroyed or created a novel illegitimate microRNA target, and only 9 SNPs had no potential effect on microRNA binding. This alteration in target recognition in presence of SNP is caused by existence of the SNPs in MRESS of the target genes. MRESS plays crucial role in determining microRNA-mRNA binding (Lewis et al. 2005), and SNP in MRESS has important impact on microRNA recognition (Clap et al. 2006). Out of 240 SNPs capable of destroying or creating a novel illegitimate MRESS, 204 SNPs were found to be true polymorphic SNPs by genotyping fish with allelic imbalance between high and low ranked families  $> 2.0$  and  $< 0.5$ . Interestingly, 125 SNPs showed significant association with growth and muscle quality phenotypes based on genotype-phenotype association analysis performed on a large number of fish population ( $n = 786$ ) (FDR-p-value  $< 0.05$ ) (Table 4) A total of 56, 21, 6 and 12 MRESS-destroying SNPs in 3' UTR were significantly associated with WBW, muscle yield, crude fat content and fillet whiteness respectively (FDR-p-value  $< 0.05$ ). In addition to growth and muscle quality phenotypes, these SNPs also explained significant variation in other important fish traits. As an example, a total of 45, 53, 43, 40, 5, 5, 1 and 24 MRESS-destroying SNPs in 3' UTR were significantly associated with fillet trim loss, mid-belly flap thickness, fish viscera weight, fillet thickness, fillet energy content, cook yield, fillet crude protein content and head weight, respectively (FDR-p-value  $< 0.05$ ).

Previous studies indicated that SNPs in MRESS of target genes have important impact in phenotype determination (Clap et al. 2006). Phenotype associated MRESS-destroying SNPs were present in various classes of genes including metabolic enzymes, transporter proteins, transcription factors, signaling molecules and muscle related proteins. Among the WBW associated SNPs are SNPs present in 3' UTR of troponin, malate

dehydrogenase, ranbp-type and c3hc4-type zinc finger-containing protein 1, phosphate carrier mitochondrial like protein, cytochrome b-c1 complex, novel protein vertebrate nebulin, ATP-dependent 6-phosphofructokinase, ankyrin repeat and soxs box protein 5 and calsequestrin-1 each SNP explained more than 2% variation in WBW phenotype (Table 4). Similarly, SNPs existing in 3' UTR of beta-sarcoglycan and gamma-adducin-like isoform x6 each explained over 2% variation in muscle yield. The presence of MRESS-destroying allele may stabilize the target genes against microRNA-mediated downregulation that may contribute to the difference in phenotype between high vs low ranked families. As an example, allele destroying MRESS in troponin fast skeletal muscle was 11× more frequent in high WBW family compared to low WBW family, and was significantly correlated with WBW (Table 4). A previous study performed in a trout population has also identified 3 SNPs present in 3' UTR of troponin C associated with growth traits (Salem et al. 2012). Similarly, allele destroying MRESS in ATP-dependent 6-phosphofructokinase, a glycolytic gene, was 5.5× more frequent in high WBW family compared to low WBW family, and significantly correlated with WBW. Consistent with our finding, trout growth traits associated SNPs from previous study were mainly present in genes involved in energy metabolism and muscle structure (Salem et al. 2012). Above findings suggest that 3' UTR SNPs capable of destroying or creating an illegitimate MRESS may have important functional consequences.

In contrast to target genes, no SNPs with allelic imbalances were detected in the microRNA mature sequences. This may be due to high degree of negative selective pressure as hundreds of genes are regulated by the same microRNA. These findings suggest

that though some microRNAs do not show variation in expression level in response to muscle trait phenotype, genetic variation are positioned in their target genes.

**Table 4:** SNPs in microRNA recognition element seed site (MRESS) of target gene, allele frequency ratio of MRESS-destroying SNPs between high vs low ranked families of different muscle traits, and correlation between the SNP and phenotype. Column 1 shows potential regulatory microRNA and column 2 shows microRNA binding site in 3' UTR of target gene. Note that only MRESS with SNP, not full microRNA sequence is shown. Complete datasets are given in supplementary dataset 5. Note: WBW: whole body weight,

MicroRNA	SNP in MRESS	SNP NCBI S.r.	Allele destroying MRESS	Frequency ratio of MRESS destroying allele (High/Low family) (WBW; MY; CFC; S; F; FWI)	Target gene harboring SNP	SNP association with phenotype ( $R^2$ ) (WBW; MY; CFC; S; F; FWI)
pma-miR-7a-3p	TGTC[C/T]TGT	2711239550	T	11.00; 3.55; 2.33; 0.96; 1.21	troponin fast skeletal muscle isoform	0.041; NA; NA; NA; NA
cgr-miR-598	TCCTAC[T/G]A	2711263652	G	0.30; 0.44; 0.11; 0.25; 0.31	malate dehydrogenase cytoplasmic-like	0.036; NA; NA; NA; NA; 0.019
efu-miR-9203a	ACTAT[C/T]AA	2711277723	T	1.62; 1.04; 3.00; 0.83; 1.00	ranbp-type and c3hc4-type zinc finger-containing protein 1	0.030; NA; NA; NA; NA
oha-miR-30e-5p	AC[C/T]GGAAGG	2711281551	C	3.24; 5.00; 1.06; 0.27; 0.25	phosphate carrier mitochondrial precursor	0.030; NA; NA; NA; NA; 0.019
ssa-miR-139-5p	CA[C/A]GTAGA	2711237958	A	0.29; 1.57; 0.37; 0.025; 0.39	novel protein vertebrate nebulin	0.029; NA; NA; NA; NA
ssa-miR-139-5p	CACT[G/A]TAGA	2711228680	A	0.29; 1.57; 0.37; 0.025; 0.39	novel protein vertebrate nebulin	0.029; NA; NA; NA; NA
dvi-miR-968-5p	TATCAT[C/T]AG	2711283561	C	5.50; 1.00; 0.56; 30.67; 2.57	ATP-dependent 6-phosphofructokinase, muscle type	0.028; NA; NA; NA; NA
ssy-miR-508	GT[A/G]GCTGG	2711210896	A	2.56; 0.68; 9.20; 2.52; 0.58	ankyrin repeat and soes box protein 5	0.028; 0.016; NA; NA; NA
ppc-miR-8229a-5p	GCTGAGG[A/T]	2711244497	T	5.09; 5.48; 4.00; 0.18; 1.17	calsequestrin-1-like	0.025; NA; NA; NA; NA
oha-miR-26-5p	GGATA[C/A]GGT	2711271866	A	1.11; 1.71; 1.40; 0.16; 0.25	fk506-binding protein 1a	0.025; 0.015; NA; NA; NA
oha-miR-24-3p	AGCAG[G/A]AAA	2711198793	A	1.83; 0.35; 0.44; 1.90; 2.20	spectrin beta non-erythrocytic 1 isoform x1	0.023; NA; NA; NA; NA
mmu-miR-3547-3p	CCCC[T/C]CTT	2711211990	C	0.31; 3.21; 0.20; 1.00; 0.27	gamma-adducin-like isoform x6	0.023; 0.027; NA; NA; NA
oha-miR-365a-2-5p	CAGAA[G/A]GA	2711268945	A	2.16; 0.55; 8.44; 0.43; 2.50	nuclear receptor subfamily 1 group d member 2	0.022; NA; NA; NA; NA
ssa-miR-196b-5p	GT[A/T]GTTGTT	2711240153	A	0.40; 0.09; 1.50; 0.53; 0.64	histone-lysine n-methyltransferase setd7-like	0.021; NA; NA; NA; NA
ppc-miR-8298-3p	CATA[A/C]TTC	2711225442	A	0.67; 1.81; 0.50; 0.39; 0.51	atp synthase lipid-binding mitochondrial precursor	0.021; 0.021; NA; NA; NA
oha-miR-181a-5p	T[G/T]AATGTT	2711263991	T	3.75; 1.89; 1.21; 2.13; 0.67	phosducin-like protein 3	0.020; NA; NA; NA; NA
eca-miR-9171	CAG[A/G]CTGT	2711198125	G	2.10; 1.00; 0.56; 1.40; 6.50	nfu1 iron-sulfur cluster scaffold mitochondrial-like	0.020; NA; NA; NA; NA
mmu-miR-7241-3p	AGTATT[C/T]G	2711278279	T	1.75; 5.50; 0.42; 1.25; 0.60	3-hydroxyacyl-CoA dehydratase 1	0.020; 0.013; NA; NA; NA
cgr-miR-29a-5p	GTGTACG[G/T]C	2711192504	G	2.89; 1.00; 11.33; 3.33; 1.47	phosphorylase b kinase gamma catalytic subunit	0.019; 0.015; NA; NA; NA
efu-miR-9186e	TCT[C/A]TGGTA	2711207216	C	2.05; 3.08; 21.11; 0.37; 2.16	manganese superoxide dismutase	0.019; 0.015; NA; NA; NA
ccr-miR-729	TACCA[C/T]C	2711280997	C	5.47; 2.86; 2.75; 21.11; 0.40	profilin-2-like isoform x1	0.017; NA; NA; NA; NA
cfa-miR-8804	TCTA[T/C]CTA	2711247534	C	0.64; 0.32; 0.11; 0.43; 0.79	14-3-3 protein beta alpha-2	0.017; 0.017; NA; NA; NA
ssa-miR-210-5p	[T/G]TACATTA	2711275155	T	3.26; 0.49; 1.07; 0.90; 6.51	60s ribosomal protein 117	0.015; 0.022; NA; NA; NA
eca-miR-9011	[T/A]CCTGTACA	2711195787	T	0.29; 8.80; 0.90; 0.44; 1.14	monocarboxylate transporter 9	0.013; NA; NA; NA; NA
api-miR-3015a	TTGAAAAC[C/A]	2711235305	C	1.32; 2.67; 3.78; 0.58; 0.82	parvalbumin-7-like isoform x2	0.013; NA; NA; NA; NA; 0.012
ssa-miR-93a-5p	AGCA[T/C]TTTG	2711192683	T	0.99; 0.72; 7.82; 0.25; 5.09	protein nap homolog 2-like	0.013; NA; 0.019; NA; 0.025
mmu-miR-6975-3p	[G/A]CAGACGAC	2711260618	G	0.46; 0.08; 2.53; 0.98; 6.33	wolframin	0.011; 0.012; NA; NA; NA
hsa-miR-8086	ATGTAT[T/G]CC	2711280737	T	0.08; 0.17; 0.74; 0.66; 0.19	muscleblind-like protein 1-like	NA; 0.010; NA; NA; NA
osa-miR818f	ACAATCT[A/G]T	2711274522	G	0.21; 0.13; 0.40; 5.44; 1.68	nexilin isoform x1	NA; 0.011; NA; NA; NA
mmu-miR-6989	GCAACT[C/T]CAA	2711191913	T	0.21; 0.11; 0.74; 2.70; 1.00	arrestin domain-containing protein 2	NA; 0.012; NA; NA; NA
ggp-miR-4738	AGCAGC[G/A]T	2711275288	G	4.11; 12.50; 2.48; 0.18; 0.71	sarcosine mitochondrial-like	NA; 0.012; NA; NA; NA
prd-miR-7957d-3p	TC[T/A]GGACAT	2711274238	T	4.74; 15.63; 2.48; 0.18; 0.60	profilin-2-like isoform x2	NA; 0.013; NA; NA; NA
oha-miR-133b-5p	GTGCAC[G/T]T	2711249814	G	0.09; 0.18; 0.25; 38.00; 0.85	inactive dual specificity phosphatase 27	NA; 0.014; NA; NA; NA
ppc-miR-8304a-3p	[C/A]TTTGG	2711194746	A	0.17; 2.71; 0.14; 1.90; 0.50	atlastin-2 isoform 1	NA; NA; 0.017; NA; NA
ipu-miR-34b	TGTT[A/C]ACT	2711267634	C	1.64; 2.87; 3.71; 0.50; 1.41	creatine kinase m-type-like	NA; NA; 0.018; NA; NA

MY: muscle yield of WBW (%), CF: muscle crude fat content (%), SF: muscle shear force; and FWI: fillet whiteness index.

## **CONCLUSION**

Improvement of muscle growth and quality in salmonids has long been sought by aquaculture industries around the globe. So far, little progress has been made toward genetic improvement of muscle growth and quality in salmonids, probably due to the complex interrelated relationship among muscle quality traits (Johnston et al. 2000, Mørkøre et al. 2001), and influence of several genetic and non-genetic factors. However, previous studies performed in different aquaculture species indicated existence of good genetic variation for muscle growth and quality in trout and other salmonids (Gjedrem 1983, Blanc and Choubert 1993). In this study, we investigated microRNAs expression and genetic polymorphism in microRNA-binding sites in target genes that is associated with phenotypic variation in muscle growth and quality in Rainbow trout population produced after 4 generations of family-based phenotypic selection at NCCCWA Rainbow trout breeding station.

We quantified DE microRNAs between fish families showing contrasting phenotypes for 5 traits in Rainbow trout: WBW, muscle % of WBW, crude-fat, shear force and FWI. Muscle specific myogenic microRNAs (e. g. mir-1, mir-206 and mir-133) as well as several non-muscle specific microRNAs showed regulated expression in response to muscle yield and other phenotypes. Most of the DE microRNAs between high vs low muscle yield families were also DE between high vs low families of fillet whiteness and muscle shear force phenotype. This observation may be, in part, due to the interrelated nature of these muscle growth phenotypes (Johnston et al. 2000, Mørkøre et al. 2001).

Biological pathways such as development, cell cycle, muscle growth, transcription and muscle proteolysis were significantly overrepresented in the list of DE microRNA targets suggesting that DE microRNAs may regulate growth and muscle quality traits by post-transcriptionally regulating the genes involved in these pathways. Presence of common cis regulatory motifs for myogenic TFs in DE microRNAs and their respective target genes may suggest that myogenic TFs regulate expression of both myogenic microRNAs and their target genes.

Due to crucial role of MRESS in microRNA-mediated gene regulation, SNPs creating or disrupting MRESS in target gene have important functional consequences (Clop et al. 2006, Sethupathy and Collins 2008). In this study, we have identified 125 true SNPs in 3' UTR capable of abrogating or creating MRESS on several metabolic and growth-related genes, and the SNPs explained significant variation in growth, muscle yield and other muscle phenotypes. However, functional significance of these SNPs in microRNA target recognition needs to be experimentally validated. In order to make present study more robust in identifying potential genetic markers for muscle and growth phenotypes, we applied dual approach of analyzing differential microRNA expression as well as genetic variation analysis in their target genes. We believe that this approach is more productive as high negative selective pressure impose constraints on expression/genetic variation in microRNA. This argument is supported by our finding of several SNPs on microRNA binding sites in growth related genes and transcription factors though microRNAs did not show variation in expression level in response to variation in WBW. In this study, we performed genome wide approach to investigate variation in expression and variation in

target recognition of growth and muscle-important microRNAs, and the study may help identifying suitable genetic marker for genetic selection of these traits.

In another study performed in the same set of fish, an unexpectedly a low number of protein coding genes were observed as DE between high and low ranked families of these phenotypes (data will be published elsewhere); however, in this study I observed significant number of DE microRNAs and their target protein coding genes that are associated with these phenotypes. These findings suggest that variation in these muscle quality phenotypes may be explained largely by variation in microRNA expression and genetic variation affecting recognition of microRNA targets rather than by direct regulation of mRNA expression. Consequently, the current study pinpoints a greater role of microRNAs in post-transcriptional regulation of muscle important traits and genes in Rainbow trout. This study provides insights into the underlying biological basis of muscle quality-trait variation in Rainbow trout, and will help in design of future functional studies aimed at muscle quality improvement in trout.

## REFERENCES

- Armstrong, D. D. & K. A. Esser (2005) Wnt/beta-catenin signaling activates growth-control genes during overload-induced skeletal muscle hypertrophy. *Am J Physiol Cell Physiol*, 289, C853-9.
- Bagga, S., J. Bracht, S. Hunter, K. Massirer, J. Holtz, R. Eachus & A. E. Pasquinelli (2005) Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*, 122, 553-63.
- Berthelot, C., F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noel, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J. M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J. N. Volff, C. Genet, P. Wincker, O. Jaillon, H. Roest Crollius & Y. Guiguen (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*, 5, 3657.
- Blanc, J.-M. & G. Choubert (1993) Genetic Variation of Flesh Color in Pan-Size Rainbow Trout Fed Astaxanthin. *Journal of Applied Aquaculture*, 1993.
- Brack, A. S., I. M. Conboy, M. J. Conboy, J. Shen & T. A. Rando (2008) A temporal switch from notch to Wnt signaling in muscle stem cells is necessary for normal adult myogenesis. *Cell Stem Cell*, 2, 50-9.
- Chen, L., P. Wu, X. H. Guo, Y. Hu, Y. L. Li, J. Shi, K. Z. Wang, W. Y. Chu & J. S. Zhang (2014) miR-143: a novel regulator of MyoD expression in fast and slow muscles of *Siniperca chuatsi*. *Curr Mol Med*, 14, 370-5.
- Clop, A., F. Marcq, H. Takeda, D. Pirottin, X. Tordoier, B. Bibe, J. Bouix, F. Caiment, J. M. Elsen, F. Eychenne, C. Larzul, E. Laville, F. Meish, D. Milenkovic, J. Tobin, C. Charlier & M. Georges (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet*, 38, 813-818.
- Cossu, G. & U. Borello (1999) Wnt signaling and the activation of myogenesis in mammals. *EMBO J*, 18, 6867-72.
- Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson & T. R. Gingeras (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15-21.
- Einen, S. (1998) Quality characteristics in raw and smoked fillets of Atlantic salmon, *Salmo salar*, fed high-energy diets. *Aquaculture Nutrition*, 4, 99-108
- Farré, D., R. Roset, M. Huerta, J. E. Adsuara, L. Roselló, M. M. Albà & X. Messeguer (2003) Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. *Nucleic Acids Res*, 31, 3651-3.
- Flynt, A. S., N. Li, E. J. Thatcher, L. Solnica-Krezel & J. G. Patton (2007) Zebrafish miR-214 modulates Hedgehog signaling to specify muscle cell fate. *Nat Genet*, 39, 259-63.
- Georges, M., A. Clop, F. Marcq, H. Takeda, D. Pirottin, S. Hiard, X. Tordoier, F. Caiment, F. Meish, B. Bibé, J. Bouix, J. M. Elsen, F. Eychenne, E. Laville, C. Larzul, D. Milenkovic, J. Tobin & A. C. Charlier (2006) Polymorphic microRNA-target

- interactions: a novel source of phenotypic variation. *Cold Spring Harb Symp Quant Biol*, 71, 343-50.
- Gjedrem, T. (1983) Genetic variation in quantitative traits and selective breeding in fish and shellfish. *Aquaculture*, 33, 51-72.
- Hsu, R. J., C. Y. Lin, H. S. Hoi, S. K. Zheng, C. C. Lin & H. J. Tsai (2010) Novel intronic microRNA represses zebrafish myf5 promoter activity through silencing dickkopf-3 gene. *Nucleic Acids Res*, 38, 4384-93.
- Huang, d. W., B. T. Sherman & R. A. Lempicki (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37, 1-13.
- (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44-57.
- Institute, N. F. 1991. A Manual of Standard Methods for Measuring and Specifying the Properties of Surimi. ed. W. National Fisheries Institute, DC.
- Jin, W., M. V. Dodson, S. S. Moore, J. A. Basarab & L. L. Guan (2010) Characterization of microRNA expression in bovine adipose tissues: a potential regulatory mechanism of subcutaneous adipose tissue development. *BMC Mol Biol*, 11, 29.
- Johnston, I. (1999) Muscle development and growth: potential implications for flesh quality in fish. *Aquaculture*, 177, 99-115.
- Johnston, I. A., R. Alderson, C. Sandham, A. Dingwall, D. Mitchell, C. Selkirk, D. Nickell, R. Baker, B. Robertson, D. Whyte & J. Springate (2000) Muscle fibre density in relation to the colour and texture of smoked Atlantic salmon (*Salmo salar* L.). *Aquaculture*, 189, 335-349.
- Johnston, I. A., H. T. Lee, D. J. Macqueen, K. Paranthaman, C. Kawashima, A. Anwar, J. R. Kinghorn & T. Dalmay (2009) Embryonic temperature affects muscle fibre recruitment in adult zebrafish: genome-wide changes in gene and microRNA expression associated with the transition from hyperplastic to hypertrophic growth phenotypes. *J Exp Biol*, 212, 1781-93.
- Juanchich, A., P. Bardou, O. Rué, J. C. Gabillard, C. Gaspin, J. Bobe & Y. Guiguen (2016) Characterization of an extensive rainbow trout miRNA transcriptome by next generation sequencing. *BMC Genomics*, 17, 164.
- Kofler, R., R. V. Pandey & C. Schlötterer (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, 27, 3435-6.
- Krek, A., D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel & N. Rajewsky (2005) Combinatorial microRNA target predictions. *Nat Genet*, 37, 495-500.
- Leeds, T. D., R. L. Vallejo, G. M. Weber, D. G. Pena & J. S. Silverstein (2016) Response to five generations of selection for growth performance traits in rainbow trout (*Oncorhynchus mykiss*). *Aquaculture*, 465, 341-351.
- Lewis, B. P., C. B. Burge & D. P. Bartel (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120, 15-20.



- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin & G. P. D. P. Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078-9.
- Ling, B. M., N. Bharathy, T. K. Chung, W. K. Kok, S. Li, Y. H. Tan, V. K. Rao, S. Gopinadhan, V. Sartorelli, M. J. Walsh & R. Taneja (2012) Lysine methyltransferase G9a methylates the transcription factor MyoD and regulates skeletal muscle differentiation. *Proc Natl Acad Sci U S A*, 109, 841-6.
- Liu, Q., H. Fu, F. Sun, H. Zhang, Y. Tie, J. Zhu, R. Xing, Z. Sun & X. Zheng (2008) miR-16 family induces cell cycle arrest by regulating multiple cell cycle genes. *Nucleic Acids Res*, 36, 5391-404.
- Lopes, C. T., M. Franz, F. Kazi, S. L. Donaldson, Q. Morris & G. D. Bader (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26, 2347-8.
- Luo, W., E. Li, Q. Nie & X. Zhang (2015) Myomaker, Regulated by MYOD, MYOG and miR-140-3p, Promotes Chicken Myoblast Fusion. *Int J Mol Sci*, 16, 26186-201.
- Manor, M. L., B. M. Cleveland, P. B. Kenney, J. Yao & T. Leeds (2015) Differences in growth, fillet quality, and fatty acid metabolism-related gene expression between juvenile male and female rainbow trout. *Fish Physiology and Biochemistry*, 41, 533-547.
- Marancik, D., G. Gao, B. Paneru, H. Ma, A. G. Hernandez, M. Salem, J. Yao, Y. Palti & G. D. Wiens (2014) Whole-body transcriptome of selectively bred, resistant-, control-, and susceptible-line rainbow trout following experimental challenge with *Flavobacterium psychrophilum*. *Front Genet*, 5, 453.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytzky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly & M. A. DePristo (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20, 1297-303.
- Merico, D., R. Isserlin, O. Stueker, A. Emili & G. D. Bader (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*, 5, e13984.
- Messeguer, X., R. Escudero, D. Farré, O. Núñez, J. Martínez & M. M. Albà (2002) PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics*, 18, 333-4.
- Mishima, Y., C. Abreu-Goodger, A. A. Staton, C. Stahlhut, C. Shou, C. Cheng, M. Gerstein, A. J. Enright & A. J. Giraldez (2009) Zebrafish miR-1 and miR-133 shape muscle gene expression and regulate sarcomeric actin organization. *Genes Dev*, 23, 619-32.
- Mørkøre, T., J. I. Vallet, M. Cardinal, M. C. Gomez-Guillen, P. Montero, O. J. Torrissen, R. Nortvedt, S. Sigurgisladottir & M. S. Thomassen (2001) Fat content and fillet shape of Atlantic Salmon: Relevance for processing, yield and quality of raw and smoked products. *Journal of food science*, 66, 1348-1354.
- Olsen, P. H. & V. Ambros (1999) The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol*, 216, 671-80.

- Paneru, B., R. Al-Tobasei, Y. Palti, G. D. Wiens & M. Salem (2016) Differential expression of long non-coding RNAs in three genetic lines of rainbow trout in response to infection with *Flavobacterium psychrophilum*. *Sci Rep*, 6, 36032.
- Ponsuksili, S., Y. Du, F. Hadlich, P. Siengdee, E. Murani, M. Schwerin & K. Wimmers (2013) Correlated mRNAs and miRNAs from co-expression and regulatory networks affect porcine muscle and finally meat properties. *BMC Genomics*, 14, 533.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly & P. C. Sham (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81, 559-75.
- Qin, L., Y. Chen, Y. Niu, W. Chen, Q. Wang, S. Xiao, A. Li, Y. Xie, J. Li, X. Zhao, Z. He & D. Mo (2010) A deep investigation into the adipogenesis mechanism: profile of microRNAs regulating adipogenesis by modulating the canonical Wnt/beta-catenin signaling pathway. *BMC Genomics*, 11, 320.
- Rueda, A., G. Barturen, R. Lebrón, C. Gómez-Martín, Á. Alganza, J. L. Oliver & M. Hackenberg (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res*, 43, W467-73.
- Salem, M., P. B. Kenney, C. E. Rexroad, 3rd & J. Yao (2006) Microarray gene expression analysis in atrophying rainbow trout muscle: a unique nonmammalian muscle degradation model. *Physiol Genomics*, 28, 33-45.
- Salem, M., M. L. Manor, A. Aussanasuwannakul, P. B. Kenney, G. M. Weber & J. Yao (2013) Effect of sexual maturation on muscle gene expression of rainbow trout: RNA-Seq approach. *Physiol Rep*, 1, e00120.
- Salem, M., B. Paneru, R. Al-Tobasei, F. Abdouni, G. H. Thorgaard, C. E. Rexroad & J. Yao (2015) Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout. *PLoS ONE*.
- Salem, M., R. L. Vallejo, T. D. Leeds, Y. Palti, S. Liu, A. Sabbagh, C. E. Rexroad & J. Yao (2012) RNA-Seq identifies SNP markers for growth traits in rainbow trout. *PLoS One*, 7, e36264.
- Schmittgen, T. D. & K. J. Livak (2008) Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc*, 3, 1101-8.
- Sethupathy, P. & F. S. Collins (2008) MicroRNA target site polymorphisms and human disease. *Trends Genet*, 24, 489-97.
- Steelman, C. A., J. C. Recknor, D. Nettleton & J. M. Reecy (2006) Transcriptional profiling of myostatin-knockout mice implicates Wnt signaling in postnatal skeletal muscle growth and hypertrophy. *FASEB J*, 20, 580-2.
- Tobasei, R., B. Paneru, T. Leeds, B. Kenney & M. Salem (2016) Genome-wide discovery of long non-coding RNAs in rainbow trout. *PLoS one*.
- Wang, X. H. (2013) MicroRNA in myogenesis and muscle atrophy. *Curr Opin Clin Nutr Metab Care*, 16, 258-66.
- Wang, Y., X. Li & H. Hu (2011) Transcriptional regulation of co-expressed microRNA target genes. *Genomics*, 98, 445-52.
- Wu, L., J. Fan & J. G. Belasco (2006) MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A*, 103, 4034-9.

- Yan, B., J. T. Guo, C. D. Zhu, L. H. Zhao & J. L. Zhao (2013a) miR-203b: a novel regulator of MyoD expression in tilapia skeletal muscle. *J Exp Biol*, 216, 447-51.
- Yan, B., C. D. Zhu, J. T. Guo, L. H. Zhao & J. L. Zhao (2013b) miR-206 regulates the growth of the teleost tilapia (*Oreochromis niloticus*) through the modulation of IGF-1 gene expression. *J Exp Biol*, 216, 1265-9.
- Yuan, Y., Y. Shen, L. Xue & H. Fan (2013) miR-140 suppresses tumor growth and metastasis of non-small cell lung cancer by targeting insulin-like growth factor 1 receptor. *PLoS One*, 8, e73604.
- Zhang, B., Q. Wang & X. Pan (2007) MicroRNAs and their regulatory roles in animals and plants. *J Cell Physiol*, 210, 279-89.

**APPENDICES**

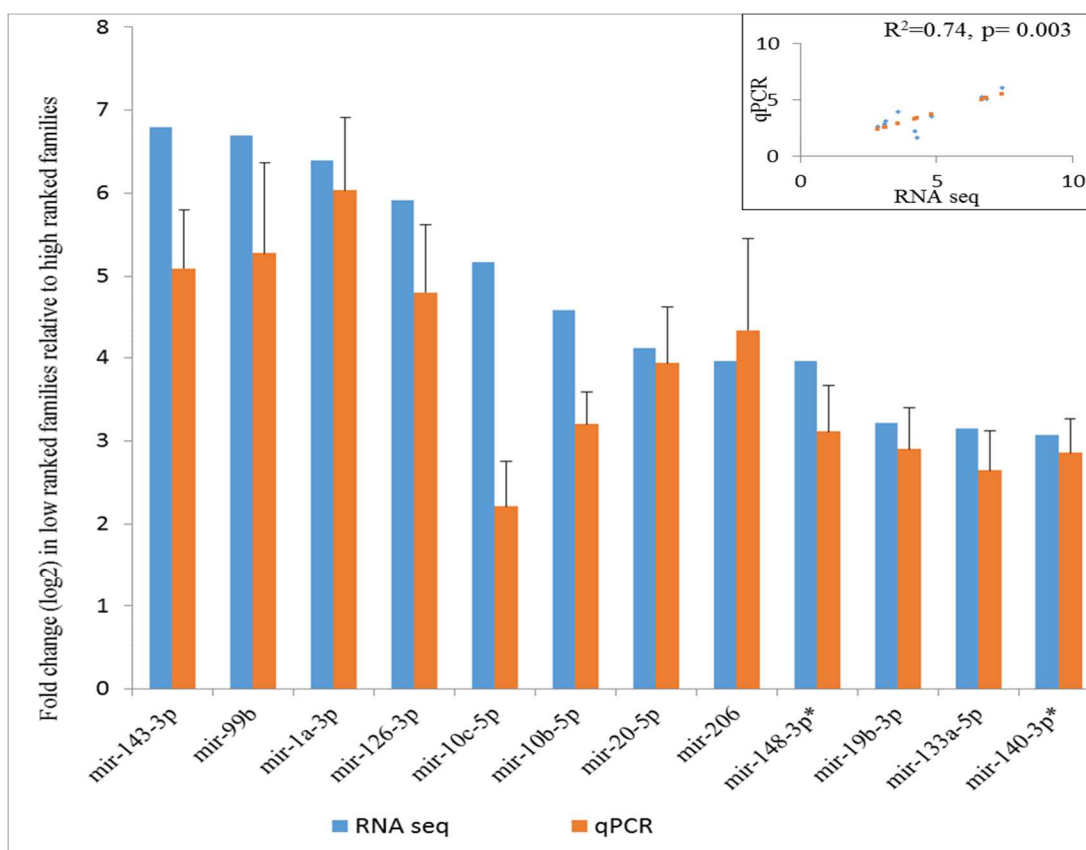
## APPENDIX A

**CORRELATION BETWEEN DIFFERENT GROWTH AND MUSCLE QUALITY TRAITS. CORRELATION WAS CALCULATED FROM PHENOTYPIC DATA OF ABOUT 500 INDIVIDUAL FISH OF USDA-ARS-NCCCWA'S GROWTH SELECTED LINE USED IN THIS STUDY. TABLE SHOWS CORRELATION (R) BETWEEN EACH PAIR OF TRAITS AND P VALUE INSIDE THE BRACKETS. NEGATIVE (-) VALUE INDICATES NEGATIVE CORRELATION**

	<b>Whole body weight</b>	<b>Muscle yield</b>	<b>Muscle shear force</b>	<b>Muscle whiteness</b>
<b>Muscle yield</b>	0.568 (< 0.001)			
<b>Muscle shear force</b>	0.212 (< 0.001)	0.179 (< 0.001)		
<b>Muscle whiteness</b>	0.032 (> 0.001)	-0.152 (< 0.001)	-0.173 (< 0.001)	
<b>Muscle crude fat content</b>	0.573 (< 0.001)	0.247 (< 0.001)	0.003 (> 0.001)	0.224 (< 0.001)

## APPENDIX B

**REAL TIME PCR VALIDATION OF FOLD CHANGE OF 12 MICRORNAS DE BETWEEN HIGH AND LOW MUSCLE YIELD GROUP. RNA-SEQ AND QPCR SHOW CONSISTENCY IN EXPRESSION PATTERN OF MICRORNAS. FOR QPCR ANALYSIS, TOTAL RNA FROM 16 INDIVIDUAL FISH FROM EACH HIGH RANKED AND LOW RANKED FAMILY (N = 32) WAS USED. ERROR BARS ON QPCR DATA REPRESENT STANDARD DEVIATION. FOLD CHANGE WAS STATISTICALLY SIGNIFICANT BY RNA SEQ (FDR-P>0.05) AS WELL AS BY QPCR (P>0.05). FOR QPCR DATA P VALUE WAS CALCULATED USING NON-PARAMETRIC MANN-WHITNEY U TEST**



## APPENDIX C

## ENRICHED GENE PATHWAYS (IN BIOLOGICAL PROCESS CATEGORIES)

## AMONG TARGET GENES OF DE MICRORNAS BETWEEN HIGH VS LOW

## MUSCLE YIELD

Enriched GO terms in biological process category for muscle DE microRNA targets			
Term	Count	Fold Enrichment	FDR
(GO:0061061)~muscle structure development	33	3.16	4.86E-05
(GO:0007517)~muscle organ development	22	4.35	6.05E-05
(GO:0014706)~striated muscle tissue development	23	4.04	1.09E-04
(GO:0060537)~muscle tissue development	23	4.01	1.24E-04
(GO:0051146)~striated muscle cell differentiation	21	3.4	6.30E-03
(GO:0007519)~skeletal muscle tissue development	15	4.54	6.31E-03
(GO:0042692)~muscle cell differentiation	23	3.14	7.90E-03
(GO:0060538)~skeletal muscle organ development	15	4.09	2.13E-02
(GO:0055001)~muscle cell development	19	3.31	2.73E-02
(GO:0055002)~striated muscle cell development	18	3.45	2.74E-02
GO:0008045~motor neuron axon guidance	12	5.858717041	0.004520748
GO:0016569~covalent chromatin modification	27	5.37049062	2.04E-09
GO:0030030~cell projection organization	14	5.24559549	0.001984335
GO:0007067~mitotic nuclear division	25	4.9726765	1.01E-07
GO:0007049~cell cycle	47	4.67431591	8.91E-16
GO:0060271~cilium morphogenesis	16	4.603277675	0.001705472
GO:0007275~multicellular organism development	159	4.574507189	2.61E-60
GO:0051301~cell division	30	4.559850527	9.40E-09
GO:0006974~cellular response to DNA damage stimulus	29	4.209303459	1.79E-07
GO:0030154~cell differentiation	61	4.199999075	1.09E-18
GO:0007399~nervous system development	40	4.027867965	1.28E-10
GO:0008380~RNA splicing	16	4.027867965	0.010493515
GO:0016055~Wnt signaling pathway	29	3.678997512	4.97E-06
GO:0006397~mRNA processing	23	3.399668374	0.001236411
GO:0015031~protein transport	40	3.288055482	1.10E-07
GO:0006351~transcription, DNA-templated	150	3.222294372	4.91E-36
GO:0042384~cilium assembly	29	3.200223863	1.24E-04
GO:0006915~apoptotic process	27	3.129566477	6.03E-04
GO:0007507~heart development	30	2.715416606	0.002636313
GO:0016567~protein ubiquitination	36	2.636422668	4.07E-04
GO:0006355~regulation of transcription, DNA-templated	160	1.964813642	2.78E-14

## APPENDIX D

### ENRICHED GENE PATHWAYS (IN BIOLOGICAL PROCESS CATEGORIES)

#### AMONG TARGET GENES OF DE MICRORNAS BETWEEN HIGH VS LOW

#### CRUDE FAT CONTENT

Enriched GO terms in biological process category for fat DE microRNA targets			
Term	Count	Fold Enrichment	FDR
GO:0007275~multicellular organism development	41	4.677858676	6.66E-13
GO:0030154~cell differentiation	15	4.095686145	0.027802
GO:0006351~transcription, DNA-templated	39	3.322420601	1.60E-07
GO:0006355~regulation of transcription, DNA-templated	43	2.09404441	0.007846



## APPENDIX E

## SHEAR FORCE ASSOCIATED MICRORNAS AND THEIR TARGET GENES

## CODING FOR COLLAGEN AND COLLAGEN REGULATORS

microRNA	Target mRNA/gene ID	Gene name
mir-122-5p	GSONMT00038907001	collagen alpha-2 chain-like
mir-132-3p	GSONMT00019296001	collagen alpha-1 chain-like
mir-23b-5p	GSONMT00003289001	collagen type xi alpha 1 short isoform
mir-196b-5p	GSONMT00000976001	collagen type i alpha 2
mir-199-5p	GSONMT00021453001	collagen alpha-2 chain
mir-21-3p	GSONMT00040586001	collagen type i alpha 1
mir-7-5p	GSONMT00044213001	collagen alpha-5 chain isoform x1
let-7d-5p	GSONMT00064790001	collagen alpha-1 chain b-like
mir-96-5p	GSONMT00032499001	procollagen c-endopeptidase enhancer 2-like
mir-1a-3p	GSONMT00004527001	procollagen- $\alpha$ 1(I) procollagen-1
mir-148-3p	GSONMT00058103001	collagen alpha-1 chain
mir-148-3p*	GSONMT00022102001	collagen alpha-1 chain-like isoform x1
mir-19b-3p	GSONMT00037024001	72 kda type iv collagenase precursor
mir-16-5p	GSONMT00066815001	collagen alpha-1 chain-like isoform x2
mir-125a-3p	GSONMT00025592001	procollagen galactosyltransferase 1-like
mir-200a-3p	GSONMT00070223001	collagen alpha-3 chain
mir-222-3p	GSONMT00063749001	acetylcholinesterase collagenic tail peptide
mir-1a-3p	GSONMT00060667001	f-box only protein 42
mir-1a-3p	GSONMT00033781001	glucokinase
mir-132-3p	GSONMT00025862001	rna-directed dna polymerase from mobile element jockey-like
mir-200b-3p	GSONMT00015752001	complement c1q tumor necrosis factor-related protein 1-like
mir-27b-5p	GSONMT00010370001	bchain crystal structure of anticalin n7a in complex with oncofetal fibronectin fragment fn7b8
mir-125a-3p	GSONMT00037416001	complement c1q subcomponent subunit b-like
mir-125a-3p	GSONMT00031844001	egl nine 1-like protein
mir-183-5p	GSONMT00054212001	serpin h1-like isoform x1
mir-18a-5p	GSONMT00003567001	thrombospondin-1-like
mir-200a-3p	GSONMT00058525001	forkhead box protein j2-like
mir-200a-3p	GSONMT00025830001	probable histone deacetylase 1-b-like
mir-222-3p	GSONMT00030989001	endothelin-converting enzyme 2-like
mir-222-3p	GSONMT00021172001	lumican precursor
mir-222-3p	GSONMT00044809001	complement c1q-like protein 2-like
mir-30d-5p	GSONMT00020641001	tgf-beta receptor type-1-like
mir-30d-5p	GSONMT00037954001	transforming growth factor beta-3-like
mir-7-5p	GSONMT00030838001	complement c1q tumor necrosis factor-related protein 7-like
mir-203-3p	GSONMT00035274001	tgf-beta receptor type-1
mir-1a-3p	GSONMT00028793001	protein-lysine 6-oxidase-like
mir-148-3p	GSONMT00038792001	prolyl 3-hydroxylase 2-like
mir-20-5p	GSONMT00081582001	mothers against decapentaplegic homolog 4
mir-20-5p	GSONMT00016866001	coiled-coil and c2 domain-containing protein 2a
mir-140-3p	GSONMT00019727001	translocating chain-associated membrane protein 2
mir-140-3p	GSONMT00067514001	nucleolar gtp-binding protein 1
mir-143-5p	GSONMT00075404001	disintegrin and metalloproteinase domain-containing protein 9

## APPENDIX F

## ENRICHED GENE PATHWAYS (IN BIOLOGICAL PROCESS CATEGORIES)

## AMONG TARGET GENES OF DE MICRORNAS BETWEEN HIGH VS LOW

## FILLET WHITENESS INDEX

GO Term	Count	Fold Enrichment	FDR
GO:0006351~transcription, DNA-templated	237	3.715870458	9.55E-75
GO:0007275~multicellular organism development	202	4.241669488	5.59E-74
GO:0006355~regulation of transcription, DNA-templated	256	2.294455362	1.37E-36
GO:0007399~nervous system development	63	4.630139218	6.84E-23
GO:0030154~cell differentiation	75	3.768937092	2.77E-21
GO:0016569~covalent chromatin modification	41	5.952128801	8.96E-19
GO:0007049~cell cycle	55	3.992281513	3.60E-16
GO:0015031~protein transport	55	3.299742883	5.17E-12
GO:0006397~mRNA processing	37	3.991615577	5.88E-10
GO:0051301~cell division	36	3.993651078	1.26E-09
GO:0008380~RNA splicing	26	4.777127765	3.66E-08
GO:0007067~mitotic nuclear division	29	4.210042322	7.24E-08
GO:0006974~cellular response to DNA damage stimulus	34	3.601881502	1.41E-07
GO:0006810~transport	149	1.564378103	3.53E-05
GO:0006915~apoptotic process	34	2.876322639	7.30E-05
GO:0007507~heart development	39	2.576428457	1.49E-04
GO:0016568~chromatin modification	16	5.08500918	2.15E-04
GO:0002376~immune system process	16	4.951193149	3.28E-04
GO:0016567~protein ubiquitination	43	2.298366365	8.36E-04
GO:0048511~rhythmic process	9	8.819312796	0.001666454
GO:0030901~midbrain development	14	4.988702188	0.002225424
GO:0008045~motor neuron axon guidance	14	4.988702188	0.002225424
GO:0045893~positive regulation of transcription, DNA-templated	25	2.969465588	0.003346071
GO:0016055~Wnt signaling pathway	29	2.685145103	0.003724562
GO:0031047~gene silencing by RNA	12	5.64436019	0.004016625
GO:0030097~hemopoiesis	23	2.846936061	0.019569321
GO:0006914~autophagy	16	3.549912069	0.039154273
GO:0010002~cardioblast differentiation	7	9.145954011	0.043091861
GO:0006417~regulation of translation	14	3.919694576	0.047995316
GO:0009880~embryonic pattern specification	11	4.974996962	0.04891121

## CHAPTER V

### A CLOSER LOOK AT THE MUSCLE “DEGRADOME” IN RAINBOW TROUT: FUNCTIONAL INTERPLAY AMONG LNC-RNAS, MICRORNAS AND PROTEIN CODING GENES

#### ABSTRACT

In fish, coding and non-coding genes involved in muscle atrophy are not fully characterized. In addition, the functional interplays that exist between different classes of non-coding RNAs and protein coding genes that regulate muscle atrophy remains unknown. Muscle degradation triggered by energetic demands at sexual maturation represents an excellent model system for studying the muscle deterioration in salmonids. Using a RNA-Seq approach, we identified 852 mRNAs, 1,198 lncRNAs and 28 microRNAs that were differentially expressed (DE) in atrophying muscle in Rainbow trout during sexual maturation. Muscle atrophy appeared to be mediated by many genes encoding ubiquitin-proteasome system, autophagy related proteases, lysosomal proteases and transcription factors. Four transcripts encoding atrogen-1 showed exceptional upregulation in atrophying muscle suggesting their important role in bulk muscle proteolysis during atrophy. DE lncRNA and mRNA genes tend to be co-localized in genome, and showed evidence of extensive ‘lncRNA-mRNA’, ‘lncRNA-microRNA’, ‘mRNA-microRNA’ and ‘lncRNA-protein’ physical interactions. Co-localized or physically interacted lncRNA and protein coding genes were more strongly correlated in expression ( $R > 0.70$ ) compared to random lncRNA and protein coding genes. DE genes showing potential functional interactions comprised highly correlated ‘lncRNA-mRNA-microRNA’ gene network that I designated as ‘degradome’, which appeared to be the

global gene regulatory network implicated in muscle atrophy. This study pinpoints the existence of extensive coding and non-coding RNA interactions during muscle atrophy, and suggest several elements for future mechanistic studies that should reveal how individual genes in the interaction network contribute to the fish muscle atrophy.

## **INTRODUCTION**

Sexual maturation, starvation and several pathological conditions cause fish skeletal muscle atrophy that negatively affects growth performance and fillet qualities (Salem et al. 2006a, Aussanasuwannakul et al. 2012, Salem et al. 2013). Strategy of improving growth performance and fillet qualities by reducing protein turnover requires an understanding of muscle proteolytic system. Previously, several studies were performed to identify the protein coding genes associated with skeletal muscle atrophy in fish (Salem et al. 2006a, Aussanasuwannakul et al. 2012, Salem et al. 2013). A previous microarray study identified about 200 protein coding genes differentially expressed (DE) during sexual maturation associated muscle atrophy in trout (Salem et al. 2006a). The same study revealed upregulated expression of catheptic and collagenase proteolytic pathways during muscle atrophy. Yet, another study found increase in the catalytic activity of calpain and 28S proteasome subunit during starvation induced skeletal muscle atrophy in trout (Salem et al. 2007). However, some of the previous findings were inconsistent and do not provide a comprehensive dataset of protein coding genes associated with muscle atrophy. As an example, some studies have reported downregulation of ubiquitin-proteasome system during atrophy (Salem et al. 2006a), while others have reported its upregulation (Cleveland and Evenhuis 2010, Tacchi et al. 2010). These studies either investigated a single protein coding gene (Cleveland and Evenhuis 2010, Tacchi et al. 2010) or limited sets of protein

coding genes (Salem et al. 2006a) due to lack of a holistic approach that was available at the time. In addition, none of the previous studies have investigated the role of microRNAs and lncRNAs in skeletal muscle atrophy in Rainbow trout. An approach that is more robust is needed to discover all the potential candidate genes involved in muscle atrophy.

MicroRNAs binds 3' -UTR of mRNA which leads to downregulation of the gene by various mechanisms such as translation suppression (Olsen and Ambros 1999), target mRNA breakdown (Bagga et al. 2005) and deadenylation of poly A tail (Wu, Fan and Belasco 2006). There are evidences that a single microRNA can regulate hundreds of genes whereas the same gene can be regulated by several microRNAs (Krek et al. 2005). MicroRNAs are known to regulate muscle proteolysis and muscle atrophy in different mammalian species (Wang 2013). For example, mir-486 regulates disease related muscle atrophy in mice by regulating FOXO1 transcription factor (Xu et al. 2012). MicroRNA, mir-182 indirectly regulates expression of key muscle atrophy genes including atrogen-1, cathepsin and autophagy related genes by targeting transcription factor FOXO3 (Hudson et al. 2014). Muscle specific microRNA mir-1 regulates dexamethasone mediated muscle atrophy by targeting heat shock protein 70 (HSP70) (Kukreti et al. 2013). However, microRNAs that regulate muscle atrophy were not investigated before in aquaculture salmonids.

LncRNAs are new classes of non-coding RNAs with critical gene regulatory roles (Rinn and Chang 2012). LncRNAs are known to regulate genes by direct physical interaction with microRNAs, mRNAs and proteins. Several lncRNAs bind microRNAs by sequence complementarity, that leads to sequestration of cellular microRNA (sponge effect) and lncRNA-mRNA competition for microRNA binding (Yoon, Abdelmohsen and

Gorospe 2014). For example, lncRNA H19 binds and sponges away let-7 family microRNAs from repressing its protein coding targets (Kallen et al. 2013). Similarly, muscle specific lncRNA linc-MD1 competes with MAML1 and MEF2C to bind microRNAs mir-135 and mir-133, respectively (Cesana et al. 2011). LncRNA MALAT1 modulates mir-133 mediated downregulation of serum response factor (SRF) by sharing mir-133 binding site (Han et al. 2015). LncRNA's direct physical interaction with mRNA leading to mRNA decay (Gong and Maquat 2011) and translation suppression (Yoon et al. 2012). Some lncRNAs hybridize in 3' UTR of target mRNA and facilitate mRNA decay (Gong and Maquat 2011). On the other hand, lincRNA-p21 directly bind *JUNB* and *CTNNB1* mRNAs and suppress their translation (Yoon et al. 2012). LncRNA's physical interaction with proteins modulates the stability (Taniue et al. 2016), cellular availability (sequestration) (Hirose et al. 2014), activity (Tripathi et al. 2010) and cellular localization (Tripathi et al. 2010) of the protein. For example, lncRNA UPAT1 binds to UHRF1 protein and interferes with its ubiquitination and subsequent degradation (Taniue et al. 2016). LncRNA MALAT1 binds to SR proteins and regulates their phosphorylation and hence cellular localization (Tripathi et al. 2010). Both 'lncRNA-microRNA' and 'lncRNA-protein coding genes' interactions are known to regulate development (Cesana et al. 2011), disease (Li et al. 2016) and cancers (Wang et al. 2010, Ma et al. 2015); however, their involvement in skeletal muscle atrophy remains unknown.

To identify coding and non-protein coding genes involved in sexual maturation associated muscle atrophy, we sequenced mRNAs, lncRNAs and microRNAs from atrophying skeletal muscle of gravid fish and normal skeletal muscle of sterile fish, and performed differential gene expression between the two groups. We investigated functional

interactions between differentially expressed (DE) lncRNAs, microRNAs and protein coding genes in terms of expression correlation, genome co-localization and physical interaction to investigate gene-regulatory circuits during muscle atrophy. This study provides the first genome-wide lncRNA-mRNA-microRNA interaction network during fish muscle degradation and will help understand how energetic demand at sexual maturation triggers skeletal muscle atrophy.

## **MATERIALS AND METHODS**

### **Fish population and muscle sampling**

We previously described fish population used in this study in different studies (Salem et al. 2006a). Briefly, mature sterile (3N: triploid) and fertile (2N: diploid) female Rainbow trout (about 500 gram) were obtained from Flowing Springs Trout Farm (Delray, WV) during sexual maturation season and were cultured in identical raceways. Water from a common spring was circulated in raceways at temperature  $13 \pm 3^{\circ}\text{C}$ . Ad libitum (Zeiglar Gold; Zeigler Bros., Gardeners, PA) was fed to both groups via demand feeder until sampling. At the time of muscle sampling gonado-somatic index (GSI) of fertile fish was  $15.8 \pm 0.3$  ( $n = 5$ ) compared to  $0.3 \pm 0.2$  ( $n = 5$ ) in sterile fish confirming gravid stage of fertile fish. White muscle tissue of 8 fish (4 fertile and 4 sterile) was collected from dorsal musculature, flash frozen in liquid nitrogen and were stored at  $-80^{\circ}\text{C}$  until RNA extraction. Total RNA was extracted from the muscle using TRIZol method (Invitrogen, Carlsbad, CA). For phenotype measurement, boneless and skinless muscle fillet was obtained in a filleting procedure. Muscle yield was measured as percentage of whole body weight. Small piece of muscle fillet was used for compositional and proximate analysis (e.g. crude protein content, crude fat content, shear force, moisture content and pH).

### **Library construction and sequencing**

Libraries were generated with Truseq Ribo-Zero gold protocol following the manufacturer's recommendations (Illumina Inc, CA, USA). One sequencing library was prepared from each fish and was provided a unique barcode. Equal amount of barcoded libraries from all 8 fish were pooled and sequenced using Illumina Hiseq 2000 sequencing platform in a single lane ( $2 \times 100$  reads). Similarly, for microRNA sequencing, Illumina's Tru Seq small RNA library preparation kit was used to prepare one barcoded library from each 8 fish and libraries were pooled and sequenced in a single lane of Illumina Hiseq 2000 sequencing platform.

### **Discovery of LncRNAs**

LncRNAs from sequencing reads were identified by using the pipeline we described previously (Al-Tobasei, Paneru and Salem 2016).

### **Identification of DE mRNA, lncRNAs and microRNAs**

Read mapping and identification of DE genes was performed in CLC genomics workbench. For protein coding genes, sequencing reads from every fish were mapped to mRNA reference from Rainbow trout genome (Berthelot et al. 2014) and transcriptome assemblies (Salem et al. 2015). The expression value of each transcript was calculated in terms of TPM (transcript per million), and DE mRNAs between gravid and sterile fish were identified using EDGE test (FDR-P -value  $< 0.01$ , fold change:  $> 3$  or  $< -3$ ). For lncRNAs, previously published Rainbow trout lncRNAs (Al-Tobasei et al. 2016) and additional lncRNAs assembled from this sequencing project were used as a reference. Read mapping and identification of DE lncRNA was done as described for mRNAs. For microRNAs, sequencing adapters were trimmed and reads were mapped to miRBase microRNA



reference (release 21) (mismatch  $\leq 2$ , additional/missing upstream/downstream bases  $\leq 2$ ). The total read count for each microRNA was calculated, and used to identify DE microRNAs by EDGE test (FDR-p-value  $< 0.01$ , fold change:  $> 3$  or  $< -3$ ).

### **Real time PCR validation of DE transcripts**

Expression of 4 atrogen-1 isoforms, 7 random lncRNAs and 3 random microRNAs was individually validated by real time PCR. Total RNA from the same 8 fish used for sequencing was used to make template cDNA for qPCR analysis. Contaminating DNA in RNA sample was removed by DNase treatment and cDNA was synthesized by Verso cDNA Kit (Thermo Scientific, Hudson, NH). Transcript abundance was quantified per manufacturer's instruction using DyNAmo Flash SYBR Green Master Mix (Thermo Scientific, Hudson, NH) in Bio Rad CFX96™ System (Bio Rad, Hercules, CA). For microRNAs, miScript II RT kit (Qiagen, Valencia, CA, USA) was used to synthesize cDNA, and miScript<sup>R</sup> SYBR<sup>R</sup> green (Qiagen, Valencia, CA, USA) (Ramachandra et al. 2008) was used to quantify microRNA in Bio Rad CFX96™ System. Endogenous control used for normalization were actin for mRNA, and U6 for both lncRNA and microRNA. Fold changes in gene expression was calculated by using  $\Delta\Delta C_t$  method as described previously (Marancik et al. 2014, Paneru et al. 2016).

### **Identification of tissue specific genes**

Expression pattern of DE genes was investigated across 13 vital tissues: red muscle, white muscle, spleen, liver, skin, testis, brain, intestine, stomach, kidney, head kidney, gill and fat. To identify tissue specific genes, we used a statistical approach described by (Li et al. 2014). Normalized expression value (z score) of every gene was calculated in each tissue from TPM counts. Each gene was classified as 'specific' to a tissue if z score was greater

than 1.5 in that tissue and less than 1.5 in remaining 12 tissues which is similar to a previous study (Li et al. 2014). We used the same approach to identify genes ‘specific’ to particular month during pre-spawning and spawning season (Salem et al. 2013) as described in result section.

### **LncRNA, mRNA and microRNA co-expression**

Sample size of 8 (from 2 ploidy groups) was not large enough to study correlation. As a consequence, we included sequencing reads from 22 fish families used to study variation in the growth and muscle quality phenotypes (harvest year 2010). These fish families were produced by selection of specific traits (4 generations) at USDA/NCCCWA which are described in detail (Gonzalez-Pena et al. 2016). Briefly, fish families were reared till about 13 month as described previously (Leeds et al. 2016). Single-sire × single-dam mating was used to produce full sib families and eggs were reared in spring water. Water temperature was varied from 7-13°C to synchronize the hatching time. Each family was reared in a separate 200-L tank at ~ 600 alevins/tank density. The random culling of fish was performed every month to maintain stock density of < 50 kg/m<sup>3</sup>. At the age of 5 month, each fish was given unique identification PIT (passive integrated transponder) tag, and reared in a big 1,000-L commercial tanks. Commercial fishmeal-based diet (16% fat, 42% protein; Ziegler Bros Inc., Gardners, PA) was fed using automatic feeder. Feeding rate was gradually reduced from 2.5% of body weight to 0.5% of body weight as fish grew older. WBW of all fish belonging to 98 families was measured, and families were ranked based on their WBW measurements. Second- or third-ranked fish from each family was selected for muscle sampling so that WBW of sampled fish is adjusted around median of each family. Selected fish were randomly assigned to one of the 5 harvest group (~100

fish/harvest group) and each harvest groups were sampled in 5 consecutive weeks. Fish were anesthetized in 100 mg/L tricaine methanesulfonate and weighted, slaughtered and eviscerated. Muscle samples were separated from dorsal musculature and were stored in liquid nitrogen until RNA extraction and phenotype measurement.

TPM values (for lncRNAs and mRNAs) and the total count (for microRNAs) were calculated in all 30 samples (8 samples from 2 ploidy groups and 22 samples from fish families) by mapping reads to corresponding references mentioned above. TPM and total count values were normalized by using a scaling method as previously described (Bolstad et al. 2003), and normalized values were used to calculate expression correlation coefficients using Pearson correlations. LncRNA-mRNA-microRNA expression network was constructed using Expression Correlation in cytoscape (Lopes et al. 2010).

### **Identification of microRNA-generating lncRNAs**

Rainbow trout genome reference (Berthelot et al. 2014) was annotated with Rainbow trout lncRNA reference sequences mentioned above. Rainbow trout pre-microRNA sequences (about 64 nts long) from recently published source (Juanchich et al. 2016) were aligned to the lncRNA-annotated genome assembly. Pre-microRNA sequences that perfectly align (with no mismatch or gap) within annotated lncRNA loci in the genome were reported.

### **LncRNA and mRNA targets of microRNAs**

In the case of mRNAs, microRNA binding sites were searched in 3' UTR, while in case of lncRNAs, target sites were searched in the entire sequence length. Three target prediction algorithms: miRanda, PITA and TargetSpy were used to find target genes using sRNAtoolbox (Rueda et al. 2015). If the same target site is predicted by all 3 tools, it was

considered as a potential microRNA target site. For all tools, minimum energy threshold was chosen as -20 Kcal/mole. Threshold scores chosen were 150 for Miranda and 0.99 for TargetSpy.

### **Prediction of lncRNA-mRNA and lncRNA-protein interaction**

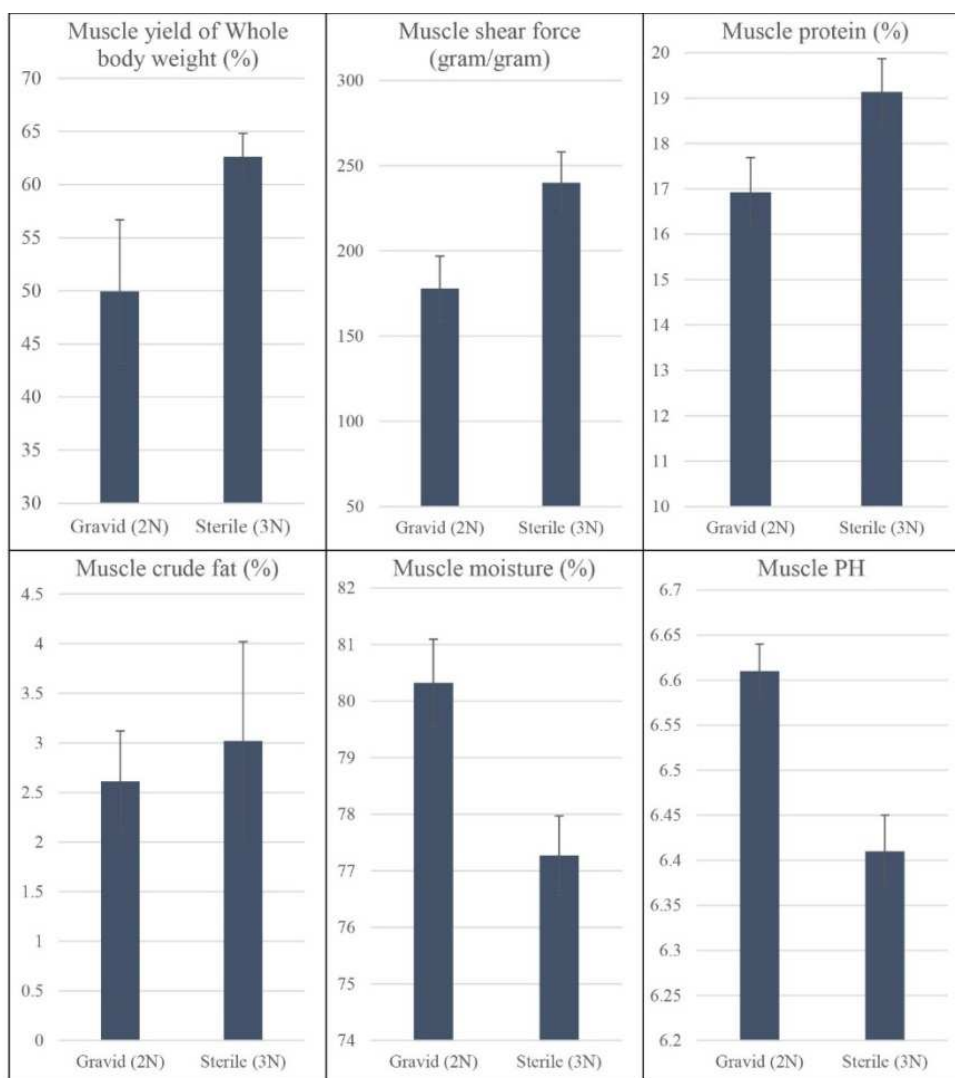
lncRNA-mRNA direct physical interaction was predicted using IntaRNA-RNA-RNA interaction tool (Busch, Richter and Backofen 2008). All lncRNA-mRNA interactions were recorded at interaction energy threshold  $< -100$  Kcal/mole. lncRNA-protein interactions were predicted using CatRapid Omics tool (Agostini et al. 2013). Interaction strength and discriminative power (a measure of predictability of interaction) both were set at a minimum of 96% to consider lncRNA-protein interaction as putative interaction.

## **RESULTS AND DISCUSSION**

### **Atrophying and control skeletal muscle have different compositional and textural characteristics**

Differential expression of coding and non-coding genes was performed in atrophying skeletal muscle from diploid (2N) gravid fish in comparison to that of sterile triploid (3N) fish. The same set of skeletal muscle samples were used in our laboratory in several previous studies (Salem et al. 2006a, Salem et al. 2006b, Salem et al. 2010). Compared to normal skeletal muscle from sterile fish, atrophying muscle from gravid fish had less extractable muscle per whole body weight ( $49.9\% \pm 6.7\%$  vs  $62.6\% \pm 2.2\%$ ,  $p = 0.01$ ), muscle protein ( $16.9\% \pm 0.7\%$  vs  $\sim 19.1\% \pm 0.7\%$ ,  $p = 0.01$ ) and muscle shear force ( $178 \pm 19$  gram/gram vs  $240 \pm 18$  gram/gram,  $p = 0.01$ ). On the other hand, atrophying skeletal muscle had higher moisture content ( $80.3\% \pm 0.7\%$  vs  $77.2\% \pm 0.6\%$ ) and PH

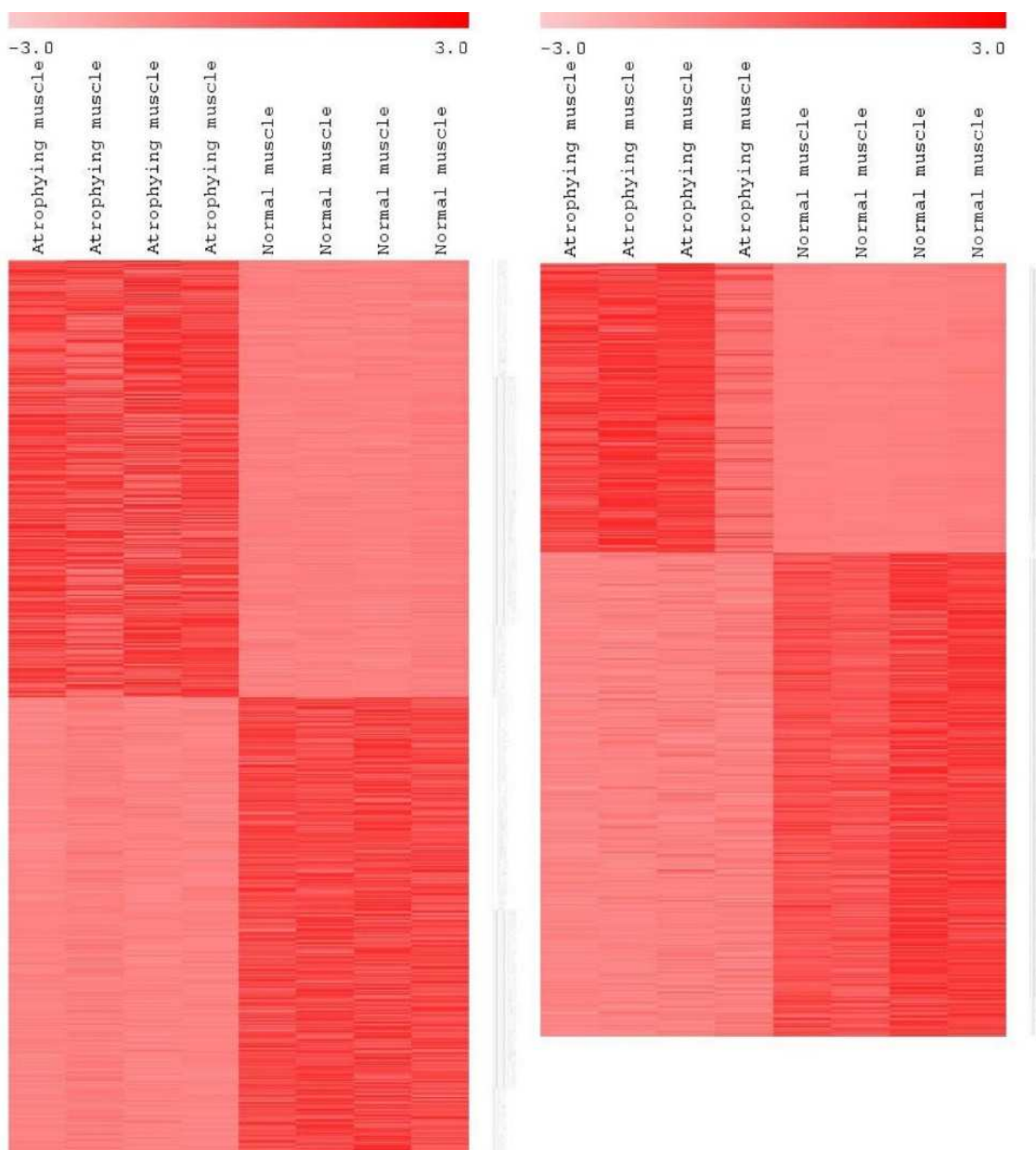
( $6.61 \pm 0.03$  vs  $6.41 \pm 0.04$ ) (Figure 1). Atrophying muscle also had lower crude fat content than normal muscle, but the difference was not statistically significant ( $p = 0.30$ ). These textural and compositional difference between two groups of muscle result from extensive muscle atrophy in gravid fish triggered by energetic demand of sexual maturation.



**Figure 1:** Comparison of different muscle phenotypes between atrophying skeletal muscle from gravid diploid (2N) fish and normal skeletal muscle from sterile triploid (3N) fish.

**mRNAs, lncRNAs and microRNAs are DE between atrophying and control muscle**

To identify genes likely involved in sexual maturation associated skeletal muscle atrophy, we performed deep lncRNA, mRNA and microRNA sequencing, and quantified DE genes between atrophying skeletal muscle from gravid fish and normal skeletal muscle from sterile fish. A total of 852 mRNAs, 1,198 lncRNAs and 28 microRNAs were DE between these two groups (FDR-p-value < 0.01, fold change: > 3 or < -3) (Figure 2 and Table 1). Total of 1,053 transcripts (352 mRNAs, 689 lncRNAs and 12 microRNAs) were upregulated and 1,025 transcripts (500 mRNAs, 509 lncRNAs and 16 microRNAs) were downregulated in atrophying skeletal muscle. Previous microarray based approach performed in the same set of muscle samples identified only 82 upregulated and 120 downregulated protein coding genes (Salem et al. 2006a), suggesting discovery of a large number of additional candidate genes involved in muscle atrophy. DE protein coding genes, lncRNAs and microRNAs are described in separate sections below.



**Figure 2:** Heat map of DE lncRNAs (left) and protein coding genes (right) between atrophying muscle of gravid fish and normal skeletal muscle of sterile fish. Value of color limit represents normalized expression values (Z scores). Fold change in gene expression was considered significant at: FDR-p-value < 0.01, fold change: > 3 or < -3. Darker red and lighter red colors represent higher and lower level of expression respectively.

**Table 1:** DE microRNAs between atrophying muscle of gravid fish and normal skeletal muscle of sterile fish. Positive and negative value of fold change represent upregulation and downregulation respectively in atrophying skeletal. Fold change was considered significant at cut off: 3 > or < -3, FDR-p-value < 0.01.

MicroRNA name	MicroRNA sequence	Fold change	FDR p-value
let-7j	TGAGGTAGTAGGTTGGATAGTT	-1056.3	0.00069
mir-7641-1	TTGATCTCGGAAGCTAAGC	-9.3	0.00002
mir-2187	TTTAATTAGTATAGCCTGTATT	-8.1	0.00800
mir-7551	GGGGCCTGAGTCCTTCTGG	-6.8	0.00878
mir-181a-2	CCATCGACCGTTGACTGTACC	-5.5	0.00211
mir-1a-2	ACATACTTCTTTATGTACCCATA	-5.2	0.00600
mir-7641	CTGAACACGCCCGATCTCGT	-4.7	0.00172
mir-1386	CTCCTGGCTGGCTCGCCA	-3.8	0.00603
let-7c-1	CTGTACAACCTTCTAGCTTTCC	-3.8	0.00211
let-7a-3	CTATACAACCTTACTGTCTTTCC	-3.6	0.00700
mir-203b	AGTGGTCCTAAACATTTTAC	-3.6	0.00800
mir-1-3	ACATACTTCTTTATGCGTCCATA	-3.6	0.00900
mir-148a	AAGTTCTGTGATACACTTCGACT	-3.5	0.00069
mir-125b-1	ACGGGTAGGCTCTTGGGAGCT	-3.3	0.00256
mir-15b	CGAATCATGATGTGCTGCTACT	-3.2	0.00603
mir-133a-1	GCTGGTAAAATGGAACCAAT	-3.0	0.00620
mir-132b	ACCATGGCTGTAGACTGTTACC	3.0	0.00800
let-7d	TGAGGTAGTAGGTTGTAAAGTT	3.3	0.00977
mir-146a	TGAGAACTGAATCCATAGATGG	3.4	0.00610
mir-132-1	ACCGTGGCTTTAGATTGTTACT	3.7	0.00720
miR-29c-3p	ACCGATTTCAAATGGTGCTA	4.2	0.00005
mir-29c	TAGCACCATTTGAAATCGGTTA	5.1	0.00001
let-7	TGAGGTAGTCGGTTGTAAAGA	5.2	0.00069
mir-457b	AGCAGCACATAAATACTGGAG	5.7	0.00025
mir-29b-2	TGACTGATTTCTCTGGTGTTTAGA	7.0	0.00000
mir-29b	TAGCACCATTTGAAATCAGT	7.6	0.00154
mir-29b-1	TAGCACCATTTGAAATCAGTGTT	9.6	0.00001
mir-29a	CTGGTTTCACATGGTGGTTTAGA	11.7	0.00013

### ***Protein coding genes***

Many genes that promote proteolysis were significantly upregulated in atrophying skeletal muscle. At least 37 genes involved in protein ubiquitination, 22 genes involved in autophagy related proteolysis, and 15 lysosomal and other proteases (cathepsin D,



cathepsin B, cathepsin L and cathepsin Z) showed upregulation in atrophying muscle (Table 2). On the other hand, genes that negatively regulate ubiquitin-proteasome system (ubiquitin carboxyl-terminal hydrolase 10, ubiquitin-like domain-containing ctd phosphatase 1 and uridine-cytidine kinase 2) and autophagy (cdgsh iron-sulfur domain-containing protein 2) were downregulated. Amino acid and fat biosynthetic genes showed downregulation while genes involved in amino acid catabolism and transport were highly upregulated (Appendix A). Similarly, genes associated with muscle sarcomere and extracellular matrix were downregulated, in consistent with the loss of muscle mass and shear force during atrophy evidenced from phenotype comparison. As an example, 47 collagen related genes and 24 non-collagen extracellular matrix protein genes were significantly downregulated. Previous study also showed similar expression pattern of genes involved in protein ubiquitination, autophagy-lysosome system, extracellular matrix and sarcomere structure during muscle atrophy in mammal (Llano-Diez et al. 2011). At least 53 transcription factors (TFs) or transcription regulators were also DE, of them, 28 were upregulated and 25 were downregulated (Appendix B). While proteolytic role of majority of the TFs was unknown, 8 highly upregulated TFs had known function in protein catabolism. On the other hand, development related TFs like myoD were downregulated. These findings suggest that muscle atrophy is triggered by upregulation of proteolytic and catabolic genes concomitant with downregulation of muscle sarcomere, extracellular matrix, muscle development and biosynthetic genes.

**Table 2:** Selected proteolytic genes highly upregulated in atrophying skeletal muscle of gravid female Rainbow trout relative to normal skeletal muscle of same-aged sterile Rainbow trout. Fold change was considered significant at cut off:  $3 >$  or  $< -3$ , FDR-p-value  $< 0.01$ .

DE mRNA ID	DE mRNA name	Fold change	FDR p-value correction
<b>Genes involved in ubiquitin-mediated protein degradation</b>			
GSONMT00016768001	f-box only protein 32/fbxo32/atrogin-1	377.71	6.90926E-16
TCONS_00058870	F-box only protein 32/fbxo32/atrogin-1	213.38	9.39594E-07
GSONMT00031929001	f-box only protein 32/fbxo32/atrogin-1	152.44	1.62926E-15
TCONS_00098636	f-box only protein 32/fbxo32/atrogin-1	110.61	1.82E-08
GSONMT00049913001	kelch-like protein 38-like	54.13	3.12816E-06
GSONMT00006333001	kelch-like protein 33-like	37.99	6.8219E-05
GSONMT00021608001	zinc finger and btb domain-containing protein 16-a-like	35.37	2.36468E-08
GSONMT00076944001	tribbles homolog 2	9.59	0.000285129
GSONMT00082158001	otu domain-containing protein 1	9.40	1.47461E-06
GSONMT00079892001	tumor protein p53-inducible nuclear protein 2	7.05	3.30565E-07
GSONMT00000505001	thioredoxin-interacting protein	6.98	2.58513E-05
TCONS_00090611	E3 ubiquitin-protein ligase HERC2-like	6.76	0.00594187
TCONS_00080006	speckle-type POZ protein	6.55	0.000676541
GSONMT00074639001	ubiquitin carboxyl-terminal hydrolase 25-like isoform x2	6.52	2.03969E-06
GSONMT00036946001	ubiquitin-conjugating enzyme e2 g1	6.37	0.000790452
GSONMT00064758001	e3 ubiquitin-protein ligase znr12	6.24	0.004335714
GSONMT00009231001	ddb1- and cul4-associated factor 6-like isoform x4	6.18	2.0242E-05
<b>Lysosomal proteases</b>			
GSONMT00080266001	cathepsin b	4.96	0.000187881
GSONMT00063049001	cathepsin L1	8.46	3.03244E-07
GSONMT00049973001	cathepsin z precursor	3.49	0.00333059
TCONS_00051616	cathepsin D	3.89	4.21867E-05
<b>Autophagy related proteases</b>			
GSONMT00065684001	protein sog3-like isoform x3	63.81	2.42588E-09
GSONMT00024835001	transmembrane protease serine 5-like	21.41	0.000668215
GSONMT00078909001	serine threonine-protein kinase ulk2-like isoform x1	12.20	2.65111E-05
GSONMT00059371001	cysteine protease atg4b	12.07	4.24878E-06
GSONMT00069267001	autophagy-related protein 9a-like isoform x1	11.17	0.000167319
GSONMT00012216001	gamma-aminobutyric acid receptor-associated 1	9.65	6.68987E-07
GSONMT00067581001	serine threonine-protein kinase ulk2	9.31	0.005801126
GSONMT00031082001	autophagy-related protein 2 homolog a-like	7.12	0.000259099
GSONMT00037970001	autophagy-related protein 2 homolog b-like	5.97	0.001297892
GSONMT00075003001	beclin 1-associated autophagy-related key regulator	5.77	0.002235139

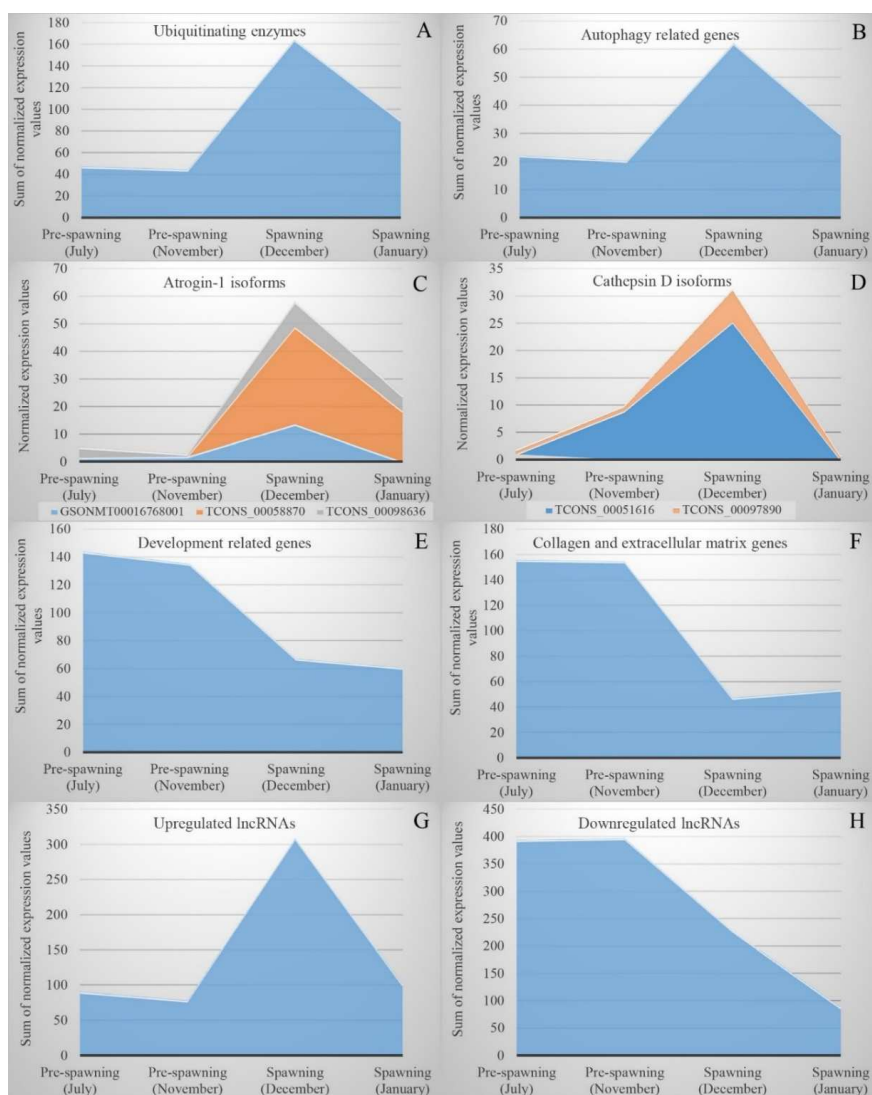
Ubiquitin proteasome system appeared to be the major proteolytic system governing muscle atrophy. F-box only protein 32 (FBXO32) (atrogin-1), an E3 ubiquitin ligase, was the most highly upregulated genes in atrophying muscle suggesting that it might be the major player of bulk muscle proteolysis during atrophy. Four mRNAs coding for atrogin-1 (GSONMT00016768001, TCONS\_00058870, GSONMT00031929001 and TCONS\_00098636) showed 378, 313, 152 and 111-fold upregulation respectively (Table 2). Their expression was validated by real time PCR (data not shown). Overexpression of atrogin-1 during starvation induced skeletal muscle atrophy has been reported previously in Rainbow trout (Cleveland and Evenhuis 2010), Atlantic salmon (Tacchi et al. 2010) and mammals (Gomes et al. 2001).

As reported previously in trout, as fish progress from pre-spawning to spawning month, severity of skeletal muscle atrophy increases as indicated by loss of muscle mass, muscle protein and muscle shear force (Salem et al. 2006b). To further investigate the potential contribution of DE genes in sexual maturation associated muscle atrophy, we investigated expression pattern of DE genes over 4 months during pre-spawning (July, November) and spawning season (December and January). Transcript abundance of ubiquitin-proteasome system genes and autophagy-related proteolytic genes remained constant in July and November, sharply increased in December and then declined in January (Figure 3, A-B). This trend in expression level of proteases nicely positively correlated with the severity of muscle atrophy during the time points. Late December represents the time with peak level of sexual maturation associated muscle atrophy followed by in January. Expression level of genes coding for different atrogin-1 isoforms and cathepsin D also showed highest level of expression in December and then decline in

January (Figure 3, C-D). Like atrogin-1, cathepsin D is involved in sexual maturation associated muscle atrophy and shows increase in transcript level, without significant change in activity during atrophy (Salem et al. 2006b). Extracellular matrix protein genes and development related genes showed opposite expression trend (Figure 3, E-F) in consistent with the loss of muscle stiffness and development during atrophy. These findings suggest that DE genes identified in the study may serve as faithful candidate that drive sexual maturation associated muscle atrophy in fish.

#### ***Long non-coding RNAs (lncRNAs)***

Unlike mRNAs, none of the DE lncRNAs had homology to previously annotated genes with known role in muscle atrophy. Out of 1,198 DE lncRNAs, 237 and 10 lncRNAs had sequence homology with lncRNAs from Atlantic salmon and Zebra fish respectively (sequence identity: > 80%, E value: <  $E^{-10}$ , query cover: > 50 nucleotides) (data not shown), but their functional annotation was not available in any species. Like protein coding mRNA, expression level of DE lncRNAs nicely correlated with the severity of muscle atrophy during pre-spawning and spawning months. Transcript abundance of upregulated lncRNAs remained constant during pre-spawning months, but increased drastically in December and then declined in January (Figure 3, G). On the other hand, transcript abundance of downregulated lncRNAs showed opposite trend as expected (Figure 3, H). These findings suggest that expression level of these DE lncRNAs may be involved in sexual maturation associated muscle atrophy.



**Figure 3:** Transcript level of different classes of DE genes during pre-spawning and spawning months in skeletal muscle of diploid gravid fish: all ubiquitinating genes combined (A), all autophagy related genes combined (B), atrogin-1 isoforms (C), cathepsin D isoforms (D), all development related genes combined (E), all collagen and extracellular matrix related genes combined (F), all upregulated lncRNAs combined (G) and all downregulated lncRNAs combined (H). Note that expression level of each gene in gravid fish (2N) was normalized by expression level of respective gene in sterile fish (3N).

### ***MicroRNAs***

Of 28 DE microRNAs, differential expression of mir-1, mir-133 and mir-29 during mammalian muscle atrophy has been reported previously (Wang 2013, Georgantas et al. 2014), but the rest of the DE microRNAs were identified for the first time. Though only 17 out of 665 predicted target genes were DE, all major pathways that showed significant alteration during muscle atrophy in this study, were well represented in the target gene list. Sixteen downregulated microRNAs targeted a total of 206 different protein coding genes. Consistent with the downregulation of microRNAs in atrophying muscle, 26 of the target genes were proteolytic enzymes including the genes involved in ubiquitin-proteasome and autophagy-lysosome mediated proteolysis (Appendix C). Twelve upregulated microRNAs targeted 468 different protein coding genes. In consistence with their upregulated expression during atrophy, 101 target genes were directly involved in extracellular matrix, muscle structure or development (data not shown). Above findings suggest that some genes involved in muscle atrophy may not be necessarily regulated at transcription level, and its fate is determined at post-transcription by regulated expression of microRNA.

Let-7j was the most highly downregulated microRNA (-1056× fold) in atrophying muscle (Table 1). It targeted 63 different protein coding genes which had wide range of functions (data not shown). Consistent with its downregulation in atrophying muscle, some of its targets were proteolytic genes such as e3 ubiquitin-protein ligase nrdp1, ubiquitin-conjugating enzyme e2 and protein vprbp. Different isoforms of mir-29 were highly upregulated in atrophying muscle (Table 1). Mir-29a, the most highly upregulated microRNA in atrophying muscle (11.7× fold) targeted 78 genes; the highest number of predicted target genes among DE microRNAs. In consistent with its upregulation in

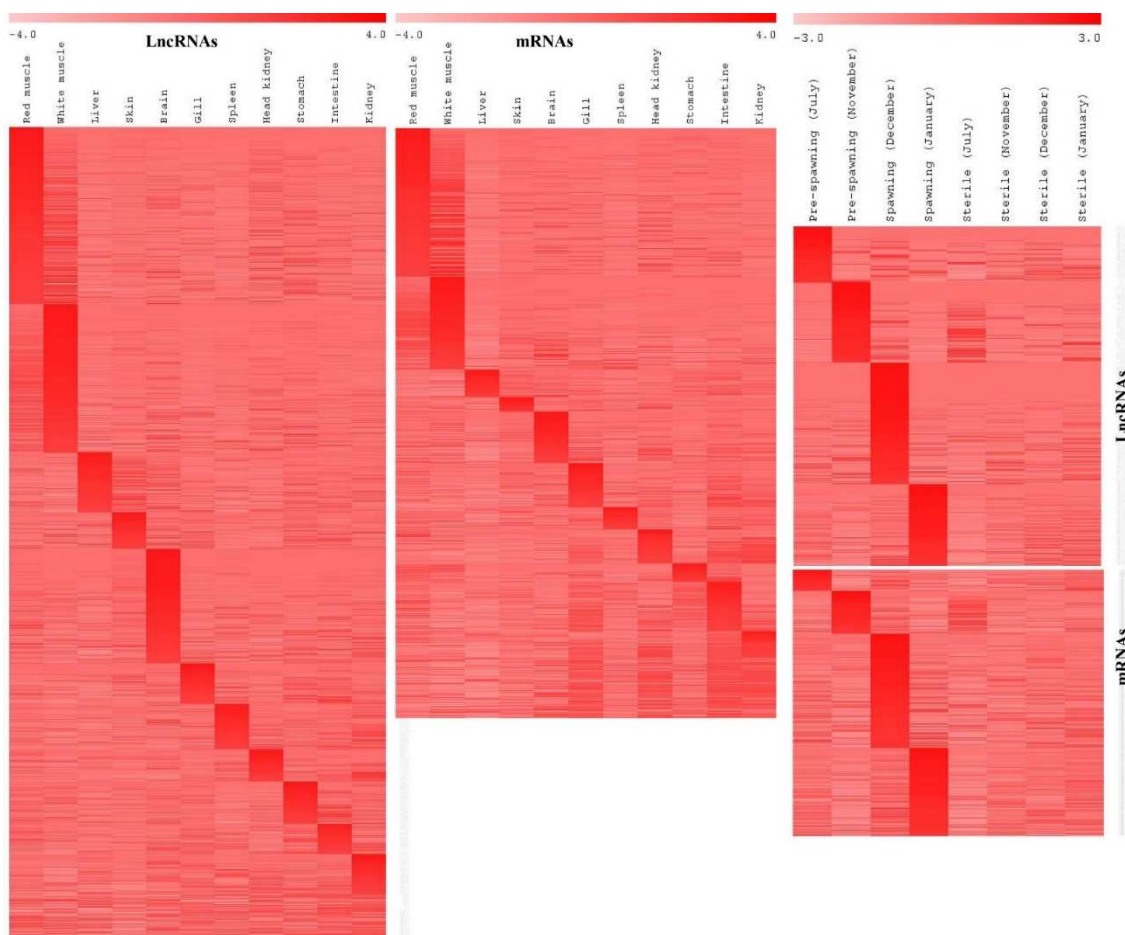
atrophy muscle, its targets included genes involved in muscle sarcomere (e. g. myosin), extracellular matrix (e. g. collagen), fat biosynthesis (e. g. long-chain-fatty-acid-- ligase acsbg2 and acyl-coenzyme a thioesterase 11), protein synthesis (e. g. 60s ribosomal protein 17) and development (e. g. prospero homeobox protein 1). These findings suggest that DE microRNAs may contribute to the muscle atrophy by regulating various proteolytic and other genes during muscle atrophy.

### **DE lncRNAs and mRNAs show spatio-temporal expression pattern**

LncRNAs show strict spatial (tissue specific) and temporal (time dependent) expression pattern (Pauli et al. 2012). To investigate tissue specificity of DE genes, we studied their expression pattern across 13 vital tissues including red and white muscle (see method section for classification of tissue specific genes). About 40% (475/1,198) of DE lncRNAs, and ~41% (348/852) of DE mRNAs were ‘specific’ to red or white muscle tissues (Figure 4). This indicated about 2.5×-fold enrichment of muscle specific lncRNAs, and about 5.5× fold enrichment of muscle specific mRNAs in the DE gene list as only ~16% (8,498/51,644) of non-DE lncRNAs and ~6% (4,583/61,412) of non-DE mRNAs in trout genome show muscle ‘specific’ expression (data not shown). Interestingly, majority of the most highly upregulated mRNAs were muscle ‘specific’. As an example, 47 out of 61 mRNAs with fold change greater than 15, had muscle restricted expression pattern, which included all 4 mRNAs encoding atrogen-1 isoforms. Muscle specific expression of atrogen-1 has also been reported previously in mammals (Gomes et al. 2001). In addition to muscle specific protein coding genes, some muscle specific lncRNAs such as linc-MD1 play important role in regulation of muscle specific genes (Cesana et al. 2011). In addition to tissue specific expression, ~37% (442/1,198) of DE lncRNAs and ~40% (342/852) of

DE mRNAs were 'specific' to one particular month during pre-spawning and spawning season in gravid fish, with no significant alteration in expression level in the absence of muscle atrophy in sterile fish (Figure 4). Above findings suggest that significant proportion of DE transcriptome in atrophying muscle is contributed by muscle specific genes expression, that show temporal expression pattern in response to the extent of muscle atrophy during sexual maturation.





**Figure 4:** Heat map showing tissue specific expression pattern of DE lncRNAs (left), tissue specific expression pattern of DE mRNAs (middle) and temporal expression pattern of DE lncRNAs and DE mRNAs during pre-spawning and spawning months (right). Value of color limit represents normalized expression values (Z scores). Darker red and lighter red colors represent higher and lower level of expression respectively.

#### **DE lncRNA and mRNA genes co-localize in genome and correlate in expression**

Functionally related lncRNAs and mRNA physically co-localize in genome (Paneru et al. 2016). Out of 1,198 DE lncRNAs, 235 (~20%) lncRNAs were either overlapped or neighbored (< 50 KB) by DE mRNA genes. There were a total of 246

(~29%) DE mRNA genes that were overlapping or neighboring to DE lncRNAs (Appendix D and E). These findings suggest that DE lncRNAs and DE mRNAs tend to co-localize or cluster together in genome. However, mere physical proximity does not necessarily lead to functional links (Cabili et al. 2011, Guttman et al. 2011). To test the functional significance of physical proximity, we computed expression correlation between all neighboring/overlapping lncRNA-mRNA genes. Out of total 380 neighboring/overlapping lncRNA-mRNA, ~37.4% (142) lncRNA-mRNA had strongly correlated expression pattern ( $R > 0.84$ ). On the other hand, out of 1,020,696 all possible DE lncRNA-mRNA combinations (1,198 DE lncRNA x 852 DE mRNA = 1,020,696 combinations) regardless of genomic co-localization, only ~16.7% (168,207) lncRNA-mRNA had strong expression correlation ( $R > 0.84$ ). The difference was statistically significant (Chi square p value  $< 0.001$ ) suggesting that co-localized lncRNA and protein coding genes tend to be more frequently correlated in expression than genes that are far in the genome (Appendix D and E). However, the degree of expression correlation between lncRNA and protein coding genes as a function of physical proximity, although statistically significant, was negative and weak ( $R = -0.35$ , p value  $< 0.001$ ). Next, we investigated if ‘strand orientation’ (sense or antisense) had correlation with ‘type of expression correlation’ (negative or positive), and found no significant correlation (Chi square p value  $> 0.05$ ). These findings suggest that DE lncRNA and DE mRNA gene tend co-localize or cluster together in genome, and often show correlated expression pattern.

### **DE lncRNAs potentially sponge or generate microRNAs**

‘LncRNA-microRNA’ binding has important functional consequences in microRNA mediated gene regulation as it sponges cellular microRNAs and causes

lncRNA-microRNA competition for mRNA binding (Yoon et al. 2014). In order to identify DE lncRNAs that potentially compromise microRNA mediated gene regulation, we searched for high confidence microRNA binding sites on DE lncRNAs. About 29% (350/1,198) of DE lncRNAs had binding sites for 314 of trout microRNAs including 4 DE microRNAs; mir-29, mir-133, mir-132 and let-7. Some of these microRNAs mir-133, mir-23, mir-29, mir-214, mir-221, mir-21, mir-19 and mir-199 are known to regulate muscle atrophy or proteolysis (Eisenberg et al. 2007, Ardite et al. 2012). Out of 314 microRNAs with potential binding sites on DE lncRNAs, 134 microRNAs also had binding sites on 154 of the 852 DE mRNAs that included mRNA encoding atrogen-1, cathepsins, serine proteases and several enzymes of ubiquitin proteasome system (Table 3). Atrogen-1 (GSONMT00016768001), the most highly upregulated genes in atrophying muscle, had multiple binding sites in 3' UTR for mir-22-3p. Two upregulated lncRNAs, Omy200063021 and TCONS\_00145202 also had high confidence binding site for the same microRNA (Table 4). Sharing of microRNA binding site between DE lncRNAs and DE mRNAs may have important functional consequences including lncRNA-mRNA competition for microRNA binding and sequestration of cellular microRNA by lncRNAs (Yoon et al. 2014). Next we asked if these functional interactions potentially exist in cell, lncRNA and mRNA that share microRNA binding sites should have correlated expression pattern. As expected, lncRNAs and mRNAs that share microRNA binding sites had strongly correlated expression pattern (Table 3). Interestingly, strong correlations ( $R > 0.84$ ) were significantly more frequent between microRNA binding site sharing lncRNA-mRNA pairs than other lncRNA-mRNA pairs that don't share microRNA binding site (Chi square p value:  $< 0.001$ ). This finding suggests that lncRNA and mRNA potentially interact

via common microRNA binding that may lead to correlated expression. Alternatively, correlated expression may form the basis of real lncRNA-mRNA competition to bind a common microRNA. In consistent with this argument, majority (~98%) of the correlations between lncRNA-mRNA pairs were positive suggesting potential requirement of lncRNA/mRNA co-expression for competition to exist. In addition, positive correlation may result from stabilization of mRNA by lncRNA from microRNA mediated decay. However, as translation suppression is the major mechanism of microRNA mediated gene regulation in animal (Selbach et al. 2008), 'lncRNA-protein level' correlation studies warrant revealing the true biological significance of this lncRNA-microRNA-mRNA interaction on the fate of mRNA. LncRNAs serve as precursors of microRNAs and other classes of small non-coding RNA (sRNAs) (Consortium 2012, Pauli et al. 2012). Out of 1,198 DE lncRNAs, 2 lncRNAs harbored microRNA loci within them. TCONS\_00148395 harbored mir-27 loci and Omy200105075 harbored microRNA aly-miR-398c-like loci. Both mir-27 and aly-miR-398c-like positively correlated in expression with their potential host lncRNAs with correlation (R) of 0.60 and 0.81 respectively. This finding suggests that these microRNAs could be generated by post-transcription processing of lncRNA transcripts. Though these microRNAs were not DE in atrophying muscle, microRNA mir-27, potentially generated from highly upregulated lncRNA TCONS\_00148395 (76× fold), is known to suppress adipogenesis (Lin et al. 2009), a pathway suppressed in atrophying muscle based on expression pattern of genes in present study. Interestingly, only DE predicted target gene of mir-27 was cyclic amp-dependent transcription factor atf-5, which is an important cofactor for adipogenesis (Zhao et al. 2014). This finding suggest that some

DE lncRNAs may contribute indirectly to the muscle atrophy by generating microRNAs capable of regulating various processes undergoing during atrophy.

**Table 3:** DE lncRNAs and mRNAs sharing microRNA binding sites and expression correlation between them.

MicroRNA Name	DE lncRNA with microRNA binding site	DE protein coding (mRNA) gene with microRNA binding site	mRNA-lncRNA correlation (R)
miR-22-3p	TCONS_00145202	GSONMT00016768001: f box only protein 32/atrogin-1	0.97
pma-miR-7a-3p like	TCONS_00148576	GSONMT00051340001: cysteine protease atg4b-like	0.91
bta-miR-7865 like	Omy200187283	GSONMT00005406001: ankyrin repeat and soxs box protein 2	0.98
miR-9404-5p	Omy100165236	GSONMT00043294001: kelch repeat and btb domain-containing protein 12-like	0.92
lin-4-5p	TCONS_00094218	GSONMT00029001001: class e basic helix-loop-helix protein 40-like	0.90
mmu-miR-29c-3p like	TCONS_00005193	GSONMT00013716001: myomegalin-like	0.94
hsa-miR-5007-5p like	TCONS_00004475	GSONMT00018163001: amp deaminase 3-like	0.93
mmi-miR-7189-3p like	TCONS_00068350	GSONMT00018181001: alanine aminotransferase 2-like	0.97
mmu-miR-3968 like	TCONS_00079982	GSONMT00018460001: large neutral amino acids transporter small subunit 4-like	0.94
pma-miR-192-3p like	Omy200145928	GSONMT00026025001: protein slowmo homolog 2-like	0.97
mir-221-3p	gill_00047762	GSONMT00032971001: phosphomannomutase 1-like	0.95
omy-miR-XXXX-5p	TCONS_00105663	GSONMT00038618001: endophilin-b1 isoform xl	0.93
omy-miR-XXXX-5p	TCONS_00029772	GSONMT00042478001: ring finger protein 122-like	0.99
mmu-miR-7074-3p like	gill_00051074	GSONMT00048379001: mitochondrial import receptor subunit tom22 homolog	0.90
cel-miR-1822-3p like	TCONS_00148690	GSONMT00049456001: sestrin-1-like isoform xl	0.91
eca-miR-9140 like	TCONS_00011543	GSONMT00050732001: sestrin-1-like isoform xl	0.99
omy-miR-XXXX-3p	gill_00023434	GSONMT00054562001: cation transport regulator-like protein 1-like	0.94
pma-miR-192-3p like	TCONS_00093963	GSONMT00060243001: low quality protein: ethanolaminephosphotransferase 1-like	0.93
pma-miR-192-3p like	TCONS_00093963	GSONMT00060255001: atp-dependent rna helicase an3-like isoform xl	0.95
cfa-miR-8844 like	gill_00080908	GSONMT00062364001: fatty acid synthase	0.96
miR-877-3p	TCONS_00008946	GSONMT00062643001: large neutral amino acids transporter small subunit 4-like	1.00
pma-miR-4543 like	TCONS_00075648	GSONMT00063472001: ras-related protein rab-7a	0.94
omy-miR-XXXX-5p	TCONS_00105663	GSONMT00065167001: cgmp-dependent protein kinase 1-like isoform xl	0.92
aly-miR-4235 like	TCONS_00105663	GSONMT00065334001: dual specificity protein phosphatase 22-b-like	0.99
eca-miR-8977 like	TCONS_00005364	GSONMT00065439001: ras-related protein rab-7a	0.90
gga-miR-1653 like	gill_00062140	GSONMT00068219001: intermediate filament family orphan 2	0.91
pma-miR-7a-3p like	Omy100004693	GSONMT00068926001: stromal interaction molecule 1-like	0.93
bta-miR-7865 like	TCONS_00012873	GSONMT00070501001: growth hormone receptor isoform l precursor	0.91
hsa-miR-5007-5p like	TCONS_00181081	GSONMT00070874001: insulin-induced gene 1	0.99
miR-XXXX-3p	TCONS_00004265	GSONMT00072008001: ras-related protein rab-5a-like	0.95
miR-XXXX-5p	TCONS_00056721	GSONMT00074474001: translocon-associated protein subunit delta precursor	0.92
cfa-miR-8844 like	TCONS_00016065	GSONMT00079999001: calcium-binding and coiled-coil domain-containing protein 1-like	0.98
omy-miR-XXXX-3p	Omy200142397	GSONMT00021507001: mitogen-activated protein kinase kinase kinase mlt-like	-0.88
omy-miR-XXXX-3p	TCONS_00004475	GSONMT00007327001: wd repeat-containing protein 55	-0.84

### **DE lncRNA may physically interact with DE protein coding genes at transcript and protein level**

Direct 'lncRNA-mRNA' physical interactions play crucial role in protein coding genes regulation as it leads to mRNA degradation (Gong and Maquat 2011) and translation inhibition (Yoon et al. 2012). To investigate the existence of such interactions between DE lncRNAs and DE mRNAs, we predicted their physical interactions. Strikingly, at interaction energy threshold  $< -100$  Kcal/mole, 1,151 potential physical interactions existed between DE lncRNA and DE mRNAs (Table 4). Some of these physical interactions appeared to be so strong that there was up to 150 'DE lncRNA-DE mRNA' hybrid length with near perfect complementarity. Strong expression correlations ( $R > 0.95$ ) were strikingly more frequent among DE lncRNA-DE mRNA showing evidence of direct physical interaction than among DE lncRNA-DE mRNA without the evidence of direct physical interactions (Chi square p value  $< 0.001$ ). These findings suggest that functional consequences of lncRNA-mRNA physical interaction may lead to correlated expression. As shown in table 5, DE lncRNAs appeared to physically interact with almost all major proteolytic genes including atrogin-1, cathepsins and many ubiquitinating enzymes. This finding suggests that lncRNA-mRNA physical interaction is most likely to regulate skeletal muscle atrophy.

In addition to interacting with mRNA transcript, lncRNA also physically interact with proteins, which play crucial role in gene regulation (Rinn et al. 2007, Tripathi et al. 2010). In order to investigate such interactions during muscle atrophy, we predicted direct physical interaction of DE lncRNAs with a proteome of DE protein coding genes. A total of about 1,000 physical interactions existed between DE lncRNA and proteome of DE

protein coding genes at interaction threshold: interaction strength  $\geq 96\%$  and discriminative power  $\geq 96\%$  (Table 4). ‘DE lncRNA-DE proteins’ showing evidence of direct physical interactions showed more frequent expression correlation (at transcript level) than other ‘DE lncRNA-DE protein coding gene’ without the evidence of direct physical interaction (Chi square p value  $< 0.001$ ). Several proteolytic proteins including atrogen-1, cathepsin, and several enzymes of ubiquitin-proteasome system showed evidence of physical interaction with different DE lncRNAs. In addition to proteolytic proteins, 28 (out of 53) DE transcription factors (TFs) showed evidence of physical interactions with DE lncRNAs. LncRNA’s physical interaction with protein modulates stability (Taniue et al. 2016), availability (sequestration) (Hirose et al. 2014), covalent modification/activity (Tripathi et al. 2010) and proper cellular localization (Tripathi et al. 2010) of the protein. Physical interaction of DE lncRNAs with DE proteolytic genes may play important role to determine the fate of DE proteins and hence muscle atrophy.

**Table 4:** Differentially expressed (DE) lncRNA-DE mRNA physical interaction statistics and expression correlation between them (top). Physical interaction statistics between DE lncRNA and proteome of DE protein coding genes (bottom).

DE lncRNA	DE mRNA	lncRNA-mRNA hybrid length (nts)	Interaction energy (Kcal/mole)	Expression correlation (R)
gill_00018678	f-box protein 32/atrogin-1	149	-229.635	0.96
TCONS_00015745	f-box only protein 32/atrogin-1	147	-144.674	0.95
TCONS_00044636	cathepsin D	147	-116.681	0.97
TCONS_00071240	ubiquitin carboxyl-terminal hydrolase 25-like	149	-170.963	0.95
gill_00080545	ubiquitin carboxyl-terminal hydrolase 10	146	-191.917	0.92
gill_00030058	collagen alpha-1 chain like	149	-240.644	0.94
TCONS_00028182	myosin-binding protein slow-type-like	128	-107.449	0.91
TCONS_00055397	ATP-dependent 6-phosphofructokinase	149	-148.776	0.91
gill_00034918	dnaJ homolog subfamily B member 1-like	150	-206.625	0.90
gill_00043112	serine threonine-protein kinase ulk2-like	149	-166.14	0.99
DE lncRNA	DE protein	Interaction strength (%)	Discriminative power (%)	Expression correlation (R)
TCONS_00034255	f-box only protein 32/atrogin-1	100	96	0.75
Omy100083321	cathepsin L1	99	97	0.99
TCONS_00034255	cathepsin z precursor	99	97	0.97
gill_00058188	ubiquitin carboxyl-terminal hydrolase 10	100	100	0.94
Omy100109323	e3 ubiquitin-protein ligase trim63-like	98	97	0.99
TCONS_00181081	collagen alpha-1 chain-like	100	99	0.96
gill_00048471	autophagy-related protein 9a-like	99	98	0.95
TCONS_00025350	cyclic amp-dependent transcription factor atf-5	99	96	0.97
Omy200129177	kelch-like protein 38-like	100	100	0.97
TCONS_00015874	camp-responsive element modulator isoform	100	96	0.95
TCONS_00044055	ccaat enhancer-binding protein delta	100	96	0.87

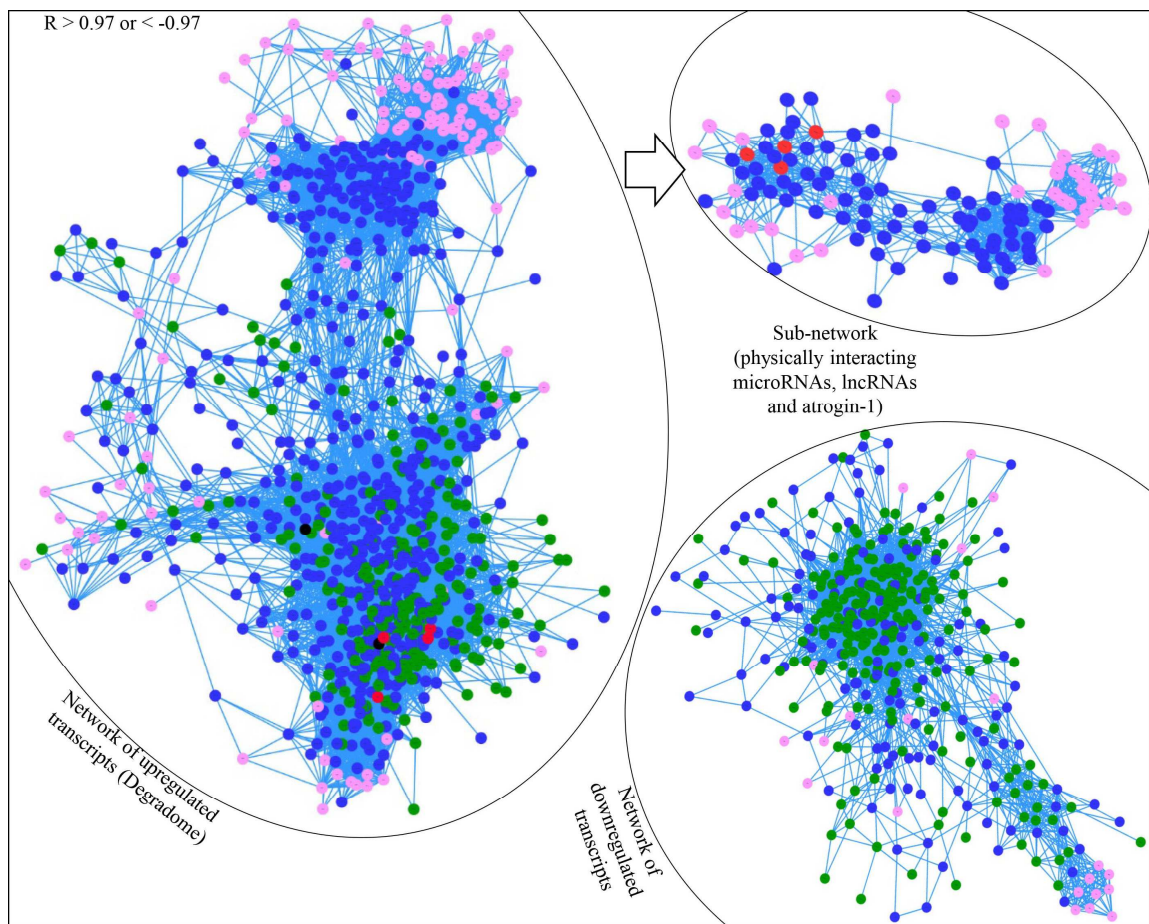
### DE lncRNAs, mRNAs and microRNAs comprise a muscle “degradome”

LncRNAs, microRNAs and mRNAs comprise interacted gene regulatory networks in cell (Jalali et al. 2013), probably due to regulation of microRNA and lncRNAs by each other (Yoon et al. 2014). To investigate the functional interplay among coding and non-coding RNAs during skeletal muscle atrophy, we computed lncRNA-mRNA-microRNA



interaction networks based on their expression pattern across 30 samples. At correlation threshold  $R > 0.97$  or  $< -0.97$ , about 50% (1,584) of transcripts comprised strongly correlated gene networks (Figure 5) suggesting strong co-expression pattern among DE transcripts. Interestingly, majority of the correlated transcripts clustered in one of two major networks; the first comprised of downregulated transcripts and the second comprised of upregulated transcripts. The first network consisted of 430 transcripts (137 lncRNAs, 219 mRNAs and 74 microRNAs). The second network consisted of 960 transcripts (559 lncRNAs, 235 mRNAs and 166 microRNAs). The second network appeared to be interacted gene “reactome” regulating muscle proteolysis as almost all upregulated proteolytic genes including enzymes of ubiquitin proteasome system, autophagy related proteolytic genes and other proteases such as cathepsins were in the network. Similarly, majority of the upregulated microRNAs including mir-29, let-7 and mir-132 were in the network. Interestingly, 4 atrogin-1 transcripts (the most highly upregulated transcripts) and mir-29a (the most highly upregulated microRNA) were in the center of the network suggesting that they may be one of the key members in the network. Central position of the atrogin-1 in the network is also supported by previous studies that report atrogin-1 as a key regulator of muscle atrophy (Gomes et al. 2001, Cleveland and Evenhuis 2010, Tacchi et al. 2010). We named this network as ‘the Rainbow trout muscle degradome’ as it appeared to be a regulatory network of muscle degrading genes. The network was largely comprised of the genes that showed evidence of physical interaction with each other. A sub-network of degradome shown in figure 5 consist of 4 atrogin-1 transcripts, all DE lncRNAs that bind to atrogin-1, and microRNAs that either bind to lncRNAs and/or atrogin-1 mRNAs. This finding suggests that DE lncRNA, protein coding genes and

microRNAs interact with each other by multiple mechanisms which may lead to highly correlated gene regulatory 'lncRNA-mRNA-microRNA' network. Above findings also suggests that genes likely involved in muscle atrophy may be induced simultaneously in a regulated fashion and work as a highly correlated complex gene network.



**Figure 5:** Gene expression network of DE lncRNAs (blue node), DE mRNAs (green node) and microRNAs (pink node) ( $R > 0.97$  or  $< -0.97$ ). Note that the majority of the DE genes are clustered in one of the two major networks. The larger network (degradome) comprises of upregulated genes and smaller network comprises of downregulated genes. In the network of upregulated transcripts, 4 atrogen-1 transcripts (red nodes) and mir-29 isoforms (black nodes) are in the center of the network. Sub-network drawn from the larger network contains 4 atrogen-1 transcripts, lncRNAs that bind to atrogen-1, and microRNAs that either bind to the lncRNAs and/or atrogen-1. Note: edges that connect nodes (genes) represent correlated expression at  $R$  cut off  $0.97 >$  or  $< -0.97$ ; the shorter the length, the stronger the expression correlation.

## CONCLUSION

Sexual maturation associated skeletal muscle atrophy serves as an excellent model to study piscine muscle proteolysis (Salem et al. 2006a, Salem et al. 2010, Salem et al. 2013). Previous efforts of investigating fish muscle proteolysis have provided limited information as these studies relied on individual or limited set of protein coding genes (Salem et al. 2006a, Cleveland and Evenhuis 2010, Tacchi et al. 2010). In present study, we used deep lncRNA, mRNA and microRNA sequencing approach to investigate genes and gene regulatory network that regulate muscle proteolysis in fish. By thorough investigation of atrophying muscle transcriptome, we elucidated that the fish muscle atrophy, like mammalian counterpart, is regulated mainly by ubiquitin-proteasome system. In addition, large number of autophagy-lysosomal proteases and transcription factors appeared to take part in bulk muscle proteolysis during atrophy. Atrophying muscle showed upregulation of proteolytic genes concomitant with downregulation of genes involved in muscle sarcomere, extracellular matrix, protein and fat biosynthesis, and development. This trend in expression pattern of genes in atrophying muscle nicely correlated with the measured muscle phenotypes (e. g. muscle mass, protein content and muscle shear force) of atrophying muscle suggesting essential role of DE genes in muscle atrophy. Present study identified large number of new candidate coding and non-coding genes in addition to the genes identified by previous microarray and proteomic approaches (Salem et al. 2006a, Salem et al. 2010) suggesting that the sequencing approach has identified vast majority of faithful candidate genes likely involved in muscle proteolysis.

In present study, we characterized lncRNAs potentially involved in fish muscle proteolysis, and investigated lncRNA-mRNA, lncRNA-microRNA and mRNA-

microRNA interactions that potentially regulate muscle atrophy. Majority of DE lncRNAs co-localized or clustered with DE protein coding genes in genome. DE lncRNAs appeared to extensively physically interact with DE protein coding genes at its transcript and protein level. Similarly, DE lncRNA also showed potential to bind and sequester cellular microRNAs implicated in muscle proteolysis. LncRNA, mRNA and microRNAs that showed evidence of above interactions comprised a highly correlated gene network in atrophying muscle. This finding indicates that the vast majority of the genes involved in muscle proteolysis are expressed simultaneously by a common gene transcription program. Important to mention, 'DE lncRNA-DE protein coding' gene pairs that either co-localized in genome or showed evidence of direct physical interaction or competition for a common microRNA binding, were more frequently correlated in expression than random 'DE lncRNA-DE protein coding' gene pairs. This is perhaps the first genome wide study that provided links between expression correlation and potential functional interactions between lncRNA and mRNA in genome wide scale in any species. The present study has investigated potential coding and non-coding RNA interactions during muscle atrophy and will help understand how energetic demand of sexual maturation triggers skeletal muscle atrophy in fish.

## REFERENCES

- Agostini, F., A. Zanzoni, P. Klus, D. Marchese, D. Cirillo & G. G. Tartaglia (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, 29, 2928-30.
- Al-Tobasei, R., B. Paneru & M. Salem (2016) Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout. *PLoS One*, 11, e0148940.
- Ardite, E., E. Perdiguero, B. Vidal, S. Gutarra, A. L. Serrano & P. Muñoz-Cánoves (2012) PAI-1-regulated miR-21 defines a novel age-associated fibrogenic pathway in muscular dystrophy. *J Cell Biol*, 196, 163-75.
- Aussanasuwannakul, A., G. M. Weber, M. Salem, J. Yao, S. Slider, M. L. Manor & P. B. Kenney (2012) Effect of sexual maturation on thermal stability, viscoelastic properties, and texture of female rainbow trout, *Oncorhynchus mykiss*, filets. *J Food Sci*, 77, S77-83.
- Bagga, S., J. Bracht, S. Hunter, K. Massirer, J. Holtz, R. Eachus & A. E. Pasquinelli (2005) Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*, 122, 553-63.
- Berthelot, C., F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noel, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J. M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J. N. Volff, C. Genet, P. Wincker, O. Jaillon, H. Roest Crolius & Y. Guiguen (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*, 5, 3657.
- Bolstad, B. M., R. A. Irizarry, M. Astrand & T. P. Speed (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19, 185-93.
- Busch, A., A. S. Richter & R. Backofen (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24, 2849-56.
- Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev & J. L. Rinn (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 25, 1915-27.
- Cesana, M., D. Cacchiarelli, I. Legnini, T. Santini, O. Sthandier, M. Chinappi, A. Tramontano & I. Bozzoni (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147, 358-69.
- Cleveland, B. M. & J. P. Evenhuis (2010) Molecular characterization of atrogin-1/F-box protein-32 (FBXO32) and F-box protein-25 (FBXO25) in rainbow trout (*Oncorhynchus mykiss*): Expression across tissues in response to feed deprivation. *Comp Biochem Physiol B Biochem Mol Biol*, 157, 248-57.
- Consortium, E. P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57-74.
- Eisenberg, I., A. Eran, I. Nishino, M. Moggio, C. Lamperti, A. A. Amato, H. G. Lidov, P. B. Kang, K. N. North, S. Mitrani-Rosenbaum, K. M. Flanigan, L. A. Neely, D. Whitney, A. H. Beggs, I. S. Kohane & L. M. Kunkel (2007) Distinctive patterns of

- microRNA expression in primary muscular disorders. *Proc Natl Acad Sci U S A*, 104, 17016-21.
- Georgantas, R. W., K. Streicher, S. A. Greenberg, L. M. Greenlees, W. Zhu, P. Z. Brohawn, B. W. Higgs, M. Czapiga, C. A. Morehouse, A. Amato, L. Richman, B. Jallal, Y. Yao & K. Ranade (2014) Inhibition of myogenic microRNAs 1, 133, and 206 by inflammatory cytokines links inflammation and muscle degeneration in adult inflammatory myopathies. *Arthritis Rheumatol*, 66, 1022-33.
- Gomes, M. D., S. H. Lecker, R. T. Jagoe, A. Navon & A. L. Goldberg (2001) Atrogin-1, a muscle-specific F-box protein highly expressed during muscle atrophy. *Proc Natl Acad Sci U S A*, 98, 14440-5.
- Gong, C. & L. E. Maquat (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, 470, 284-8.
- Gonzalez-Pena, D., G. Gao, M. Baranski, T. Moen, B. M. Cleveland, P. B. Kenney, R. L. Vallejo, Y. Palti & T. D. Leeds (2016) Genome-Wide Association Study for Identifying Loci that Affect Fillet Yield, Carcass, and Body Weight Traits in Rainbow Trout (*Oncorhynchus mykiss*). *Front Genet*, 7, 203.
- Guttman, M., J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root & E. S. Lander (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 477, 295-300.
- Han, X., F. Yang, H. Cao & Z. Liang (2015) Malat1 regulates serum response factor through miR-133 as a competing endogenous RNA in myogenesis. *FASEB J*, 29, 3054-64.
- Hirose, T., G. Virnicchi, A. Tanigawa, T. Naganuma, R. Li, H. Kimura, T. Yokoi, S. Nakagawa, M. Bénard, A. H. Fox & G. Pierron (2014) NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Mol Biol Cell*, 25, 169-83.
- Hudson, M. B., J. A. Rahnert, B. Zheng, M. E. Woodworth-Hobbs, H. A. Franch & S. R. Price (2014) miR-182 attenuates atrophy-related gene expression by targeting FoxO3 in skeletal muscle. *Am J Physiol Cell Physiol*, 307, C314-9.
- Jalali, S., D. Bhartiya, M. K. Lalwani, S. Sivasubbu & V. Scaria (2013) Systematic transcriptome wide analysis of lncRNA-miRNA interactions. *PLoS One*, 8, e53823.
- Juanchich, A., P. Bardou, O. Rué, J. C. Gabillard, C. Gaspin, J. Bobe & Y. Guiguen (2016) Characterization of an extensive rainbow trout miRNA transcriptome by next generation sequencing. *BMC Genomics*, 17, 164.
- Kallen, A. N., X. B. Zhou, J. Xu, C. Qiao, J. Ma, L. Yan, L. Lu, C. Liu, J. S. Yi, H. Zhang, W. Min, A. M. Bennett, R. I. Gregory, Y. Ding & Y. Huang (2013) The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Mol Cell*, 52, 101-12.
- Krek, A., D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel & N. Rajewsky (2005) Combinatorial microRNA target predictions. *Nat Genet*, 37, 495-500.
- Kukreti, H., K. Amuthavalli, A. Harikumar, S. Sathiyamoorthy, P. Z. Feng, R. Anantharaj, S. L. Tan, S. Lokireddy, S. Bonala, S. Sriram, C. McFarlane, R. Kambadur & M. Sharma (2013) Muscle-specific microRNA1 (miR1) targets heat shock protein 70 (HSP70) during dexamethasone-mediated atrophy. *J Biol Chem*, 288, 6663-78.

- Leeds, T. D., R. L. Vallejo, G. M. Weber, D. G. Pena & J. S. Silverstein (2016) Response to five generations of selection for growth performance traits in rainbow trout (*Oncorhynchus mykiss*). *Aquaculture*, 465, 341-351.
- Li, J. J., H. Huang, P. J. Bickel & S. E. Brenner (2014) Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res*, 24, 1086-101.
- Li, N., M. Ponnusamy, M. P. Li, K. Wang & P. F. Li (2016) The Role of MicroRNA and LncRNA-MicroRNA Interactions in Regulating Ischemic Heart Disease. *J Cardiovasc Pharmacol Ther.*
- Lin, Q., Z. Gao, R. M. Alarcon, J. Ye & Z. Yun (2009) A role of miR-27 in the regulation of adipogenesis. *FEBS J*, 276, 2348-58.
- Llano-Diez, M., A. M. Gustafson, C. Olsson, H. Goransson & L. Larsson (2011) Muscle wasting and the temporal gene expression pattern in a novel rat intensive care unit model. *BMC Genomics*, 12, 602.
- Lopes, C. T., M. Franz, F. Kazi, S. L. Donaldson, Q. Morris & G. D. Bader (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26, 2347-8.
- Ma, M. Z., B. F. Chu, Y. Zhang, M. Z. Weng, Y. Y. Qin, W. Gong & Z. W. Quan (2015) Long non-coding RNA CCAT1 promotes gallbladder cancer development via negative modulation of miRNA-218-5p. *Cell Death Dis*, 6, e1583.
- Marancik, D., G. Gao, B. Paneru, H. Ma, A. G. Hernandez, M. Salem, J. Yao, Y. Palti & G. D. Wiens (2014) Whole-body transcriptome of selectively bred, resistant-, control-, and susceptible-line rainbow trout following experimental challenge with *Flavobacterium psychrophilum*. *Front Genet*, 5, 453.
- Olsen, P. H. & V. Ambros (1999) The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol*, 216, 671-80.
- Paneru, B., R. Al-Tobasei, Y. Palti, G. D. Wiens & M. Salem (2016) Differential expression of long non-coding RNAs in three genetic lines of rainbow trout in response to infection with *Flavobacterium psychrophilum*. *Sci Rep*, 6, 36032.
- Pauli, A., E. Valen, M. F. Lin, M. Garber, N. L. Vastenhouw, J. Z. Levin, L. Fan, A. Sandelin, J. L. Rinn, A. Regev & A. F. Schier (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*, 22, 577-91.
- Ramachandra, R. K., M. Salem, S. Gahr, C. E. Rexroad & J. Yao (2008) Cloning and characterization of microRNAs from rainbow trout (*Oncorhynchus mykiss*): their expression during early embryonic development. *BMC Dev Biol*, 8, 41.
- Rinn, J. L. & H. Y. Chang (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem*, 81, 145-66.
- Rinn, J. L., M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal & H. Y. Chang (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129, 1311-23.



- Rueda, A., G. Barturen, R. Lebrón, C. Gómez-Martín, Á. Alganza, J. L. Oliver & M. Hackenberg (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res*, 43, W467-73.
- Salem, M., P. B. Kenney, C. E. Rexroad & J. Yao (2006a) Microarray gene expression analysis in atrophying rainbow trout muscle: a unique nonmammalian muscle degradation model. *Physiol Genomics*, 28, 33-45.
- (2006b) Molecular characterization of muscle atrophy and proteolysis associated with spawning in rainbow trout. *Comp Biochem Physiol Part D Genomics Proteomics*, 1, 227-37.
- (2010) Proteomic signature of muscle atrophy in rainbow trout. *J Proteomics*, 73, 778-89.
- Salem, M., M. L. Manor, A. Aussanasuwannakul, P. B. Kenney, G. M. Weber & J. Yao (2013) Effect of sexual maturation on muscle gene expression of rainbow trout: RNA-Seq approach. *Physiol Rep*, 1, e00120.
- Salem, M., B. Paneru, R. Al-Tobasei, F. Abdouni, G. H. Thorgaard, C. E. Rexroad & j. Yao (2015) Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout. *PLoS ONE*.
- Salem, M., J. Silverstein, C. E. Rexroad & J. Yao (2007) Effect of starvation on global gene expression and proteolysis in rainbow trout (*Oncorhynchus mykiss*). *BMC Genomics*, 8, 328.
- Selbach, M., B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin & N. Rajewsky (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455, 58-63.
- Tacchi, L., R. Bickerdike, C. J. Secombes, N. J. Pooley, K. L. Urquhart, B. Collet & S. A. Martin (2010) Ubiquitin E3 ligase atrogin-1 (Fbox-32) in Atlantic salmon (*Salmo salar*): sequence analysis, genomic structure and modulation of expression. *Comp Biochem Physiol B Biochem Mol Biol*, 157, 364-73.
- Taniue, K., A. Kurimoto, H. Sugimasa, E. Nasu, Y. Takeda, K. Iwasaki, T. Nagashima, M. Okada-Hatakeyama, M. Oyama, H. Kozuka-Hata, M. Hiyoshi, J. Kitayama, L. Negishi, Y. Kawasaki & T. Akiyama (2016) Long noncoding RNA UPAT promotes colon tumorigenesis by inhibiting degradation of UHRF1. *Proc Natl Acad Sci U S A*, 113, 1273-8.
- Tripathi, V., J. D. Ellis, Z. Shen, D. Y. Song, Q. Pan, A. T. Watt, S. M. Freier, C. F. Bennett, A. Sharma, P. A. Bubulya, B. J. Blencowe, S. G. Prasanth & K. V. Prasanth (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell*, 39, 925-38.
- Wang, J., X. Liu, H. Wu, P. Ni, Z. Gu, Y. Qiao, N. Chen, F. Sun & Q. Fan (2010) CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res*, 38, 5366-83.
- Wang, X. H. (2013) MicroRNA in myogenesis and muscle atrophy. *Curr Opin Clin Nutr Metab Care*, 16, 258-66.
- Wu, L., J. Fan & J. G. Belasco (2006) MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A*, 103, 4034-9.

- Xu, J., R. Li, B. Workeneh, Y. Dong, X. Wang & Z. Hu (2012) Transcription factor FoxO1, the dominant mediator of muscle wasting in chronic kidney disease, is inhibited by microRNA-486. *Kidney Int*, 82, 401-11.
- Yoon, J. H., K. Abdelmohsen & M. Gorospe (2014) Functional interactions among microRNAs and long noncoding RNAs. *Semin Cell Dev Biol*, 34, 9-14.
- Yoon, J. H., K. Abdelmohsen, S. Srikantan, X. Yang, J. L. Martindale, S. De, M. Huarte, M. Zhan, K. G. Becker & M. Gorospe (2012) LincRNA-p21 suppresses target mRNA translation. *Mol Cell*, 47, 648-55.
- Zhao, Y., Y. D. Zhang, Y. Y. Zhang, S. W. Qian, Z. C. Zhang, S. F. Li, L. Guo, Y. Liu, B. Wen, Q. Y. Lei, Q. Q. Tang & X. Li (2014) p300-dependent acetylation of activating transcription factor 5 enhances C/EBP $\beta$  transactivation of C/EBP $\alpha$  during 3T3-L1 differentiation. *Mol Cell Biol*, 34, 315-24.

**APPENDICES**

## APPENDIX A

**SELECTED DE GENES INVOLVED IN FAT/AMINO ACID BIOSYNTHESIS,  
AMINO ACID TRANSPORT/CATABOLISM, MUSCLE STRUCTURE AND  
MYOGENESIS. POSITIVE VALUE OF FOLD CHANGE INDICATES  
UPREGULATION AND NEGATIVE VALUE INDICATES  
DOWNREGULATION IN ATROPHYING MUSCLE. FOLD CHANGE WAS  
CONSIDERED SIGNIFICANT AT CUT OFF: 3> OR < -3, FDR-P-VALUE <0.01**

DE mRNA ID	DE mRNA name	Fold change	correction
<b>Fat and fatty acid biosynthesis</b>			
GSONMT00062364001	fatty acid synthase	-29.43	1.52624E-11
GSONMT00030942001	fatty acid synthase	-22.75	9.82781E-06
GSONMT00057481001	fatty acid synthase	-19.30	5.8159E-05
GSONMT00049946001	acetyl- carboxylase alpha	-17.00	4.84828E-07
GSONMT00008756001	Fatty acyl- desaturase-like	-10.20	9.28048E-07
GSONMT00062538001	pyrroline-5-carboxylate reductase 2-like isoform x1	-9.66	6.03535E-05
GSONMT00066615001	long-chain-fatty-acid- ligase 4-like	-7.11	7.44715E-05
GSONMT00079100001	phosphoserine phosphatase	-6.49	0.000516507
GSONMT00004169001	glycerol-3-phosphate acyltransferase mitochondrial	-6.22	2.0319E-08
GSONMT00075321001	diacylglycerol o-acyltransferase 2	-5.67	0.004296087
<b>Amino acid biosynthesis</b>			
TCONS_00069986	acidic amino acid decarboxylase GADL1-like	-34.21	1.91285E-14
TCONS_00121062	acidic amino acid decarboxylase GADL1-like	-25.06	1.01175E-10
TCONS_00095286	acidic amino acid decarboxylase GADL1-like	-15.83	8.98274E-11
TCONS_00051857	pyrroline-5-carboxylate reductase 1, mitochondrial-like	-11.10	3.59445E-11
GSONMT00062538001	pyrroline-5-carboxylate reductase 2-like isoform x1	-9.66	6.03535E-05
TCONS_00139838	pyrroline-5-carboxylate reductase 1, mitochondrial-like	-9.31	0.009522559
GSONMT00026184001	s-adenosylmethionine synthase isoform type-2	-8.96	1.20074E-10
<b>Amino acid transport and catabolism</b>			
GSONMT00055546001	large neutral amino acids transporter small subunit 2-like	110.57	2.51528E-09
GSONMT00023744001	large neutral amino acids transporter small subunit 4-like	45.38	3.38306E-10
GSONMT00033966001	large neutral amino acids transporter small subunit 4-like	42.40	1.03851E-09
GSONMT00009952001	large neutral amino acids transporter small subunit 2	36.52	1.03889E-05
GSONMT00062643001	large neutral amino acids transporter small subunit 4-like	15.97	2.43482E-07
GSONMT00018460001	large neutral amino acids transporter small subunit 4-like isoform x2	15.88	1.83876E-05
<b>Myogenic genes and structural components of skeletal muscle fiber and extracellular matrix</b>			
GSONMT00049811001	myosin heavy fast skeletal muscle- partial	-30.38	6.0668E-08
GSONMT00076512001	collagen alpha-1 chain-like	-24.73	0.000381791
GSONMT00073107001	immunoglobulin-like and fibronectin type iii domain-containing protein 1-like	-19.09	0.003187507
GSONMT00023975001	fras1-related extracellular matrix protein 2-like	-13.98	2.5078E-08
GSONMT00007676001	gla-rich protein	-12.52	1.59864E-11
GSONMT00000976001	collagen type i alpha 2	-11.47	3.42209E-06
GSONMT00057539001	myosin heavy chain	-11.40	1.05538E-05
GSONMT00065898001	troponin i	-9.95	0.000365555
GSONMT00070983001	myostatin	-6.17	0.000204828
GSONMT00026553001	protein unc-45 homolog b-like	-4.45	6.94585E-05
GSONMT00025527001	myoblast determination protein 2 (myoD2)	-3.24	0.008344459

**APPENDIX B**

**DIFFERENTIALLY EXPRESSED (DE) TRANSCRIPTION FACTORS AND/OR  
TRANSCRIPTION REGULATORS IN SKELETAL MUSCLE BETWEEN  
GRAVID AND STERILE FISH. POSITIVE VALUE OF FOLD CHANGE  
INDICATES UPREGULATION AND NEGATIVE VALUE INDICATES  
DOWNREGULATION IN ATROPHYING MUSCLE. FOLD CHANGE WAS  
CONSIDERED SIGNIFICANT AT CUT OFF: 3 > OR < -3, FDR-P-VALUE < 0.01**

<b>DE transcription factor</b>	<b>old chang</b>	<b>FDR-p-value</b>
protein sog3-like isoform x3	63.81	2.42588E-09
response gene to complement 32 protein	17.27	6.65591E-08
camp-responsive element modulator isoform x3	14.03	2.83053E-05
ccaat enhancer-binding protein delta	8.25	3.95985E-07
nuclear protein 1	7.88	1.9406E-08
nuclear factor erythroid 2-related factor 1	7.28	0.0013038
thioredoxin-interacting protein	6.98	2.58513E-05
ras-related protein rab-5a-like	5.68	0.002551483
doublesex and mab-3 related transcription factor 2b	5.03	0.007871657
dual specificity protein phosphatase 22-b-like	4.62	0.000355571
cyclin-dependent kinase inhibitor 1b	4.53	0.000842541
krueppel-like factor 9	4.50	1.71231E-05
camp-dependent protein kinase inhibitor gamma	-4.04	0.000696213
protein arginine n-methyltransferase 1-like isoform x1	-4.05	0.000219971
lumican precursor	-4.05	0.000312624
actin-like protein 6a	-4.18	0.004353264
nuclear receptor subfamily 4 group a member 1-like	-4.33	4.75119E-05
protein arginine n-methyltransferase 1-like isoform x1	-4.37	8.37719E-05
dnaj homolog subfamily c member 2 isoform x1	-4.62	0.001659932
cyclic amp-dependent transcription factor atf-5	-4.74	0.002028069
protein aflq	-4.85	0.000485396
dnaj homolog subfamily c member 2-like	-5.34	1.60145E-07
tribbles homolog 2	-6.43	1.81387E-05
cyclic amp-dependent transcription factor atf-5	-7.30	5.37973E-08
zinc finger protein 648-like	-7.92	3.2885E-07
ruvb-like 2-like isoform x1	-8.89	4.61953E-06
methylosome protein 50	-10.29	0.002132849
cold shock domain-containing protein c2-like	-17.74	0.002866634

## APPENDIX C

**SELECTED TARGETS OF DOWNREGULATED MICRORNAS THAT ARE  
INVOLVED IN MUSCLE PROTEOLYSIS**

<b>MicroRNA</b>	<b>Target gene</b>
mir-15b	peroxisomal leader peptide-processing protease
mir-7551	autophagy-related protein 101-like
mir-15b	ubiquitin-conjugating enzyme e2 variant 1
mir-1386	e3 ubiquitin-protein ligase znrf2-like
mir-15b	small ubiquitin-related modifier 1 precursor
mir-15b	e3 ubiquitin-protein ligase trim23
mir-1386	probable e3 ubiquitin-protein ligase rnf144a
mir-1386	e3 ubiquitin-protein ligase rnf180
mir-125b-1	e3 ubiquitin-protein ligase rnf8
mir-7641	e3 ubiquitin-protein ligase trim39-like
mir-15b	e3 ubiquitin-protein ligase march9-like
mir-125b-1	e3 ubiquitin isg15 ligase trim25-like
mir-7551	parafibromin
mir-15b	jnk1 mapk8-associated membrane protein isoform x1
mir-1386	endoplasmic reticulum lectin 1-like
mir-7551	proteasome activator complex subunit 4b-like isoform x2
mir-203b	pol polyprotein
mir-1386	serpin h1-like isoform x1
mir-125b-1	a disintegrin and metalloproteinase with thrombospondin motifs 8
mir-7551	matrix metalloproteinase-14
mir-1386	serpin h1-like isoform x1
mir-7641-1	frizzled- partial
mir-7551	microtubule-associated proteins 1a 1b light chain 3c-like
mir-15b	kelch-like protein diablo-like isoform x2
mir-7551	cullin-3 isoform 1
mir-1386	dipeptidyl peptidase 9-like

**APPENDIX D**

**SELECTED OVERLAPPING DIFFERENTIALLY EXPRESSED LNCRNA-  
PROTEIN CODING GENE PAIRS AND THEIR EXPRESSION CORRELATION**

<b>DE lncRNA</b>	<b>DE mRNA</b>	<b>Correlation (R)</b>
TCONS_00181081	GSONMT00070874001	0.99
gill_00052726	GSONMT00071779001	0.99
gill_00039399	GSONMT00023146001	0.98
TCONS_00175630	GSONMT00004158001	0.96
TCONS_00040794	GSONMT00033306001	0.96
gill_00007443	GSONMT00035250001	0.96
Omy200196248	GSONMT00081504001	0.93
Omy200162242	GSONMT00082573001	0.92
gill_00047660	GSONMT00016221001	0.89
gill_00023727	GSONMT00063187001	0.89
TCONS_00147638	GSONMT00070501001	0.89
Omy300076644	GSONMT00078417001	0.83

**APPENDIX E**

**SELECTED NEIGHBORING (< 50 KB DISTANCE) DIFFERENTIALLY  
EXPRESSED LNCRNA-PROTEIN CODING GENE PAIRS AND THEIR  
EXPRESSION CORRELATION**

<b>DE lncRNA</b>	<b>DE mRNA</b>	<b>Genomic distance</b>	<b>Correlation (R)</b>
gill_00031094	GSONMG00004464001	782	0.9990
gill_00042997	GSONMG00017163001	1721	0.9982
gill_00074276	GSONMG00037970001	1436	0.9966
gill_00039399	GSONMG00023147001	7666	0.9965
gill_00064660	GSONMG00077992001	1090	0.9964
gill_00007443	GSONMG00035251001	5097	0.9942
gill_00076634	GSONMG00034517001	9138	0.9936
Omy300080206	GSONMG00077992001	1569	0.9935
TCONS_00020586	GSONMG00046771001	44237	0.9925
gill_00077286	GSONMG00072161001	3144	0.9917
gill_00064659	GSONMG00077992001	1569	0.9911
Omy300051792	GSONMG00017163001	1721	0.9903
gill_00044605	GSONMG00000432001	1574	0.9902
Omy100165236	GSONMG00001899001	13871	0.9882
gill_00066004	GSONMG00062364001	545	0.9879
gill_00058188	GSONMG00004169001	1565	0.9878
gill_00047763	GSONMG00078104001	7804	0.9877
gill_00051074	GSONMG00074372001	546	0.9856
TCONS_00077288	GSONMG00057626001	1778	0.9848
gill_00077570	GSONMG00034517001	8236	0.9821
gill_00064423	GSONMG00051154001	1622	0.9813
gill_00079467	GSONMG00052539001	13815	-0.8246
Omy200060094	GSONMG00042231001	49277	-0.8364
TCONS_00001433	GSONMG00074929001	48641	-0.8382
gill_00057229	GSONMG00066040001	19368	-0.8399
gill_00052726	GSONMG00027735001	30983	-0.8429
gill_00069566	GSONMG00076698001	13428	-0.8622
Omy300084686	GSONMG00076698001	14124	-0.8698
Omy300068824	GSONMG00080266001	1882	-0.8715
gill_00064424	GSONMG00016295001	3931	-0.8725
gill_00069565	GSONMG00076698001	14124	-0.8747
TCONS_00000940	GSONMG00028360001	1901	-0.9084
gill_00088655	GSONMG00037382001	23955	-0.9178
gill_00076366	GSONMG00069466001	17304	-0.9413



## PROJECT CONCLUSION

Over the past decade, international efforts on genomic research in rainbow trout have provided valuable information including reference genome (Berthelot et al. 2014) reference. However, genomic information about rainbow trout is still limited. In a recently published first genome draft (Berthelot et al. 2014), not all mRNA genes appear to be reported. In addition, lncRNAs, which comprise the significant portion of the transcribed genome and play important gene regulatory roles, were not reported. In this project, we sequenced and assembled the transcriptome of mRNA and lncRNAs genes from diverse vital tissues to annotate the genome reference. The transcriptome assembly approach identified ~11,000 protein coding genes previously not reported by the genome sequencing project and also identified ~54,000 rainbow trout lncRNA transcripts for the first time. We investigated structural, expression, and sequence conservation characteristics of lncRNA transcripts for the first time in aquaculture species. Compared to protein coding genes, lncRNA transcripts were shorter, strikingly one-exon biased, lowly expressed, strictly tissue specific, and less conserved in sequence. Similar characteristics have been observed for lncRNAs from mammalian species (Derrien et al. 2012). LncRNA and protein coding transcript identified in this study will serve as a reference sequence for future functional genomics studies in rainbow trout.

In addition to genome annotation, we studied the role of non-coding RNAs in important aquaculture traits including disease resistance, growth and muscle quality traits. Two important classes of non-coding RNA studied in this project, lncRNAs and microRNAs, appeared to be heavily involved in regulation of these traits. We identified several lncRNAs differentially expressed (DE) during *Flavobacterium psychrophilum*

infection; and expression of DE lncRNAs correlated with the body bacterial load and infection susceptibility of the host strain. It is worth mentioning that expression of lncRNAs more strongly correlated with body bacterial load than did expression of immunity related protein coding genes. This finding suggests that lncRNA may play crucial role in modulation of antibacterial immune response in fish. DE lncRNAs showed genomic co-localization and strong expression correlation with DE immunity related protein coding genes. In addition to Flavobacterial infection, several lncRNAs were found to be differentially regulated during sexual maturation-associated skeletal muscle atrophy. These DE lncRNAs showed evidence of strong expression correlation and evidence of direct physical interaction with proteolytic genes involved in sexual maturation-associated muscle atrophy. This finding suggests that lncRNA may regulate skeletal muscle atrophy by regulating expression, activity, and/or stability of muscle atrophy protein coding genes.

We also studied the role of microRNA expression and genetic variation in microRNA binding sites of target genes involved in growth and muscle quality traits. Several microRNAs showed differential expression between fish families with divergent phenotypes, and their expression explained significant variation in the phenotype. In addition, genetic variations in microRNA binding sites of target genes explained significant variation in growth and muscle quality traits. We identified hundreds of SNPs in 3' UTR of target genes that either destroyed or created novel illegitimate microRNA binding sites in genes important in growth and muscle quality. A genotype-phenotype association study performed in a large set of fish population showed that significant variation in phenotypes within the population was explained by these SNPs. These SNPs may control the fate of growth related target genes by altering the microRNA recognition, which in turn, may

contribute to the phenotype. In a separate study, we quantified DE protein coding genes between fish families showing divergent phenotypes for growth and muscle quality traits, which led unexpectedly small number of differentially expressed genes (to be published elsewhere). Compared to protein coding genes, a much larger fraction of trout microRNAs showed differential expression. This finding suggests that growth and muscle quality traits in trout may be controlled more by variation in microRNA expression and genetic variation in the target gene that interferes with microRNA-target recognition, rather than by transcriptional regulation of protein coding genes. Consistent with our finding, in some domesticated animals, genetic variation that impacts microRNA-mRNA interaction are proved to be an ideal genetic marker to enhance muscle growth (Cloup et al. 2006). In the present study, we have provided the first evidence of microRNA association, in terms of expression and genetic variation, with important muscle quality traits in trout.

In this study, we also investigated functional interplay between coding and non-coding RNAs, and the potential role of such interactions in phenotype. Using sexual maturation-associated skeletal muscle atrophy as a model system, we elucidated that microRNA, lncRNA, and protein coding genes extensively cross talk among each other by multiple mechanisms. We observed lncRNA-mRNA, lncRNA-microRNA, and microRNA-mRNA cross talks in terms of genomic co-localization, expression correlation, and direct physical interaction. LncRNAs cross-talked with microRNAs by potentially binding or generating microRNAs. These interactions are crucial in gene regulation as lncRNA-microRNA binding makes cellular microRNAs unavailable for mRNA downregulation (Kallen et al. 2013). On the other hand, generation of microRNAs from lncRNA transcripts has the opposite effect (Dey, Pfeifer and Dutta 2014). Likewise,

lncRNAs cross-talked with protein coding genes by potentially competing with mRNAs to bind a common microRNA, and by potential direct physical interaction with mRNA and protein sequences. Direct 'lncRNA-mRNA' physical interactions play a crucial role in regulation of protein coding genes (Gong and Maquat 2011), which can have multiple consequences including mRNA decay (Gong and Maquat 2011) and translation suppression (Yoon et al. 2012). Similarly, lncRNA-protein direct physical interaction modulates stability (Taniue et al. 2016), availability (sequestration) (Hirose et al. 2014), activity (Tripathi et al. 2010), and proper cellular localization (Tripathi et al. 2010) of the protein. Though such mechanisms have been discovered in mammalian species, this is the first study in aquaculture fish that has investigated coding-non-coding RNA interactions. In the same study, we observed that functionally related protein coding genes and lncRNAs tend to co-localize or cluster together in genome. LncRNA-mRNA gene pairs that either co-localize in genome or show evidence of direct physical interaction or potentially compete for the same microRNAs are often strongly correlated in expression than random lncRNA-mRNA gene pairs. It also suggests that such coding-non-coding RNA interactions have important functional consequences in cells.

To conclude, the present study explores the role of non-coding RNAs in regulation of important aquaculture traits in rainbow trout and suggests that non-coding RNA-mediated gene regulation plays a critical role in regulation of these traits.

## PROJECT REFERENCES

- Al-Tobasei, R., B. Paneru & M. Salem (2016) Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout. *PLoS One*, 11, e0148940.
- Asche, F., H. Håvard, T. Ragnar & T. Sigbjørn (2009) The salmon disease crisis in Chile. *Marine Resource Economics*, 24, 405-411.
- Bagga, S., J. Bracht, S. Hunter, K. Massirer, J. Holtz, R. Eachus & A. E. Pasquinelli (2005) Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell*, 122, 553-63.
- Ballarino, M., V. Cazzella, D. D'Andrea, L. Grassi, L. Bisceglie, A. Cipriano, T. Santini, C. Pinnarò, M. Morlando, A. Tramontano & I. Bozzoni (2015) Novel long noncoding RNAs (lncRNAs) in myogenesis: a miR-31 overlapping lncRNA transcript controls myoblast differentiation. *Mol Cell Biol*, 35, 728-36.
- Berthelot, C., F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noël, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J. M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J. N. Volff, C. Genêt, P. Wincker, O. Jaillon, H. Roest Crollius & Y. Guiguen (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*, 5, 3657.
- Cabianca, D. S., V. Casa, B. Bodega, A. Xynos, E. Ginelli, Y. Tanaka & D. Gabellini (2012) A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell*, 149, 819-31.
- Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev & J. L. Rinn (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 25, 1915-27.
- Carpenter, S., D. Aiello, M. K. Atianand, E. P. Ricci, P. Gandhi, L. L. Hall, M. Byron, B. Monks, M. Henry-Bezy, J. B. Lawrence, L. A. O'Neill, M. J. Moore, D. R. Caffrey & K. A. Fitzgerald (2013) A long noncoding RNA mediates both activation and repression of immune response genes. *Science*, 341, 789-92.
- Carson, L. & J. Schmidtke (1995) Characteristics of *Flexibacter psychrophilus* isolated from Atlantic salmon in Australia. *Diseases of Aquatic Organisms*, 21, 157-161.
- Chen, L., P. Wu, X. H. Guo, Y. Hu, Y. L. Li, J. Shi, K. Z. Wang, W. Y. Chu & J. S. Zhang (2014) miR-143: a novel regulator of MyoD expression in fast and slow muscles of *Siniperca chuatsi*. *Curr Mol Med*, 14, 370-5.
- Clark, M. B., A. Choudhary, M. A. Smith, R. J. Taft & J. S. Mattick (2013) The dark matter rises: the expanding world of regulatory RNAs. *Essays Biochem*, 54, 1-16.
- Clop, A., F. Marcq, H. Takeda, D. Pirottin, X. Tordoir, B. Bibe, J. Bouix, F. Caiment, J. M. Elsen, F. Eychenne, C. Larzul, E. Laville, F. Meish, D. Milenkovic, J. Tobin, C. Charlier & M. Georges (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet*, 38, 813-818.
- Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C.

- A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow & R. Guigó (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, 22, 1775-89.
- Dey, B. K., K. Pfeifer & A. Dutta (2014) The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration. *Genes Dev*, 28, 491-501.
- FAO (2016) The State of World Fisheries and Aquaculture." *Food and Agriculture organization of United Nations*.
- Fatica, A. & I. Bozzoni (2014) Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet*, 15, 7-21.
- Gjedrem, T. (1997) Flesh quality improvement in fish through breeding. *Aquaculture Int*, 5, 197-206.
- (2005) *Selection and breeding programs in aquaculture*. Dordrecht: Springer.
- . 2008a. *Selection and Breeding Programs in Aquaculture*. New York: Springer.
- (2010) The first family-based breeding program in aquaculture. *Reviews in Aquaculture*, 2, 2-15.
- Gjedrem, T., Ø (2008b) Improving farmed fish quality by selective breeding. *Improving farmed fish quality and safety*, 265-274.
- Gong, C. & L. E. Maquat (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, 470, 284-8.
- Gómez, E., J. Méndez, D. Cascales & J. A. Guijarro (2014) Flavobacterium psychrophilum vaccine development: a difficult task. *Microb Biotechnol*, 7, 414-23.
- Hirose, T., G. Virnicchi, A. Tanigawa, T. Naganuma, R. Li, H. Kimura, T. Yokoi, S. Nakagawa, M. Bénard, A. H. Fox & G. Pierron (2014) NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Mol Biol Cell*, 25, 169-83.
- Houston, R. D., S. C. Bishop, A. Hamilton, D. R. Guy, A. E. Tinch, J. B. Taggart, A. Derayat, B. J. McAndrew & C. S. Haley (2009) Detection of QTL affecting harvest traits in a commercial Atlantic salmon population. *Anim Genet*, 40, 753-5.
- Jacob, F., B. Rønsholdt, N. Alsted & T. Borresen (1995) Fillet texture of rainbow trout as affected by feeding strategy, slaughtering procedure and storage post mortem. *Water science and technology*, 31, 225-231.
- Johnston, I. A., H. T. Lee, D. J. Macqueen, K. Paranthaman, C. Kawashima, A. Anwar, J. R. Kinghorn & T. Dalmy (2009) Embryonic temperature affects muscle fibre recruitment in adult zebrafish: genome-wide changes in gene and microRNA expression associated with the transition from hyperplastic to hypertrophic growth phenotypes. *J Exp Biol*, 212, 1781-93.
- Kallen, A. N., X. B. Zhou, J. Xu, C. Qiao, J. Ma, L. Yan, L. Lu, C. Liu, J. S. Yi, H. Zhang, W. Min, A. M. Bennett, R. I. Gregory, Y. Ding & Y. Huang (2013) The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Mol Cell*, 52, 101-12.
- Kino, T., D. E. Hurt, T. Ichijo, N. Nader & G. P. Chrousos (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal*, 3, ra8.

- Kornfeld, J. W. & J. C. Brüning (2014) Regulation of metabolism by long, non-coding RNAs. *Front Genet*, 5, 57.
- Lewis, B. P., C. B. Burge & D. P. Bartel (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120, 15-20.
- Marancik, D., G. Gao, B. Paneru, H. Ma, A. G. Hernandez, M. Salem, J. Yao, Y. Palti & G. D. Wiens (2014) Whole-body transcriptome of selectively bred, resistant-, control-, and susceptible-line rainbow trout following experimental challenge with *Flavobacterium psychrophilum*. *Front Genet*, 5, 453.
- Mercer, T. R., M. E. Dinger, S. M. Sunkin, M. F. Mehler & J. S. Mattick (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A*, 105, 716-21.
- Michie, I. (2001) Causes of downgrading in the salmon farming industry. *Farmed fish quality*, 129-136.
- Mishima, Y., C. Abreu-Goodger, A. A. Staton, C. Stahlhut, C. Shou, C. Cheng, M. Gerstein, A. J. Enright & A. J. Giraldez (2009) Zebrafish miR-1 and miR-133 shape muscle gene expression and regulate sarcomeric actin organization. *Genes Dev*, 23, 619-32.
- Mørkøre, T., J. I. Vallet, M. Cardinal, M. C. Gomez-Guillen, P. Montero, O. J. Torrissen, R. Nortvedt, S. Sigurgisladottir & M. S. Thomassen (2001) Fat content and fillet shape of Atlantic Salmon: Relevance for processing, yield and quality of raw and smoked products. *Journal of food science*, 66, 1348-1354.
- Nematollahi, A., A. Decostere, F. Pasmans & F. Haesebrouck (2003) *Flavobacterium psychrophilum* infections in salmonid fish. *J Fish Dis*, 26, 563-74.
- NOAA (2014) Import and export of fishery products annual summary.
- Olsen, P. H. & V. Ambros (1999) The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol*, 216, 671-80.
- Pandey, R. R., T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, T. Nagano, D. Mancini-Dinardo & C. Kanduri (2008) *Kcnqlot1* antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell*, 32, 232-46.
- Prasanth, K. V., S. G. Prasanth, Z. Xuan, S. Hearn, S. M. Freier, C. F. Bennett, M. Q. Zhang & D. L. Spector (2005) Regulating gene expression through RNA nuclear retention. *Cell*, 123, 249-63.
- Ramachandra, R. K., M. Salem, S. Gahr, C. E. Rexroad & J. Yao (2008) Cloning and characterization of microRNAs from rainbow trout (*Oncorhynchus mykiss*): their expression during early embryonic development. *BMC Dev Biol*, 8, 41.
- Rinn, J. L. & H. Y. Chang (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem*, 81, 145-66.
- Salem, M., M. L. Manor, A. Aussanasuwannakul, P. B. Kenney, G. M. Weber & J. Yao (2013) Effect of sexual maturation on muscle gene expression of rainbow trout: RNA-Seq approach. *Physiol Rep*, 1, e00120.

- Salem, M., R. L. Vallejo, T. D. Leeds, Y. Palti, S. Liu, A. Sabbagh, C. E. Rexroad, 3rd & J. Yao (2012) RNA-Seq identifies SNP markers for growth traits in rainbow trout. *PLoS One*, 7, e36264.
- Silverstein, J. T., R. L. Vallejo, Y. Palti, T. D. Leeds, C. E. Rexroad, T. J. Welch, G. D. Wiens & V. Ducrocq (2009) Rainbow trout resistance to bacterial cold-water disease is moderately heritable and is not adversely correlated with growth. *J Anim Sci*, 87, 860-7.
- Steine, G., F. Alfnes & M. B. Rørå (2005) The Effect of Color on Consumer WTP for Farmed Salmon. *Marine Resource Economics*, 20, 211-219.
- Taniue, K., A. Kurimoto, H. Sugimasa, E. Nasu, Y. Takeda, K. Iwasaki, T. Nagashima, M. Okada-Hatakeyama, M. Oyama, H. Kozuka-Hata, M. Hiyoshi, J. Kitayama, L. Negishi, Y. Kawasaki & T. Akiyama (2016) Long noncoding RNA UPAT promotes colon tumorigenesis by inhibiting degradation of UHRF1. *Proc Natl Acad Sci U S A*, 113, 1273-8.
- Tripathi, V., J. D. Ellis, Z. Shen, D. Y. Song, Q. Pan, A. T. Watt, S. M. Freier, C. F. Bennett, A. Sharma, P. A. Bubulya, B. J. Blencowe, S. G. Prasanth & K. V. Prasanth (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell*, 39, 925-38.
- Tsai, M. C., O. Manor, Y. Wan, N. Mosammamaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal & H. Y. Chang (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, 329, 689-93.
- Vallejo, R. L., T. D. Leeds, B. O. Fragomeni, G. Gao, A. G. Hernandez, I. Misztal, T. J. Welch, G. D. Wiens & Y. Palti (2016) Evaluation of Genome-Enabled Selection for Bacterial Cold Water Disease Resistance Using Progeny Performance Data in Rainbow Trout: Insights on Genotyping Methods and Genomic Prediction Models. *Front Genet*, 7, 96.
- Wang, X. H. (2013) MicroRNA in myogenesis and muscle atrophy. *Curr Opin Clin Nutr Metab Care*, 16, 258-66.
- Wu, L., J. Fan & J. G. Belasco (2006) MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A*, 103, 4034-9.
- Yan, B., J. T. Guo, C. D. Zhu, L. H. Zhao & J. L. Zhao (2013a) miR-203b: a novel regulator of MyoD expression in tilapia skeletal muscle. *J Exp Biol*, 216, 447-51.
- Yan, B., C. D. Zhu, J. T. Guo, L. H. Zhao & J. L. Zhao (2013b) miR-206 regulates the growth of the teleost tilapia (*Oreochromis niloticus*) through the modulation of IGF-1 gene expression. *J Exp Biol*, 216, 1265-9.
- Yoon, J. H., K. Abdelmohsen, S. Srikantan, X. Yang, J. L. Martindale, S. De, M. Huarte, M. Zhan, K. G. Becker & M. Gorospe (2012) LincRNA-p21 suppresses target mRNA translation. *Mol Cell*, 47, 648-55.
- Zhang, B., Q. Wang & X. Pan (2007) MicroRNAs and their regulatory roles in animals and plants. *J Cell Physiol*, 210, 279-89.
- Zhu, Q. H. & M. B. Wang (2012) Molecular Functions of Long Non-Coding RNAs in Plants. *Genes (Basel)*, 3, 176-90.