Allele-specific Expression of Ribosomal Protein Genes in Interspecific Hybrid Catfish

by

Ailu Chen

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 1, 2015

Keywords: catfish, interspecific hybrids, allele-specific expression, ribosomal protein

Approved by

Zhanjiang Liu, Chair, Professor, School of Fisheries, Aquaculture and Aquatic Sciences
Nannan Liu, Professor, Entomology and Plant Pathology
Eric Peatman, Associate Professor, School of Fisheries, Aquaculture and Aquatic Sciences
Aaron M. Rashotte, Associate Professor, Biological Sciences

Abstract

Interspecific hybridization results in a vast reservoir of allelic variations, which may potentially contribute to phenotypical enhancement in the hybrids. Whether the allelic variations are related to the downstream phenotypic differences of interspecific hybrid is still an open question. The recently developed genome-wide allele-specific approaches that harness high-throughput sequencing technology allow direct quantification of allelic variations and gene expression patterns. In this work, I investigated allele-specific expression (ASE) pattern using RNA-Seq datasets generated from interspecific catfish hybrids. The objective of the study is to determine the ASE genes and pathways in which they are involved. Specifically, my study investigated ASE-SNPs, ASE-genes, parent-of-origins of ASE allele and how ASE would possibly contribute to heterosis. My data showed that ASE was operating in the interspecific catfish system. Of the 66,251 and 177,841 SNPs identified from the datasets of the liver and gill, 5,420 (8.2%) and 13,390 (7.5%) SNPs were identified as significant ASE-SNPs, respectively. With these SNPs, a total of 1,519 and 3,075 ASE-genes have been identified. Gene Ontology analysis has revealed that genes encoding for cytoplasmic ribosomal proteins were highly enriched among ASE genes. Parent-of-origins of imbalanced alleles were determined for 27 and 30 ASE ribosomal protein genes in liver and gill, respectively. Of the 27 ASE ribosomal protein genes in the liver, 13 were of channel catfish origin and 14 were of blue catfish origin. Similarly, of the 30 ASE ribosomal protein genes in the gill, 16 were of channel catfish origin, while 14 were of blue catfish origin. Therefore, it appeared that ASE was not related to selected expression of a set of ribosomal protein genes from a specific parent. However, each RP gene appeared to be almost exclusively expressed from only one parent, indicating that ribosomes in the hybrid catfish were in “hybrid” forms. It’s also observed that the expression percentage of ribosomal protein genes out of total genes in gill was smaller in hybrid catfish (19.75%) than channel catfish (25.31%), indicating that hybrid ribosomes probably worked more efficiently than their homozygous counterparts.

My study is the very first of its kind in catfish to determine if ASE exists in the interspecific hybrid system. It provides a new avenue of research to discover the genetic interactions at the transcriptional level and genome scale.

Acknowledgement

I would like give my special thanks to my advisor Dr. Zhanjiang Liu for his valuable guidance, encouragement, support and patience during my Ph.D. study. I would also like to thank all my other committee members: Dr. Nannan Liu, Dr. Eric Peatman and Dr. Arran Rashotte for making me learn more and be confidence in myself and my knowledge. My thanks also extend to the university reader, Dr. Charles Y. Chen for his support and guidance. I am very grateful for the technical assistance by Dr. Huseyin Kucuktas and Ludmilla Kaltenboeck. And I would like to thank Dr. Shikai Liu, Ruijia Wang, Luyang Sun, Lisui Bao, Chen Jiang, Chao Li, Yun Li, Jun Yao, Qifan Zeng and all the other colleagues in the laboratory for their help, collaboration, and friendship. I would like to especially thank the Chinese Scholarship Council for the financial support.

I am grateful to my father Yong Chen, my mother Guangmin He, my father-in-law Yanping Hu, my mother-in-law Difei Qi and my beloved husband Chenxi Hu for their endless love and tremendous support.

Table of Contents

Abstract ii

Acknowledgement iv

List of Tables viii

List of Figures ix

List of Abbreviations x

Chapter 1: Introduction 1

1.1 Overview 1

1.2 Selective Breeding 2

1.2.1 Selection 2

1.2.2 Overview of Selective Breeding 3

1.2.3 Achievements of Selective Breeding in Aquaculture 4

1.3 Interspecific Hybridization 6

1.3.1 Progress in Interspecific Hybridization 6

1.3.2 Generation of Interspecific Catfish Hybrids 7

1.3.3 Genomic Changes in Interspecific Hybridization 9

1.3 Next-generation Sequencing 6

1.4 Research Purpose 13

References 13

Chapter 2: Review of Literature 16

2.1 Genetic Markers 16
2.1.1 Microsatellites and SNPs 16
2.1.2. SNP Identification and Application in Catfish 18
2.2 Allele-specific expression studies 20
2.2.1 Polymorphism-directed Approach 21
2.2.2 Global Approach 23
2.2.3 ASE Studies in Aquaculture Species 25
References 27
Chapter 3: Methodology 30
3.1 RNA-Seq 30
3.2 Reference Mapping 31
3.3 SNP Identification 32
3.4 SNP Filtering 35
3.5 Identification of ASE-SNPs 36
3.6 Identification of the Parent-of-Origins 37
3.7 Ontology Analysis of ASE-genes 38
3.8 Expression Analysis of ASE-genes 40
3.9 Identification of ribosomal RNAs 41
3.10 Identification of thermal induced ASE-SNPs 42
3.11 Analysis of thermal induced ASE-genes 43
Reference 43

Chapter 4: Results 45

4.1 Identification of SNPs and ASE-SNPs 45

4.2 Types of SNPs and ASE-SNPs 46

4.3 Identification of ASE-genes 50

4.4 Ontology Analysis of ASE-genes 52

4.5 Allele Ratio Distribution of RP Genes 55

4.6 Identification of SNPs and ASE-SNPs 55

4.7 Ribosomal proteins were highly expressed in catfish 58

4.8 ASE of ribosomal RNAs 59

4.9 ASE of mitochondrial RPs 60

4.10 Identification of thermal induced ASE-SNPs and thermal induced ASE-genes ... 62

4.11 Analysis of thermal induced ASE-genes 63

Chapter 5: Discussion 65

5.1 Transition SNP percentage Decreased from Initial SNPs to ASE-SNP 65

5.2 RPs Expression and Growth and Fitness 66

5.3 Less RP Transcripts were Expressed in Hybrid Catfish 69

5.4 Different RP Genes Had Different Parent-of-origins 70

References 71

Appendix 73

List of Tables

Table 1. Example of SNP identification output from Popoolation 2.. 34

Table 2. rRNAs Accession Numbers of channel catfish and zebrafish.. 41

Table 3. Summary of SNPs, ASE-SNPs and ASE-genes from the hybrid catfish transcriptome 46

Table 4. SNP type proportion of initial SNPs and ASE-SNPs in the liver and gill tissues of hybrids catfish. 47

Table 5. Enriched GO terms of ASE-genes in the liver and gill. 54

Table 6. ASE RP Genes Identified in Two Tissues 56

Table 7. RPKM value of different tissues in different catfish. 58

Table 8. Expression Status of rRNAs.. 59

Table 9. ASE mRP Genes Identified in Two Tissues. 61

Table 10. Summary of SNPs, thermal induced-SNPs and thermal induced-genes from the hybrid catfish transcriptome. 62

Table 11. Enriched GO terms of thermal induced ASE-genes in the liver and gill 64

List of Figures

Figure 1. SNP type proportion of initial SNPs and ASE-SNPs in the liver and gill tissues of hybrids catfish. **A:** Initial SNP type proportion in the liver; **B:** Initial SNP type proportion in the gill; **C:** ASE-SNP type proportion in the liver; **D:** ASE-SNP type proportion in the gill. 48

Figure 2. Allele ratios distribution of ASE-Genes in the liver and gill of hybrid catfish. 51

List of Abbreviations

ASE	Allele-specific expression
SNP	Single nucleotide polymorphism
ASE-gene	allelic-specifically expressed genes
RP gene	Gene encodes for ribosomal protein
mRP	Mitochondrial Ribosomal Protein
rRNA	Ribosomal RNA
RPKM	Read per kb per million reads
QTL	Quantitative trait loci
RNA-Seq	Transcriptome sequencing

Chapter 1: Introduction

1.1 Overview

Catfish is the major aquaculture species in the United States, which accounts for over 60% of the total the US aquaculture production. The channel catfish (*Ictalurus punctatus*) and blue catfish (*I. furcatus*) are two major aquaculture catfish species. Their interspecific hybrids (channel catfish female × blue catfish male) have been widely produced for aquaculture because they outperform their parents in a number of production traits, including growth rate, feed conversion efficiency, disease resistance, and low oxygen resistance. Hybrid catfish production has increased substantially; they comprise about 20% of catfish harvested in 2011.

Interspecific hybridization results in a vast reservoir of allelic variations, which may potentially contribute to phenotypical enhancement in the hybrids. Whether the allelic variations are related to the downstream phenotypic differences of interspecific hybrid is still an open question. Previous allelic expression studies focused on the interaction between alleles, which have brought up exciting observations. As many cis- and trans- gene regulation patterns were observed in hybrids, allele specific expression (ASE) analysis was brought up as a hotspot study to discover the mechanisms underlying heterosis. The rich allelic variants in interspecific hybrids make them good models for ASE analysis. Previous ASE studies reported in insects, fish, mammals and several plants were focused only on a small set of ASE genes. With the development of sequencing technology, next-generation sequencing based transcriptome analysis (RNA-Seq) became routines, allowing for genome-level ASE analysis.

1.2 Selective Breeding

1.2.1 Selection

In biology, the term selection is defined as “a process in which environment or genetic influences determine which types of organism thrive better than others, regarded as a factor in revolution.” During evolution, certain traits or alleles undergo segregation. Under selection, individuals with adaptive traits tend to be kept in the population since they have potentials to contribute more offspring to the succeeding generation than others. When certain traits are characterized with genetic basis, selection can be made to maintain or even improve these traits because offspring will inherit those traits from their parents. When selection is tense and persistent, adaptive traits become universal to population or species and this process is called evolution.

Selection can be made in multiple developmental stages including eggs and sperm, embryos, juveniles and adults. Factors that lead to selection are called selective pressures. Selective pressure can be either physical (weather, nourishment, habitat space) or biological (predator, disease, mates) (Bell 1997).

Selections occur only when there is diversity of certain traits in a population. In the absence of individual variation or when variations are selective neutral, selections do not occur. Selections are divided into two groups: natural selection and artificial selection. Natural selections are further subcategorized into sexual selection, ecological selection, stabilizing selection, disruptive selection and directional selection. Sexual selection is a result of mate competition. There are two ways of sexual selection: one is intra-sexual, as in cases of competition among individuals of the same sex in a population; the other one is inter-sexual, as in cases where one sex controls reproductive access by choosing among a population of available mates. Ecological selection is natural selection via any other means than sexual selection. Sometimes natural selection is defined as synonymous with ecological selection, and sexual selection is then classified as a separate mechanism to natural selection (Mayr 1972). Natural selection results from the struggle to survive while sexual selection emerges from the struggle to reproduce. Artificial selection, or more commonly called selective breeding, refers to the process by which human breed animals or plants in order to improve or acquire certain traits.

1.2 2 Overview of Selective Breeding

In biology, selective breeding is defined as “the intentional mating of two animals in an attempt to produce offspring with desirable characteristics or for the elimination of a trait.” The process of selectively breeding of a strain is called domestication. Bred animals are known as breeds while bred plants are known as varieties, cultigens and cultivars. The hybrid offspring of two purebred animals from different breeds are called crossbreed. Selective breeding of both plants and animals has been practiced since early prehistory in species such as wheat, rice and dogs. These domesticated plants and animals have now significant different from their wild ancestors. Although selective breeding has been largely practiced by the Romans about 2000 years ago, it was established as a scientific practice in the 18th century by Robert Bakewell during the British Agricultural Revolution. His most remarkable work is selective breeding of sheep and he is known for having developed a fine-boned, long wool sheep breed (Pawson 1957).

Selective breeding in aquaculture has high potential for the genetic improvement of fish and shellfish. This potential benefit hasn't been realized until recently because the high mortality narrowed the selection to only a few broodstock, which result in inbreeding depression, thus forced breeders to use wild broodstock. For example, the practice in selective breeding towards higher growth rate often result in slow growth and high mortality (Gjedrem and Baranski 2010).

Aquaculture species are reared for particular traits such as growth rate, survival rate, meat quality, disease resistance and fecundity. Growth rate is measured as either gain of body weight or body length. It is the most important economic trait for all aquaculture species as faster growth speeds up the turnover of production. Survival rate may be associated with disease resistance and stress response. It is also very important since the number of survived organism accounts for the total production (Gjedrem 1983). Meat quality takes into account fish size, meatiness, fat percentage, flesh color, taste etc. It is directly related to the market value and how customer will like it. Fecundity is usually not considered as an important trait in aquaculture selective breeding because fish and shellfish produce large quantity of offspring compared to terrestrial livestock. However, practices found that egg quality are correlated with survival rate the early growth rate (Gjedrem 1985)

1.2.3 Achievements of Selective Breeding in Aquaculture

Many aquaculture species showed great response to selection. In salmonids, selections towards growth rate and disease resistance are very successful. It is reported that Atlantic salmon showed an increase in body weight by 30% per generation during selection. Selected fish had a twice better growth rate, a 40% higher feed intake, 20% better fed convert efficiency, and an increased protein and energy retention as compared to the wild stock (Thodesen, Grisdale-Helland et al. 1999). Selection was also performed to resist the Infectious Pancrearic Necrosis Virus (IPNNV). The results demonstrated that high-resistant species showed 29.3% survival rate compared to wild species (Storset, Strand et al. 2007). Another salmonid, rainbow trout, is also reported to have great improvement in growth rate after selection with 7% rate per generation (Kause, Ritola et al. 2005). Selective breeding of rainbow trout in Japan achieved a high IPNV resistance strain which has only 4.3% mortality (Okamoto, Tayama et al. 1993). Selection for live weight of pacific oysters showed improvement ranging from 0.4% to 25.6% compared to wild stock (Langdon, Evans et al. 2003). In pacific white shrimp, after one generation of selection, a 21% increase was observed in growth and 18.4% increase in survival to Taura Syndrome Virus (TSV) (Argue, Arce et al. 2002).

Channel catfish showed great improvement in body weight and growth rate after selection. Productions of channel catfish in ponds can be greatly enhanced by using improve lines and hybrids. There are three breeding programs to improve the growth rate in channel catfish: mass selection, intraspecific crossbreeding and interspecific hybridization (Smitherman, Dunham et al. 1983). Mass selection refers to a form of breeding with outperforming individuals to produce next generation. Intraspecific breeding is the act of breeding two varieties within the same species while interspecific hybridization refers to the crossing of two species that are from within the same genus. In early trials, single generation selections for body weight in channel catfish have been reported in grain of 11-18% respond to different lines (Bondari 1980; Bondari 1983). Disease resistance has also been improved during selection for body weight (Dunham and Smitherman 1985). Interspecific hybrid catfish, generated by crossing female channel catfish and male blue catfish display heterosis over their inbred channel and blue catfish in many traits such as growth rate and low oxygen resistance. Hybrid catfish have been reported to yield increases over channel catfish in body weight of 18-20% (Yant, Smitherman et al. 1975), catchability (Tave, Mcginty et al. 1981), and resistance to low dissolved oxygen (Dunham, Smitherman et al. 1983).

1.3 Interspecific Hybridization

1.3.1 Progress in Interspecific Hybridization

Interspecific hybridization has been a topic of interest for researchers in the field of ecology, taxonomy and systematics. It is a primary source of data for studies on speciation and adaptation (Schwenk, Brede et al. 2008). Over 150 years, extensive studies have been conducted on interspecific hybridization. Early studies were aiming at developing a theoretical framework for interspecific hybridization. Several botanists started to experimental study the interspecific hybridization by crossing experiments and field studies from 1930s (Anderson and Hubricht 1938; Anderson 1948). These studies provided evidence for that the genetic information was exchanged between hybrid species and this phenomenon was not rare. During 1950s and1960s, many plant and animal models were established for the study of hybridization. Only until 1980s to 1990s, researchers have realized that the result of interspecific hybridization can be narrowed to a hybrid zone and multiple evolutionary pathways were responsible to explain the underlying mechanism (Barton and Hewitt 1985; Harrison 1990). Hybrid zones are "locations where the hybrid offspring of two divergent taxa (species/subspecies/races) are prevalent and there is a cline in the genetic composition of populations from one taxon to the other." With the technique innovations such as polymerase chain reaction (PCR), later studies were able to investigate genetic analysis, hybrid fitness and selection in hybrid zones. In late 1990s, genetic variations at nuclear and mitochondrial loci of interspecific hybridization of animals were major focus in this field (Dowling and Secor 1997). With the development of genetic markers such as microsatellite, AFLP and SNP, the progress to analysis interspecific hybridization sped up towards adaptive radiations and introgression (Seehausen 2004).

1.3.2 Generation of Interspecific Catfish Hybrids

Catfish belong to the order Siluriformes, they are a diverse groups of ray-finned fish. Catfish are named for their predominant barbels, which resemble cat's whiskers. However, not all catfish have predominant barbell. Members in the Siluriformes order are defined by features of the skull and swimbladder. Catfish live inland and in coastal waters of every continent except Antarctica, and over half of all catfish species live in the Americas. They are often found in shallow, running freshwater environments (Bruton 1996). There are at 39 species of catfish in North America, but only six have been cultured. They are the blue catfish, *Ictalurus furcatus* (LeSueur); the white catfish, *Ictalurus catus* (Linnaeus); the black bullhead, *Ictalurus melas* (Rafinesque); the brown bullhead, *Ictalurus nebulosus* (LeSueur); the yellow bullhead, *Ictalurus natalis* (LeSueur); and the flathead catfish, *Pylodictis olivaris* (Rafinesque) (Wellborn 1988). The family Ictaluridae (Ictalurids) is a family of catfish native to North America. They are important food fish and sport fish. Ictalurid species have four pairs of barbels and no scales. Channel catfish (*Ictalurus puctatus*) were originally found only in the Gulf States and the Mississippi Valley north to the prairie provinces of Canada and Mexico but now it is the most numerous catfish species in North America. In the United States, the popularity of channel catfish for food has contributed to the growth of aquaculture of this species.

Several hybridization experiments have been conducted between the seven major North America catfish. Of these 42 different interspecific catfish hybrids, only one outperforms the predominantly cultured channel catfish. This hybrid is generated by crossing the female channel catfish and male blue catfish (C × B hybrids). However, the reciprocal cross doesn't play much heterosis as C × B hybrids. Researches on C × B hybrids demonstrated that they have been improved in many desired commercial characteristics. Dunham and Masser summarized that, the C × B hybrids is superior to channel catfish because its faster growth, better feed conversion, tolerance of low oxygen, increased resistance to many disease, tolerance of crowded growth conditions in ponds, uniformity in size and shape, higher dressout percentage and fillet yield, increased harvestability by seining and increased vulnerability to angling (Dunham and Masser 2012). Although the heterosis resulted from the interspecific hybridization of channel catfish and blue catfish has long been a topic of interest, no systematic study is conducted to explore the underlying mechanisms.

Most channel catfish reach to the sexual maturity at 3 years of age, and most blue catfish become mature at 5 years of age. Practically, 4 to 5 year old channel catfish are the most reliable for hybridization in early spawning seasons (Dunham and Masser 2012). There are three spawning strategies: open-pond spawning, pen spawning and artificial fertilization. Artificial fertilization is the most stable strategy and it can make large C × B hybrids production possible. Females are induced to ovulate with LHRHa, carp pituitary extract (CPE), or channel catfish pituitary extract (CCPE). After two round of injection, females are placed in spawning bags and suspend in holding vats for ovulation. When females are ovulating, they are narcotized with MS-222 for stripping process. Females are dried gently with towel and are hold head up and tail down during the stripping process, with the genital opening just above a metal pie pan lightly coated with vegetable shortening. The males must be sacrificed for their testes. The white parts of the testes are removed by gently cutting the connective tissue and then dried with paper towel. Testes are then blended with saline solution. Stripped eggs and sperm solutions are then mixed. The general rule is that one male fertilizes five to ten females, depending on the size and quality of the eggs.

Oxygenated water is then added to the mixture to activate fertilization (Masser and Dunham 1998).

1.3.3 Genomic Changes in Interspecific Hybridization

With the development of modern molecular technologies, many genomic and epigenetic changes have been observed in hybrids such as chromosomal rearrangements, transposable elements activation and gene expression alterations (Baack and Rieseberg 2007). These genomic changes may be caused by selection for fertility and ecological traits, and may result in phenotypical superiority in the hybrid offspring. Genomic changes have potential to stabilizing hybrid genomes and to produce novel gene expression patterns and phenotypes, thus proving an insight view of the mechanism underlying heterosis.

A series of studies on allopolyploid reveals that genomic changes occur immediately right after hybridization, including gene loss, gene silencing, gene expression alteration and tissues-specific expression (Paun, Fay et al. 2007). Allopolyploid is a polyploid individual or strain having a chromosome set composed of two or more chromosome sets derived more or less complete from different species. The generation process of allopolyploid combines hybridization and genome duplication. The genetic and epigenetic changes of allopolyploid are summarized as follows: 1) Chromosomal rearrangement of parental genomes contributes to proper meiotic pairing and isolation in the hybrid. Many hybrid plants and animals have been observed with chromosomal rearrangement and it is believed that genome rearrangements are necessary for restoring nuclear-cytoplasmic compatibly (Soltis and Soltis 1999). DNA sequence elimination, including gene deletions, is considered to be associated with chromosomal rearrangement. This process results in differentiation of homoeologous chromosomes, leading to correct meiotic pairing in hybrids

(Tate, Ni et al. 2006). 2) Transposable elements activation may lead to gene expression alteration, facilitating genomic reorganization. Active transposable elements have the potential for insertional mutagenesis and changes in phenotype while altering local patterns of gene expression. 3) Many of the gene expression showed organ or tissue specific pattern, indicating that there are differential regulation of the homoeologous combined genomes (Adams, Cronn et al. 2003).

Regulation of gene expression begins at the transcription initiation and is associated with a variety of regulatory factors including basal promoter, various protein complexes, DNA methylation and histone modification (Landry, Hartl et al. 2007). Transcription happens at the basal promoter region which is located at the 5' upstream of the transcription unit. Common promoter elements are TATA box, initiator sequence and downstream promoter elements etc. These elements interact with RNA polymerase II holoenzyme complex and general transcription factor for transcription. Regulatory modules are 5-15 bp long scattered DNA binding sites located close to the transcription unit in the upstream, but sometimes they are also located in the downstream such as in the introns. They usually interact with specific factors and result in tissue-/gene-/allele- specific expression. Enhancers are usually located at several kilobases upstream from the gene. They mediate gene expression by remodeling of chromatin structure and through protein-protein interactions with general transcriptional factors. Methylation is believed to play a crucial role in repressing gene expression. It is believed that methylation DNA sequences in the promoter region block the binding of transcription factor binding. DNA methylation is observed in mediating cell differentiation, embryonic development and gene expression (Phillips 2008). Evidence has been found in studies that show that promoter methylation varies among cell types and more promoter methylation correlate with low or no transcription (Suzuki and Bird 2008).

Regulatory elements and molecules are classified as acting in *cis* or in *trans*, which means on the same side or on the opposite side, respectively. *cis*- regulatory elements, such as transcription-factor-binding sites in the enhancer, are located on the same DNA molecule. *trans*- regulatory elements diffuse in the cell and they don't act on specific copy of genes. Both alleles at one locus are equally likely to interact with a *trans*- acting element. General transcription factors are examples of *trans*- acting elements. *cis*- regulatory elements segregate with the gene while *trans*- acting factors segregate independently (Landry, Hartl et al. 2007).

Interspecific hybridization results in a vast reservoir of allelic variations, which may potentially contribute to phenotypical enhancement in the hybrids. Whether the allelic variations are related to the downstream phenotypic differences of interspecific hybrid is still an open question. The allelic combinations in a hybrid may result in interactions that alter expression profiles, new protein-protein interactions, or epistatic interactions. Previous studies for genetic variations of gene expression were focusing on expression quantitative trait loci (eQTL). However, allele-specific approach can directly assess *cis*- regulatory variations (Pastinen 2010). Early allelic expression studies focused on the interaction between alleles, which have brought up exciting observations. Take maize as an example, studies in comparison of expression levels documented high level of allelic variations in hybrids including genetic fragment contents and repetitive elements (Fu and Dooner 2002). Guo et al used ASE to study the relative expression of two alleles in maize hybrids. Since both alleles have the access to identical *trans*- acting factors, thus the biased allelic expression suggest *cis*- acting variation between alleles. They observed allelic expression bias in F1 hybrids was observed (11/15 genes), suggesting that *cis*- variation is present in many maize genes (Guo, Rupe et al. 2004). As many *cis*- and *trans*- gene regulation patterns were observed in hybrids, allele specific expression (ASE) analysis was brought up as a hotspot to discover the mechanisms underlying heterosis (Wittkopp, Haerum et al. 2004).

1.3 Next-generation Sequencing

RNA-Seq, also called whole transcriptome shotgun sequencing (WTSS), is a transcriptomic profiling approach which employs deep-sequencing technologies to reveal a snapshot of RNA presence and quantity from a genome at a given time. The transcriptome is the complete set of transcripts in a cell. Transcriptome characterization is essential for understanding the gene structure and functional elements interruption in differential developmental stages and in response to stresses. The transcriptomics is aiming at cataloguing all kinds of transcripts, including mRNA, non-coding RNA and small RNAs, in order to identify the start site, 5' and 3' ends and post-transcriptional modifications under different conditions. RNA-Seq can be performed using different sequencing strategies including 454 pyrosequencing, Illumina (Solexa) sequencing, SOLiD sequencing etc.

To date, the most widely used is Illumina sequencing. In this method, cDNA molecules are attached to the primers on a slide, and then are amplified into local colonies through "Bridged" amplification. Four types (adenine, cytosine, guanine and thymine) of reversible terminate bases are added, each of them are fluorescently labeled with a different color and attached with a blocking group. The four bases compete for binding sites on the template DNA and non-incorporated labeled ones are washed away. A laser is used to excite the dyes and a photograph is taken for record. A chemical deblocking step is then used to remove the 3' terminal blocking group and the dye in a single step. The process is repeated until the full DNA molecule is sequenced.

1.4 Research Purpose

The rich allelic variants in interspecific hybrids make them good models for ASE analysis. The purpose of this study is to gain better understanding of the mechanism underlying heterosis via ASE analysis of interspecific catfish hybrids. I plan to investigate ASE pattern using RNA-Seq datasets generated from interspecific catfish hybrids, with the objective to determine the ASE genes and pathways in which they are involved. The study will investigate the following research questions: Are there any ASE in the interspecific catfish hybrids? If so, to what extent ASE is involved? What are the allelic-specifically expressed genes (ASE-genes)? What are the parent-of-origins of the ASE alleles? Is there any preferentially expression from one parent?

References

Adams KL, Cronn R, et al. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proceedings of the National Academy of Sciences 100(8): 4649-4654

Anderson E (1948) Hybridization of the habitat. Evolution: 1-9

Anderson E and Hubricht L (1938) Hybridization in Tradescantia. III. The evidence for introgressive hybridization. American Journal of Botany 25: 396–402

Argue BJ, Arce SM, et al. (2002) Selective breeding of Pacific white shrimp (*Litopenaeus vannamei*) for growth and resistance to Taura Syndrome Virus. Aquaculture 204(3): 447-460

Baack EJ and Rieseberg LH (2007) A genomic view of introgression and hybrid speciation. Current opinion in genetics & development 17(6): 513-518

Barton NH and Hewitt G (1985) Analysis of hybrid zones. Annual review of Ecology and Systematics: 113-148

Bell G (1997) Selection: the mechanism of evolution. Springer Science & Business Media.

Bondari K (1980) Cage performance and quality comparisons of Tilapia and divergently selected channel catfish [Ictalurus punctata]Proceedings of the Annual Conference Southeastern Association of Fish and Wildlife Agencies.

Bondari K (1983) Response to bidirectional selection for body weight in channel catfish. Aquaculture 33(1): 73-81

Bruton MN (1996) Alternative life-history strategies of catfishes. Aquatic Living Resources 9(S1): 35-41

Dowling TE and Secor CL (1997) The role of hybridization and introgression in the diversification of animals. Annual review of Ecology and Systematics: 593-619

Dunham RA and Masser MP (2012) Production of hybrid catfish. Southern Regional Aquaculture Center.

Dunham RA and Smitherman RO (1985) Improved growth rate, reproductive performance, and disease resistance of crossbred and selected catfish from AU-M and AU-K lines. Circular-Alabama Agricultural Experiment Station (USA)

Dunham RA, Smitherman RO, et al. (1983) Relative tolerance of channel x blue hybrid and channel catfish to low oxygen concentrations. The Progressive Fish-Culturist 45(1): 55-57

Fu H and Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. Proceedings of the National Academy of Sciences 99(14): 9573-9578

Gjedrem T (1983) Genetic variation in quantitative traits and selective breeding in fish and shellfish. Aquaculture 33(1): 51-72

Gjedrem T (1985) Improvement of productivity through breeding schemes. GeoJournal 10(3): 233-241

Gjedrem T and Baranski M (2010) Selective Breeding in Aquaculture: an Introduction: An Introduction. Springer Science & Business Media.

Guo M, Rupe MA, et al. (2004) Allelic variation of gene expression in maize hybrids. The Plant Cell Online 16(7): 1707-1716

Harrison RG (1990) Hybrid zones: windows on evolutionary process. Oxford surveys in evolutionary biology 7: 69-128

Kause A, Ritola O, et al. (2005) Genetic trends in growth, sexual maturity and skeletal deformations, and rate of inbreeding in a breeding programme for rainbow trout (*Oncorhynchus mykiss*). Aquaculture 247(1): 177-187

Landry C, Hartl D, et al. (2007) Genome clashes in hybrids: insights from gene expression. Heredity 99(5): 483-493

Langdon C, Evans F, et al. (2003) Yields of cultured Pacific oysters *Crassostrea gigas* Thunberg improved after one generation of selection. Aquaculture 220(1): 227-244

Masser M and Dunham R (1998) Production of hybrid catfish. Southern Regional Aquaculture Center College Station, Texas.

Mayr E (1972) Sexual selection and natural selection. Sexual selection and the descent of man: 87-104

Okamoto N, Tayama T, et al. (1993) Resistance of a rainbow trout strain to infectious pancreatic necrosis. Aquaculture 117(1): 71-76

Pastinen T (2010) Genome-wide allele-specific analysis: insights into regulatory variation. Nature Reviews Genetics 11(8): 533-538

Paun O, Fay MF, et al. (2007) Genetic and epigenetic alterations after hybridization and genome doubling. Taxon 56(3): 649

Pawson HC (1957) Robert Bakewell. Pioneer livestock breeder. Robert Bakewell Pioneer livestock breeder

Phillips T (2008) The role of methylation in gene expression. Nature Education 1(1): 116

Schwenk K, Brede N, et al. (2008) Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. Philosophical Transactions of the Royal Society B: Biological Sciences 363(1505): 2805-2811

Seehausen O (2004) Hybridization and adaptive radiation. Trends in ecology & evolution 19(4): 198-207

Smitherman RO, Dunham RA, et al. (1983) Review of catfish breeding research 1969–1981 at Auburn University. Aquaculture 33(1): 197-205

Soltis DE and Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. Trends in Ecology & Evolution 14(9): 348-352

Storset A, Strand C, et al. (2007) Response to selection for resistance against infectious pancreatic necrosis in Atlantic salmon (*Salmo salar L.*). Aquaculture 272: S62-S68

Suzuki MM and Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nature Reviews Genetics 9(6): 465-476

Tate JA, Ni Z, et al. (2006) Evolution and expression of homeologous loci in Tragopogon miscellus (Asteraceae), a recent and reciprocally formed allopolyploid. Genetics 173(3): 1599-1611

Tave D, Mcginty AS, et al. (1981) Relative harvestability by angling of blue catfish, channel catfish, and their reciprocal hybrids. North American Journal of Fisheries Management 1(1): 73-76

Thodesen J, Grisdale-Helland B, et al. (1999) Feed intake, growth and feed utilization of offspring from wild and selected Atlantic salmon (*Salmo salar*). Aquaculture 180(3): 237-246

Wellborn TL (1988) Channel catfish: life history and biology. SRAC publication (USA)

Wittkopp PJ, Haerum BK, et al. (2004) Evolutionary changes in cis and trans gene regulation. Nature 430(6995): 85-88

Yant R, Smitherman R, et al. (1975) Production of hybrid (blue x channel) catfish and channel catfish in pondsProceedings Annual Conference Southeast Association of Game and Fish Commissioners. pp. 86-91.

Chapter 2: Review of Literature

2.1 Genetic Markers

A genetic marker refers to a fragment of DNA with certain location within the genome, which often serves as a landmark for tracing a certain region of DNA. Mutations happen to all organisms as a result of normal cellular operations or environmental interactions, which lead to genetic variations. Genetic markers can be used in DNA fingerprinting, linkage mapping, parentage identification and measurement of genetic diversity etc. Genetic variations on the DNA levels include the following scenarios: base substitution (SNPs), insertions or deletions of nucleotide sequences (indels), inversion of a DNA fragment and rearrangement of DNA segment (Liu and Cordes 2004). In history, a variety of genetic markers have been applied in the field of aquaculture including: allozyme markers, mitochondrial DNA (mtDNA) markers, restriction fragment length polymorphism (RFLP), randomly amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), microsatellite, microsatellite (single tandem repeat, SSR) single nucleotide polymorphism (SNP), and expressed sequence tag (EST) markers.

2.1.1 Microsatellites and SNPs

Although being popular for a quite long time, RFLP, RAPD and AFLP are less frequently used now. In recent years, microsatellite and SNP markers are more commonly used. Microsatellites, also called Simple Sequence Reads (SSRs) or Simple Tandem Repeats (STRs), are repeating sequences of 2-6 base pairs of DNA. Microsatellites have been found to be abundant in all species. In fish, it is estimated that microsatellites occurs every 10 kb (Wright 1993). Microsatellite polymorphism can be differentiated by size variation due to different number of repeats at a given locus. Microsatellites have been used increasingly in aquaculture species for over 10 years since their elevated polymorphic information content (PIC), co-dominant expression, Mendelian inheritance, genomic abundance and broad distribution throughout the genome (Liu and Cordes 2004). However, large scale of microsatellite analysis is labor intensive. Each microsatellite locus has to be identified with flanking region sequences to design of PCR primers. Gel electrophoresis, usually polyacrylamide gel electrophoresis, is applied to separate the bands generated by different repeats. Different alleles can be told by band position at each locus.

SNP markers are caused by point mutations that give rise to different alleles containing alternative bases at a given position. A SNP within a locus may contain as many as four alleles, including A (adenosine), T (thymine), G (guanine) and C (cytosine). Most SNPs are usually restricted to two alleles and thus been regarded as bi-allelic. Like microsatellites, SNP markers are inherited as co-dominant markers. Although their PIC is not as high as microsatellites, SNPs have a much higher abundance and are more widely distributed across the genomes. SNPs are often used for characterize specific genomic locations as well as genome-wide analysis. Liu and Cordes summarized traditional methods and advanced platforms for SNP genotyping (Liu and Cordes 2004). Traditional methods includes direct sequencing, single base sequencing, allele-specific oligonucleotide (ASO), denaturing gradient gel electrophoresis (DGGE), single strand conformational polymorphism assays (SSCP) and ligation chain reaction (LCR). Advanced platforms include pyrosequencing, Taqman allelic discrimination, real-time PCR, microarray and RNA-Seq (Vignal, Milan et al. 2002).

2.1.2 SNP Identification and Applications in Catfish

Large scale of SNP identification in catfish has been conducted years ago, using technologies such as EST (expressed sequence tags) analysis, Sequenom MassARRAY, genotyping-by-sequencing, RNA-Seq. In recent years, a large quantity of SNPs has been identified in catfish. Wang et. al identified more than 33,000 putative SNPs from catfish ESTs (Wang, Sha et al. 2008). They mined SNPs from NCBI dbEST database including both channel and blue catfish ESTs. After SNP identification, Illumina Bead Array was used to verify the SNPs using catfish individuals. Another 48,000 high-quality catfish SNPs were identified from over 3000,000 putative SNPs using EST sequences (Wang, Peatman et al. 2010). These 438,321 newly sequenced ESTs were generated from 4blue catfish and 8 channel catfish libraries. This was the first genome-wide sequencing project on catfish, it was estimated that about 50% of the total catfish genes were identified from this study. These results allowed evolutionary conservation analysis of catfish with other teleost fish as well as other higher vertebrates.In blue catfish, genotyping by sequencing was used for SNP discovery (Li, Waldbieser et al. 2014). Individuals from domesticated and wild populations were used in this study. Sequenom MassARRAY was used for SNP validation.

The next-generation sequencing technology has also been applied to identify catfish SNPs. in 2011, Liu et. al identified species-specific markers for channel and blue catfish using RNA-Seq technology (Liu, Zhou et al. 2011). After *de novo* assembly, mapping and quality controls, a total of 340,000 channel catfish intra-specific SNPs, 366,269 blue catfish intra-specific SNPs, and over 420,000 common SNPs were identified. These SNPs were very useful for following studies. RNA-Seq analysis was also applied to identify disease related genomic regions. Wang et. al developed a bulked segregant RNA-Seq (BSR-Seq) analysis to identify the genomic locations which were responsible for ESC disease (Wang, Sun et al. 2013). ESC challenge experiment was conducted with multiple families of catfish, based on the time point of fish response, they were classified as susceptible and resistant fish. A total of 56,419 SNPs located on 4,304 unique genes were identified as significant SNPs between susceptible and resistant fish. Further SNP analysis allowed differentiating variation source as caused by segregation or allele-specific expression.

A catfish 250K SNP array was developed using Affymetrix Axiom genotyping technology with gene-associated SNPs, anonymous genomic SNPs and inter-specific SNPs (Liu, Sun et al. 2014). Over 640K high-quality SNPs were obtained and 250,113 were finalized on the array. The performance of the SNP array was then evaluated using wild channel catfish and hybrid catfish families. This array was very useful for genome-wide association studies (GWAS), fine QTL mapping, high-density linkage map construction, haplotype analysis and whole genome-based selection. A high density linkage map was developed using more than 50,000 SNPs, with their genotype screened by this 250K SNP array with three large families (Li, Liu et al. 2015). A total of 54,342 SNPs were placed on the linkage map. Integration of BAC-based physical map and linkage map allowed 86% of the whole genome scaffolds to be located onto the 29 linkage groups. This high density linkage map was extremely helpful for searching genomic regions related to disease and stress responses as well as genomic comparative studies. The 250K SNP array and linkage map were used for a genome-wide association study in catfish for columnaris disease resistance (Geng, Sha et al. 2015). Using marker analysis, this study found one significant region on linkage group 7, three suggestively QTL regions on linkage group7, 12 and

14 to be associated with columnaris resistance. Genes on these QTL regions were later characterized and were found to be mainly attributes to the PI3K pathway.

2.2 Allele-specific expression studies

An allele is one of the alternative forms of a gene at a particular location on a chromosome. Most living creatures on the Earth are diploids, which mean that each individual has two set of genetic materials and each locus has two alleles. Allele-specific expression refers to the phenomenon that the two alleles are unequally expressed in an organism or within specific tissues. Classically, ASE is considered to be associated with the epigenetic phenomena of X-chromosome inactivation and genomic imprinting. Genomic imprinting is an epigenetic phenomenon that certain genes are expressed in a parent-of-origin-specific pattern. Paternal imprinting means the allele inherited from the father is imprinted, or to say silenced, and only the allele from the mother is expressed. Similarly, maternal imprinting means the allele inherited from the mother is imprinted and only the allele from the father is expressed. Genomic imprinting is independent of Mendelian inheritance. It is an epigenetic process involves DNA methylation and histone modification without changing the genetic sequence (Wood and Oakey 2006). X-inactivation is one example of genomic imprinting. It is a process that one of the two copies of the X chromosomes in female mammals is inactivated. The inactive X chromosome is silenced by its being packaged in such a way that it has a transcriptionally inactive structure called heterochromatin. This process prevents female mammals from having twice as many X chromosome gene products as males (Wood and Oakey 2006). It was later observed that ASE was also a common phenomenon in non-imprinted autosomal genes. A variety of ASE studies have been conducted using different methods in different organisms and it is suggested that ASE is heritable (Yan, Yuan et al. 2002; Lo, Wang et al. 2003).

ASE also plays a role in regulation of gene expression. Gene expression is complex and is influenced by *cis*- and *trans*- acting elements, as well as epigenetic variations. The genetic variations that affect gene expression were largely focusing on expression quantitative trait loci (eQTL) mapping. eQTL mapping means mapping of genomic loci that regulate expression levels of mRNAs. A high level of success has been achieved using eQTL mapping for the characterization of gene expression patterns. For example, previous eQTL analysis indicates that ASE among different transcripts within cell lines can affect up to 30% of loci and *cis* regulation can affect ~30% of gene expression at the population level (Ge, Pokholok et al. 2009). However, direct assessment of the *cis*- regulatory variations requires ASE analysis. ASE analysis is able to distinguish between *cis*- and *trans*- regulations of gene expression. A gene that is under complete *trans*- regulation has a similar expression pattern of both alleles in the hybrid, while a gene that is under complete *ci*s- regulation exhibits unequal expression of the two alleles in the hybrids (Wittkopp, Haerum et al. 2004). Early allelic analyses were focusing on restricted individual loci, but genome-wide ASE patterns are now accessible with recent advances in genomic technologies.

2.2.1 Polymorphism-directed Approach

Pastinen summarized two general approaches for genome-wide allele-specific analysis: polymorphism-directed approach and global approach. The rapid characterization of genomic variants provides opportunity to detect allelic variants. On one hand, the polymorphism-directed approach increases the information density of the genomic data by using polymorphic genomic variants; on the other hand, these polymorphic cites can be mapped back to the genome sequences, which provide control for the technical biases in quantifying the allele ratios (Pastinen 2010).

Genome-wide genotyping arrays provide a convenient way to assessing ASE in expressed transcripts at a relatively low cost (Ge, Pokholok et al. 2009). The basic principle of SNP arrays are the same as the DNA microarray including DNA hybridization, fluorescence microscopy and solid surface DNA capture. It has widely been employed for eQTL analysis to characterize ASE. However, the coverage of ASE site is the main concern for this method because current standard SNP arrays contain only a small subset of polymorphic regulatory elements (<5%) (Pastinen 2010). It is reported that the use of unspliced RNA can cover more regulatory sequences than the use of mature RNA (Gimelbrant, Hutchinson et al. 2007). Genome-wide genotyping arrays are still considered as important method to access ASE pattern, and it is believed that with the ongoing improvements, this method will be able to characterize ASE in nearly all human genes (Ge, Pokholok et al. 2009).

Padlock probes are also used to capture known exonic polymorphism on a large scale. Padlock probes are single strand DNA molecules with two 20-nucleotide segments complementary to the targeted DNA sequence and 40-nucleotied linker sequence. When hybridized with the target, the padlock probes become circularized. Padlock probes are very useful for detecting known DNA sequences with high specificity since it leaves no gaps upon hybridization. Because the padlock probes are so precise that it can distinguish alleles with small difference, allele-specific padlock probes have been used to detect genomic DNA or cDNA variations for gene expression analysis in response to disease and drug associated studies (Banér, Isaksson et al. 2003). This method allows researchers to analyze thousands of targeted sites in the genome.

2.2.2 Global Approach

The rapid next generation sequencing technology allows researchers to characterize allelic variations in the transcripts at single-base resolution. This approach relies on the RNA-Seq technology. RNA-Seq reads provide digital estimation of allele frequencies at the polymorphic sites. Simple statistical approaches, such as binomial tests, allow the detections of biased allelic expression, and the power of these tests depends on the read number (Pastinen 2010). Although some researchers claim that there are unequal expression at specific sites and the biases are towards the reference alleles presented in the reference genome (Fontanillas, Landry et al. 2010; Heap, Yang et al. 2010), RNA-Seq is still considered as an important approach since it is the only method that provide both current allelic and total expression data. The global approach does not require prior knowledge about the genome sequence or polymorphism, instead, it can increase the information content by exploring ASE. Thus, this approach is ideal for non-model organisms without rich background genetic information.

RNA-Seq based ASE analysis has been widely applied to human and many other organisms. In human, early ASE studies were performed together with eQTL analysis. Heap et al used RNA-Seq for ASE analysis and eQTL analysis to understand disease-associated genetic variants expression in human. They reported two ASE-SNP sites that linked previous genome-wide association (GWA) results in gene expression. This study provide evidence for the power of ASE analysis in validating genomic analysis and demonstrated a method to estimate ASE with SNPs (Heap, Yang et al. 2010). Another example of human ASE study is conducted within Caucasian population. In this study, they found that analysis of SNPs from HapMap could lead to a larger discovery than arrays and ASE analysis allowed the identification of rare eQTLs (Montgomery, Sammeth et al. 2010).

ASE studies are also widely applied to hybrid plants, especially crops such as rice and maize, to explore the potential regulatory mechanism underlying heterosis. In rice, global ASE patterns have been described in developmental stages and with methylation status. Transcriptomic SNPs were analyzed for assessing ASE in super-hybrid rice at two developing stages. In this study, 17% of identified transcriptomic SNPs showed ASE pattern, suggesting *trans*- regulation mediated gene expression in the hybrid rice. They also observed a higher percentage of transition SNP over transversion SNPs in the hybrid rice, suggesting the existence of methylation (Zhai, Feng et al. 2013). The interaction of transcriptome and methylome in hybrid rice was analyzed by using both bisulfite sequencing (BS-Seq) and RNA-Seq. In this study, the authors identified the epimutation status between parents and hybrids and provide evidence that ASE genes were associated with their methylation status (Chodavarapu, Feng et al. 2012). Rice ASE has also been well characterized with transcriptomic SNP analysis using multiple reciprocal hybrids and parental lines. Global SNP analysis revealed that ~3% genes exhibit monoallelic expression, ~23% gene exhibit preferentially allelic expression and ~72% showed biallelic expression. The authors observed that ASE accounted for 79.8% of the genes displaying more than a 10-fold expression difference between F1hybrids and their corresponding parents, and almost all (97.3%) F1-specific genes (Song, Guo et al. 2013).

In RNA-Seq studies, *cis*- and *trans*-regulatory variations can be assessed by comparing the expression patterns in parents and hybrids. The general rules are summarized as follows: *cis*-

regulation is identified if the allelic expression patterns are the same between parents and hybrids. (A: a (parent) = A: a (hybrid) ≠ 50:50); *trans*-regulation is identified the parental expression are biased but the expression in hybrids are equal. (A: a (parent) ≠ 50:50, A: a (hybrid) = 50:50); if there’s both *cis*- and *trans*-regulation at one site, biased allelic expression will be observed in both parents and hybrids but the biased pattern are not the same. (A: a (parent) ≠ A: a (hybrid), A: a (parent ≠ 50:50, A: a (hybrid) ≠ 50:50) (Wittkopp, Haerum et al. 2004; Zhuang and Adams 2007). The allele frequencies in parents are usually measured using “artificial hybrids”, which is a mixture of equal amount of two parental RNAs.

2.2.3 ASE Studies in Aquaculture Species

An early study aiming at detecting hybridization between bream, roach and rudd applied allele-specific amplification (ASM) of nuclear (ITS1) and mitochondrial (cytochrome b) markers. ASM method used PCR amplification with allele-specific primers followed by gel electrophoresis. The difference of marker sizes provided evidence for hybridization identification (Wyatt, Pitts et al. 2006). This study validated the existence of both parental alleles in hybrids but provide no evidence in quantify the allelic expression. Early studies in the field of fish ASE were focusing on limited genes such as housekeeping genes, tissue-specific genes or genes played particular roles. ASE analyses were conducted in a hybrid carp species *Squalius alburnoides* in diploid and triploid fish. Pala et al screened for ASE in both diploid and triploid *S. alburnoides* individuals from different geographic locations. They applied restriction enzymatic digestion analysis to distinguish the fragment polymorphism between P and A genomes for 6 genes. The authors observed a ballelic expression for all the genes in diploids. In different tissues of triploids, 4 genes constantly exhibited ASE of A allele, 1 gene constantly showed biallelic expression and 1 gene showed tissue-specific ASE (Pala, Coelho et al. 2008). ASE pattern in *S. alburnoides* was later explored using PCR amplification followed by sequencing. The authors used 3 housekeeping genes to characterize the expression pattern of 20 pooled individuals. Their results further confirmed the biallelic expression in diploids but they demonstrated that the gene expression of the 3 gene was also biallelic in triploids, which is different from the previous study (Matos, Sucena et al. 2011). ASE patterns have been characterized in stickleback during the studies of two signaling pathways. Miller et al demonstrated the ASE pattern of the gene encoded for ligand of tyrosine-kinase receptor (Kitlg) in different tissues using RT-PCR and the Kitlg 5'-UTR size difference analysis. They observed that Kitlg gene showed biallelic expression in some tissues, but it showed allele preferential expression in other tissues and the expression level of freshwater Kitlg allele was expressed significantly lower than the marine Kitlg allele (Miller, Beleza et al. 2007). ASE in stickleback was also observed in thyroid-stimulating hormone-b2 gene (TSHb2) using pyrosequencing analysis. The result showed that there was allelic preferential expression towards the marine allele in all of the 20 stickleback hybrids (Kitano, Lema et al. 2010). Both of stickleback ASE studies provide evidence for *cis*-regulation contributes to the differential gene expression.

ASE pattern has been systematically characterized in medaka diploids and triploids. ASE pattern has been well characterized in diploid medaka in 11 selected genes using qRT-PCR technology. The researchers designed allele-specific primers for each parental genome and common primers that amplified both alleles to access the allelic and total expression of each gene. The use of reciprocal hybrids suggested that there were considerable ASE difference between them (Murata, Oda et al. 2012). A global ASE study was conducted of triploid medaka using RNA-Seq technology. The authors quantified ASE of two triploid individuals based on transcriptomic SNP analysis. They reported that 18% of total genes exhibited ASE and the most of them were located on 4 chromosomes (Garcia, Matos et al. 2014). This study is the lasted published article to date about comprehensive fish ASE analysis.

Transcriptomic SNP analysis has been applied to access ASE in platyfish, *Xiphophorus*, interspecific hybrids. Young hybrid platyfish, along with two parental lines were subjected to RNA-Seq and transcriptomic SNPs were identified for ASE analysis. A total of 27 ASE-genes were identified in the interspecific hybrids in a wide functional range (Shen, Catchen et al. 2012). In catfish, transcriptomic SNP analysis has been previously applied for bulk segregant RNA-Seq analysis to access genes responsible for ESC resistance in F2 backcross progenies. The allele ratio of each SNP sties was taken into consideration for bulk segregation analysis. This study is a successful example of applying transcriptomic SNPs in characterizing specific trait-associated genes in catfish.

References

Banér J, Isaksson A, et al. (2003) Parallel gene analysis with allele‐specific padlock probes and tag microarrays. Nucleic acids research 31(17): e103-e103

Chodavarapu RK, Feng S, et al. (2012) Transcriptome and methylome interactions in rice hybrids. Proceedings of the National Academy of Sciences 109(30): 12040-12045

Fontanillas P, Landry CR, et al. (2010) Key considerations for measuring allelic expression on a genomic scale using high‐throughput sequencing. Molecular ecology 19(s1): 212-227

Garcia TI, Matos I, et al. (2014) Novel Method for Analysis of Allele Specific Expression in Triploid *Oryzias latipes* Reveals Consistent Pattern of Allele Exclusion. PloS one 9(6): e100250

Ge B, Pokholok DK, et al. (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nature genetics 41(11): 1216-1222

Geng X, Sha J, et al. (2015) A genome-wide association study in catfish reveals the presence of functional hubs of related genes within QTLs for columnaris disease resistance. BMC genomics 16(1): 196

Gimelbrant A, Hutchinson JN, et al. (2007) Widespread monoallelic expression on human autosomes. Science 318(5853): 1136-1140

Heap GA, Yang JH, et al. (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Human molecular genetics 19(1): 122-134

Kitano J, Lema SC, et al. (2010) Adaptive divergence in the thyroid hormone signaling pathway in the stickleback radiation. Current Biology 20(23): 2124-2130

Li C, Waldbieser G, et al. (2014) SNP discovery in wild and domesticated populations of blue catfish, Ictalurus furcatus, using genotyping ‐ by ‐ sequencing and subsequent SNP validation. Molecular ecology resources 14(6): 1261-1270

Li Y, Liu S, et al. (2015) Construction of a high-density, high-resolution genetic map and its integration with BAC-based physical map in channel catfish. DNA Research 22(1): 39-52

Liu S, Sun L, et al. (2014) Development of the catfish 250K SNP array for genome-wide association studies. BMC research notes 7(1): 135

Liu S, Zhou Z, et al. (2011) Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC genomics 12(1): 53

Liu Z and Cordes J (2004) DNA marker technologies and their applications in aquaculture genetics. Aquaculture 238(1): 1-37

Lo HS, Wang Z, et al. (2003) Allelic variation in gene expression is common in the human genome. Genome research 13(8): 1855-1862

Matos I, Sucena É, et al. (2011) Ploidy mosaicism and allele-specific gene expression differences in the allopolyploid Squalius alburnoides. BMC genetics 12(1): 101

Miller CT, Beleza S, et al. (2007) cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. Cell 131(6): 1179-1189

Montgomery SB, Sammeth M, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464(7289): 773-777

Murata Y, Oda S, et al. (2012) Allelic expression changes in Medaka (*Oryzias latipes*) hybrids between inbred strains derived from genetically distant populations. PloS one 7(5): e36875

Pala I, Coelho MM, et al. (2008) Dosage compensation by gene-copy silencing in a triploid hybrid fish. Current Biology 18(17): 1344-1348

Pastinen T (2010) Genome-wide allele-specific analysis: insights into regulatory variation. Nature Reviews Genetics 11(8): 533-538

Shen Y, Catchen J, et al. (2012) Identification of transcriptome SNPs between *Xiphophorus* lines and species for assessing allele specific gene expression within F1 interspecies hybrids. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 155(1): 102-108

Song G, Guo Z, et al. (2013) Global RNA sequencing reveals that genotype-dependent allele-specific expression contributes to differential expression in rice F1 hybrids. BMC plant biology 13(1): 221

Vignal A, Milan D, et al. (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution 34(3): 275-306

Wang R, Sun L, et al. (2013) Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish. BMC genomics 14(1): 929

Wang S, Peatman E, et al. (2010) Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. Genome biology 11(1): R8

Wang S, Sha Z, et al. (2008) Quality assessment parameters for EST-derived SNPs from catfish. BMC genomics 9(1): 450

Wittkopp PJ, Haerum BK, et al. (2004) Evolutionary changes in cis and trans gene regulation. Nature 430(6995): 85-88

Wood AJ and Oakey RJ (2006) Genomic imprinting in mammals: emerging themes and established theories. PLoS genetics 2(11): e147

Wright J (1993) DNA fingerprinting of fishes. Biochemistry and molecular biology of fishes 2: 57-91

Wyatt P, Pitts C, et al. (2006) A molecular approach to detect hybridization between bream Abramis brama, roach Rutlius rutilus and rudd Scardinius erythrophthalmus. Journal of Fish Biology 69(sa): 52-71

Yan H, Yuan W, et al. (2002) Allelic variation in human gene expression. Science 297(5584): 1143-1143

Zhai R, Feng Y, et al. (2013) Identification of transcriptome SNPs for assessing allele-specific gene expression in a super-hybrid rice Xieyou9308. PloS one 8(4): e60668

Zhuang Y and Adams KL (2007) Extensive allelic variation in gene expression in Populus F1 hybrids. Genetics 177(4): 1987-1996

Chapter3: Methodology

3.1 RNA-Seq

RNA-Seq data used in this study was obtained from a previous study using F_1 interspecific hybrid catfish (Liu, Wang et al. 2013). In that study, the F_1 hybrid catfish were generated by mating a female channel catfish with a male blue catfish. A total of 300 one-year old F1 hybrid catfish fingerlings were used for experiment, 45 of which were randomly selected as control group without treatment. The rest 250 fish were subjected to a chronical heat stress challenge experiment. The first and last 45 individuals showing loss of equilibrium (LOE) were classified as intolerant and tolerant groups respectively. Liver and gill tissues were collected for RNA extraction from control, intolerant and tolerant fish.

RNA sequencing of the three group fish was performed on an Illumina HiSeq2000 instrument for 100 bp paired-end reads. Sequencing was conducted commercially at HudsonAlpha Genomic Services Lab (Huntsville, AL, USA). ABySS (version 1.3.0) (http://www.bcgsc.ca/platform/bioinfo/software/abyss) and Trans-ABySS (version 1.2.0) (http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss) software were used for transcriptome assembly. Gene annotation was conducted by BLASTX program using the assembled contigs against the National Center for Biotechnology Information (NCBI) RefSeq zebrafish protein database and the UniProt-SwissProt (UniProt) database with a cutoff E-value of 1E-6. The unannotated contigs were annotated based on BLASTX searches against nonredundant protein (Nr) database (Liu, Wang et al. 2013). The RNA-Seq short reads generated from the 45 control fish, the assembled transcriptome contigs, and their annotations and are used for ASE analysis in the present study.

3.2 Reference Mapping

Sequence mapping was carried out using CLC Genomics workbench (version 5.5.2; CLC bio, Aarhus, Denmark). Before mapping, raw sequence reads were trimmed to remove adaptor sequences, ambiguous nucleotides (number of "N" > 2), extreme short reads (< 30 bp) and low quality sequences (Quality score < 20) using CLC Genomics Workbench.

CLC genomics workbench used Phred quality scores to characterize the quality of DNA sequences. Phred quality scores (Q) were defined as a property which was logarithmically related to the base-calling error probabilities (P): $P = 10^{\frac{Q}{-10}}$. They were assigned to each nucleotide base call in automated sequencer traces. If Phred assigned a quality score of 20 to a base, the chances that this base was called incorrectly were 1 in 100, which meant 99% base call accuracy (Ewing and Green 1998; Ewing, Hillier et al. 1998). A new value was calculated for every base: *Limit* – P. This value would be negative for low quality bases, where a higher chance of incorrect base call would be. The Workbench calculated the running sum for every base. All negative sums would be converted to zero. The part of the sequence to be retained after trimming was the region between the first positive value of the running sum and the highest value of the running sum. Everything before and after this region would be trimmed off.

The clean reads from each group were then aligned with the reference assembly separately. The mapping parameters were set as mismatch cost of 2, deletion cost of 3 and insertion cost of 3. Mapping sequence shared ≥ 95% similarity with the reference sequence and over ≥ 90% of their length were included in the alignment. The mapping outputs were converted into BAM format for further analysis.

3.3 SNP Identification

SNPs were identified from the pooled data from all the three groups using SAMtools (version 0.1.18) and PoPoolation2 (version 1.201) (Li, Handsaker et al. 2009; Kofler, Pandey et al. 2011). First, ambiguously mapped reads were removed using SAMtools with the command “samtools view -q 20 -bu XXX.bam”: the option “-q 20” meant only sequences with quality score higher than 20 are kept; the option “-bu” meant the input file was in unzipped BAM format; “XXX.bam” were the names of BAM files from each group, each of control, intolerant and tolerant group was filtered separately. Then BAM file were sorted by genomic locations using the command “samtools sort XXX.bam XXX.sorted.bam”, and this required the use of “pile-up" function later to match reads within a specific genomic location. Indexing was conducted followed sorting using command “samtools index XXX.sorted.bam”. This enabled tools, including SAMtools itself, and other genomic viewers to perform efficient random access on the BAM file.

The next step was to identify the transcriptomic variants. SAMtools mpileup command “samtools mpileup -f reference _assembly.fa control.sorted.bam intolerant.sorted.bam tolerant.sorted.bam > CIT.mpileup” was used to calculate the genotype likelihoods supported by the aligned reads in our sample. The mpileup command automatically scaned every position supported by an aligned read, computed all the possible genotypes supported by these reads, and then computed the probability that each of these genotypes was truly present in our sample. The option “-f” meant the input reference file was in unzipped fasta format. The output file

"CIT.mpileup" was a merged file with all the information of these variants in three groups, and both groups-specific and shared variants were included. Note that if there was any file that was in a different location, the file path would be added before the file name in the script. A synchronized file was then generated by a perl script provided in the PoPoolation2 toolkit. Synchronized files were the main input files for PoPoolation2. They basically contained the allele frequencies for every population at every base in the reference genome. Because synchronizing the mpileup file was quite time consuming, a Java multi-threading method, which was about 78x faster as the implementation in perl, was used with the command "java -ea -Xmx7g -jar <popoolation2-path>/mpileup2sync.jar --input CIT.mpileup --output CIT_java.sync --fastq-type sanger --min-qual 20 --threads 8". The option "--min-qual 20" again ensured the sequence quality of 20 or higher. The option "--threads 8" meant this script need to use 8 CPUs. A total of six columns were generated in the synchronized file: the first column was the reference contig ID; the second column was the position within the reference contig; the third column was the reference genotype; the fourth column to the sixth filed were allele frequencies of each group respectively.

Raw SNPs were identified using a perl provide by PoPoolation2 with command "perl <popoolation2-path>/snp-frequency-diff.pl --input CIT_java.sync--output-prefix CIT". This script created two output files: "_rc" file contained the major and minor alleles for every SNP; "_pwc" file contained the differences in allele frequencies for every pairwise comparison of the groups. I used "_rc" file for following analysis since it contained detail information for individual polymorphism sites that was suitable to be analyzed in various forms. A sample of SNP identification results from Popoolation 2 was shown in Table 1.

As shown in column 9 of Table 1, sometimes the minor allele genotype would be “N”. The genotype “N” meant the there was no other genotype except for the major allele at this site in that group. For example, for the SNP at 124 bp of contig k50:1050148, “NAN” in column 9 meant there were only one allele present in the control and tolerant groups, and the minor allele genotype in intolerant group was A. This meant that the SNP marker was only detected in intolerant group. This kind of SNPs was defined as group-specific SNPs. SNPs with the presence of both alleles in all three groups were defined as common SNPs.

Table 1. Example of SNP identification output from Popoolation 2.A total of 15 columns are presented in the result file: column 1 is the reference contig ID; column 2 is the SNP position on the reference contig; column 3 is the reference genotype of this positon on the contig; column 4 is the number alleles detected for the SNP; column 5 is the genotypes of the alleles detected in the SNP; column 6 is deletion sum; column 7 is SNP type; column 8 is the genotype for the major alleles in each group. For example, “ACA” means the major genotype for control group is A, for intolerant group is C ad for tolerant group is A; column 9 is the genotype for the minor alleles in each group; column 10 is number of major alleles versus the number of the total alleles in the control group; column 11 is number of major alleles versus the number of the total alleles in the intolerant group; column 12 is number of major alleles versus the number of the total alleles in the tolerant group; column 13 is number of minor alleles versus the number of the total alleles in the control group; column 14 is number of minor alleles versus the number of the total alleles in the intolerant group; column 15 is number of minor alleles versus the number of the total alleles in the tolerant group.

k50:1032492	347	T	2	T/C	0	pop	TTT	NCN	6/6	5/7	10/10	0/6	2/7	0/10
k50:1032492	352	A	2	A/G	0	pop	AAA	GNG	4/6	8/8	8/9	2/6	0/8	1/9
k50:1032492	373	T	2	T/C	0	pop	TTT	CCN	6/7	6/10	7/7	1/7	4/10	0/7
k50:1032492	394	G	2	G/A	0	pop	GGG	NNA	7/7	9/9	4/6	0/7	0/9	2/6
k50:1032492	401	A	2	A/T	0	pop	AAA	TTT	3/6	6/9	5/6	3/6	3/9	1/6
k50:1034216	444	A	2	A/G	0	pop	AAA	GGG	4/8	8/11	5/10	4/8	3/11	5/10
k50:1040620	160	C	2	C/T	0	pop	CCC	TNT	39/41	51/51	54/55	2/41	0/51	1/55
k50:1040620	203	T	2	T/C	0	pop	TTT	CCN	37/38	23/24	38/38	1/38	1/24	0/38
k50:1043916u	142	G	2	G/A	0	pop	GGG	NNA	35/35	78/78	63/65	0/35	0/78	2/65
k50:1043916u	160	T	2	T/A	0	pop	TTT	NNA	47/47	88/88	73/75	0/47	0/88	2/75
k50:1043916u	203	G	2	G/A	0	pop	GGG	NNA	33/33	52/52	46/48	0/33	0/52	2/48
k50:1045191	271	G	2	G/A	0	pop	GGG	ANN	5/7	16/16	6/6	2/7	0/16	0/6
k50:1045990u	26	A	2	A/T	0	pop	AAA	NTN	81/81	91/93	67/67	0/81	2/93	0/67
k50:1045990u	34	C	2	C/A	0	pop	CCC	NAN	109/109	125/127	101/101	0/109	2/127	0/101
k50:1045990u	121	T	2	T/G	0	pop	TTT	GNN	172/175	182/182	166/166	3/175	0/182	0/166
k50:1048117	999	A	2	A/C	0	pop	AAA	CCC	5/9	25/38	20/30	4/9	13/38	10/30
k50:1050148	124	T	2	T/A	0	pop	TTT	NAN	58/58	82/84	86/86	0/58	2/84	0/86
k50:1050148	217	C	2	C/G	0	pop	CCC	NGN	53/53	64/66	58/58	0/53	2/66	0/58
k50:1055185	299	A	2	A/G	0	pop	AAA	NNG	15/15	10/10	15/17	0/15	0/10	2/17
k50:105944	56	G	2	G/A	0	pop	GGG	AAA	15/22	16/26	13/25	7/22	10/26	12/25
k50:1059646	356	C	2	C/A	0	pop	CCC	AAA	12/17	22/26	12/14	5/17	4/26	2/14
k50:1059646	1082	C	2	C/G	0	pop	CCC	GGG	9/10	14/17	14/20	1/10	3/17	6/20
k50:1032492	347	T	2	T/C	0	pop	TTT	NCN	6/6	5/7	10/10	0/6	2/7	0/10
k50:1032492	352	A	2	A/G	0	pop	AAA	GNG	4/6	8/8	8/9	2/6	0/8	1/9
k50:1032492	373	T	2	T/C	0	pop	TTT	CCN	6/7	6/10	7/7	1/7	4/10	0/7
k50:1032492	394	G	2	G/A	0	pop	GGG	NNA	7/7	9/9	4/6	0/7	0/9	2/6
k50:1032492	401	A	2	A/T	0	pop	AAA	TTT	3/6	6/9	5/6	3/6	3/9	1/6
k50:1034216	444	A	2	A/G	0	pop	AAA	GGG	4/8	8/11	5/10	4/8	3/11	5/10

3.4 SNP Filtering

Only common SNPs were selected for further analysis. This would help to exclude some errors arise from sequencing and increase the reliability for the SNPs I use. The information of SNPs from control group was used for following analyses. These allelic variants from control group reflected the actual expression status and had the potential to reveal the specific pattern in interspecific catfish hybrids.

Three factors that were important for excluding false SNPs caused by sequencing errors were set: 1) minimum read depth, 2) maximum read depth, and 3) minor allele read count. An optimal combination of these three factors was determined and used for screening quality SNPs. In order to identify reliable SNPs, further quality controls was applied as follows: 1) the alleles detected at each SNP site must not contain “N”; 2) each SNP must consist of only two alleles (allelic variants = 2); 3) the read number of minor allele at each SNP site ≥2; and 4) the total read number of alleles at each SNP site ≥6.

3.5 Identification of ASE-SNPs

The SNPs with imbalanced allele expression were identified by conducting two-tailed Fisher’s exact test. The hypothesis was to test whether there is a statistical difference between the number of reads for major allele and the number of reads for minor allele at each SNP site (Fisher 1922). Fisher’s exact test is performed using False Discovery Rate Calculator developed by Microsoft Research (http://research.microsoft.com/en-us/um/redmond/projects/MSCompBio/FalseDiscoveryRate/). The number of reads for major allele and the number of reads for minor allele at each SNP site, along with the mean number of two alleles, were subjected to the Fisher’s exact test. SNPs passed the Fisher’s exact test were set out in a text file with a False Discovery Rate (FDR) adjusted p-value (q value) added to the end of each row. SNPs with $q \leq 0.05$ were kept for further analysis.

Allele ratio for a SNP was calculated as read number of major allele divided by the read number of minor allele. Allele specifically expressed SNPs (ASE-SNPs) were identified if the allele ratio for a SNP was equal to or greater than 9. Since all the 45 fish in the control group were randomly selected from one family, according to Mendelian Inheritance Laws, the parental genotype and allele ratio in the offspring could be either of the following scenarios: 1) if the parental genotypes were AA × BB, the allele ratio in the offspring would be A:B = 1:1; 2) if the parental genotypes were AA × AB, the allele ratio in the offspring would be A:B = 3:1; 3) if the parental genotypes were AB × BB, the allele ratio in the offspring would be A:B = 1:3; 4) if the parental genotypes were AB × AB, the allele ratio in the offspring would be A:B = 1:1. In conclusion, the highest allele ratio in the offspring among all the parental genotypes was 3. Although there was possibility to miss some ASE-SNPs with the cutoff allele ratio of 9, once again, my protocol increased the reliability of the ASE-SNPs found.

3.6 Identification of the Parent-of-Origins

Parent-of-origins of the alleles were identified in two ways. One way was to BLAST the SNP sequences with a previously generated SNP database, which contained inter-specific SNP sites (Liu, Sun et al. 2014). Another way was to map the SNP sequences to the channel catfish and blue catfish genomes (Unpublished) and distinguish the heterozygous ones between the two species.

The previously generated SNP database contained four different SNPs: channel catfish-specific SNPs, blue catfish-specific SNPs, inter-specific catfish SNPs and anonymous genomic channel catfish SNPs. The first three kinds of SNPs were used for parent-of-origin analysis. Inter-specific catfish SNPs were SNPs present difference alleles between channel and blue catfish. They were very useful to identify the allele origins based on the genotype of each allele. The sequences of ASE-SNPs were extracted using command “fastacmd -d reference_assembly.fa -s contig_names -L start_point,end_point >> sequence_pool.fa” under Linux environment. The left and right 35 bp from the SNP site were included in the sequence. The pooled SNP sequences were used as input queries to BLAST against the SNP database and blue and channel catfish genomes. If the SNP sequences hit the SNP database or genomes were inverted, the alleles of the SNP site would be genotyped as the complement ones. For example, if a ASE-SNP sequence contained a A/G SNP hit a C/T SNP probe (Channel/Blue) sequence in SNP database inverted, the ASE-SNP sequence was read as T/C (both alleles were converted to the complementary ones). Then allele A was considered as of channel catfish origin and allele G is of blue catfish origin. Similar process was conducted using channel and blue catfish reference genomes.

3.7 Ontology Analysis of ASE-genes

The previously generated gene annotation list was used in the present study. The description and accession number were connected to the ASE-SNP containing contigs via Microsoft Accession. A gene was considered as an ASE-gene if a contig hit to this gene contained at least one ASE-SNP. The allele ratio of a gene was represented by the allele ratio of the ASE-SNP. If a contig contained multiple ASE-SNPs, the allele ratio of the gene was represented by the highest allele ratio of the ASE-SNPs.

The gene ontology (GO) analysis was a bioinformatics approach to unify the presentation of genes across all species. It aimed at controlling vocabulary which is structured as a directed acyclic graph, and described genes in an organism. Genes from any organism have been annotated to GO terms. So it also has a function of classification (Ashburner, Ball et al. 2000). The GO provided a hierarchically classification system of genes and gene products into different GO terms. These terms were grouped into three categories: molecular function (the molecular activity of a gene), biological process (the cellular or physiological pathways in which a gene involved) and cellular component (the cellular location of a gene product or where a gene functioned). Each gene could be annotated with multiple GO terms as to different categories and scales. GO could be used to functionally profile a set of genes generated from high-throughput experiments, to determine which GO terms appear more frequently than would be expected through enrichment analysis.

In this study, I applied Ontologizer (Version 2.0 http://compbio.charite.de/contao/index.php/ontologizer2.html) to analyze annotation-enriched GO terms. GO term enrichment analysis was defined as "a process for interpreting sets of genes making use of the GO system of classification, in which genes were assigned to a set of predefined bins depending on their functional characteristics". Ontologizer set two lists of genes: study set and population set. A study set contained genes that share some biological characteristic while a population was a larger list of genes, generally the whole set of gene list obtained from the experiment (Bauer, Grossmann et al. 2008). In my study, ASE-genes from the liver and the gill were used as study sets independently, and all the assembled contigs obtained from the RNA-Seq data were used as population set.

Before applied to the Ontologier, contigs of the ASE-genes were BLASTX against with Zebrafish protein database (Danio_rerio.GRCz10.pep.all.fa, ftp://ftp.ensembl.org/pub/release-80/fasta/danio_rerio/pep/). Top hits with $p \leq e^{-5}$ were kept for further analysis. The names of the top hits were then converted to ZFIN ID via Biomart, Ensembl. These ZFIN IDs were used as population set. Similarly, the ZFIN IDs were obtained for contigs of ASE-genes and these IDs were used as study sets.

The first step in Ontologizer was to set up a new project. In this step, file was set ad “Human”. The association file, gene_association.zfin (http://www.geneontology.org/gene-associations/), contained the gene mapping information to the GO terms. Its path was specified in the “annotation”. The location of the OBO-file, which defines GO structure, was also specified. Then ZFIN IDs of population and study sets were later copy-and-pasted into the blank spaces as directed. The next step was to compute the enrichment analysis. The relationship of the population set and study sets was set up as “Parent-Child-Union” and the “Benjamini-Hochberg” test was applied to compute the statistics. By clicking the “Onyologize” button, Ontologizer would start the enrichment analysis. The GO term with an adjusted $p \leq 0.05$ were considered as significant Go terms. The significant GO ID, names and *p* values were shown as the result. GO terms of different categories were marked in different colors.

3.8 Expression Analysis of ASE-genes

Gene expression level was characterized in terms of reads per kilobase per million reads (RPKM) (Mortazavi, Williams et al. 2008). RPKM value was calculated as “$RPKM = \frac{10^9 * C}{N * L}$”: C was the number of reads mapped to a gene; N was the total mapped reads in the experiment; L was exon length in base-pairs for a gene, here we used the length of the contig instead. A previously generated catfish full-length cDNA database was used as the mapping reference (Chen, Lee et al. 2010; Liu, Zhang et al. 2012). The mapping procedure was conducted via CLC genomics workbench through “NGS core tools – map reads to the reference” function. The parameters were set as previous mapping procedure and the result was exported as Excel file. The RPKM value was automatically computed by selection “Expression value” as “Gene: RPKM”.

3.9 Identification of ribosomal RNAs

Catfish ribosomes contained four kind ribosomal RNAs (rRNA): the 40S small subunit contained 18S rRNA; the 60S large subunit contained 5S rRNA, 5.8S rRNA and 28S rRNA. Channel catfish and zebrafish rRNA sequences were downloaded from NCBI (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/) and Ensembl (http://useast.ensembl.org/index.html). The accession numbers I used in this study were displayed in Table 2. These sequences were used as queries to BLAST against the RNA-Seq assembly. Top hits with $p \leq e^{-10}$ were kept for further analysis. The contigs associated with top hits were further subjected to BLASTN analysis against with Non-redundant database to make sure they were the right rRNA genes.

Table 2. rRNAs Accession Numbers of channel catfish and zebrafish. These sequences were obtained from NCBI and Ensembl.

rRNA	Catfish	Zebrafish
18S	AF021880	FJ915075.1
5S	ICTRRA	AF213516 AF213517
5.8S	\	ENSDART00000121881
28S	AF056008	AF398343

3.10 Identification of thermal-induced ASE-SNPs

SNPs from tolerant and intolerant groups were pooled into one group, named “heat group”. The two alleles at s SNP site could be named as reference and non-reference alleles based on the genotype of the RNA-Seq assembly. The number of the reference allele in the heat group was calculated as the sum of the reference allele from the intolerant and tolerant groups. So did the non-reference allele. The number of the reference alleles and the number of the non-reference alleles of the heat group and control group were subjected to Fisher’s exact test using Microsoft Research False Discovery Rate Calculator (http://research.microsoft.com/en-us/um/redmond/projects/MSCompBio/FalseDiscoveryRate/). SNPs passed the Fisher’s exact test with $q \leq 0.05$ were kept for further analysis.

Control ratio was calculated as follows: if number of reference allele ≥ number of non-reference allele, control ratio = number of reference allele / number of non-reference allele; if number of reference allele < number of non-reference allele, control ratio = -1 × (number of non-reference allele / number of reference allele). Heat ratio was calculated follow the same procedure. CH ratio was introduced to demonstrate how different the control ratio and heat ratio were: 1) for SNPs with control ratio > 0 and heat ratio > 0: if control ratio > heat ratio, CH ratio = control ratio / heat ratio; if control ratio < heat ratio, CH ratio = heat ratio / control ratio. 2) for SNPs with control ratio < 0 and heat ratio <0: if control ratio > heat ratio, CH ratio = heat ratio / control ratio; if control ratio < heat ratio, CH ratio = control ratio / heat ratio. 3) for SNPs with control ratio > 0 and heat ratio < 0: CH ratio = control ratio / (-1/heat ratio). 4) for SNPs with control ratio < 0 and heat ratio > 0: CH ratio = heat ratio / (-1/ control ratio). SNPs with $q \leq 0.05$ and CH ratio ≥ 3 were considered as thermal-induced SNPs.

3.11 Analysis of thermal-induced ASE-genes

Gene annotation was conducted follow the same procedure of ASE-genes. Contigs containing at least one thermal-induced ASE-SNP were considered as thermal induced ASE-genes. Gene ontology analysis was conducted via ontologizer (Version 2.0 http://compbio.charite.de/contao/index.php/ontologizer2.html) following the procedure of ASE-genes. First, the thermal-induced ASE-SNP containing contigs were BLAST against Zebrafish protein database (Danio_rerio.GRCz10.pep.all.fa, ftp://ftp.ensembl.org/pub/release-80/fasta/danio_rerio/pep/). Secondly, the top hits were converted into ZFIN IDs via Biomart. Thirdly, these ZFIN IDs were used as study set for enrichment analysis with "Parent-Child-Union" and the "Benjamini-Hochberg" settings. GO terms with adjusted $p \leq 0.1$were considered as significantly enriched terms.

Reference

Ashburner M, Ball CA, et al. (2000) Gene Ontology: tool for the unification of biology. Nature genetics 25(1): 25-29

Bauer S, Grossmann S, et al. (2008) Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24(14): 1650-1651

Chen F, Lee Y, et al. (2010) Identification and characterization of full-length cDNAs in channel catfish (*Ictalurus punctatus*) and blue catfish (*Ictalurus furcatus*). PLoS One 5(7): e11546

Ewing B and Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research 8(3): 186-194

Ewing B, Hillier L, et al. (1998) Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. Genome research 8(3): 175-185

Fisher RA (1922) On the interpretation of χ2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society: 87-94

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

Li H, Handsaker B, et al. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16): 2078-2079

Liu S, Sun L, et al. (2014) Development of the catfish 250K SNP array for genome-wide association studies. BMC research notes 7(1): 135

Liu S, Wang X, et al. (2013) RNA-Seq reveals expression signatures of genes involved in oxygen transport, protein synthesis, folding, and degradation in response to heat stress in catfish. Physiological genomics 45(12): 462-476

Liu S, Zhang Y, et al. (2012) Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. Bmc Genomics 13(1): 595

Mortazavi A, Williams BA, et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5(7): 621-628

Chapter 4: Results and Discussion

4.1 Identification of SNPs and ASE-SNPs

SNPs were identified by aligning the short reads to the reference assembly of RNA-Seq. After quality control, a total of 66,251 initial SNPs were identified in the liver. These initial SNPs were located on 16,210 contigs. From the 66,251 SNPs, 18,533 SNPs passed the Fisher's test with significant differences between the numbers of reads from major alleles and minor alleles at each SNP site. A total of 5,420 (8.2%) out of 66,251 SNPs identified were classified as ASE-SNPs in the liver and they are located on 3,243 contigs (Table 3).

Similarly, in gill, 177,841 SNPs were identified from 33,860 contigs. Among these, 13,399SNPs passed the Fisher's exact test and 13,390 (7.5%) SNPs exhibited 9 folds or greater difference in allele ratios and therefore, classified as ASE-SNPs in the gill. The ASE-SNPs in gill are located on 6,732 contigs (Table 3).

Table 3. Summary of SNPs, ASE-SNPs and ASE-genes from the hybrid catfish transcriptome.

	Liver	**Gill**
Number of initial SNPs	66,251	177,841
Number of Contigs Containing initial SNPs	16,210	33,860
Number of SNPs after Fisher's exact test	18,533	39,475
Number of Contigs Containing SNPs after Fisher's exact test	6,942	13,399
Number of ASE-SNPs	5,420 (8.2%)	13,390 (7.5%)
Number of Contigs Containing ASE-SNPs	3,243	6,732
Number of ASE-SNPs with ASE-genes	3,955	9,500
Number of Contigs with ASE-genes	2,326	4,703
Number of ASE-genes	1,519	3,075

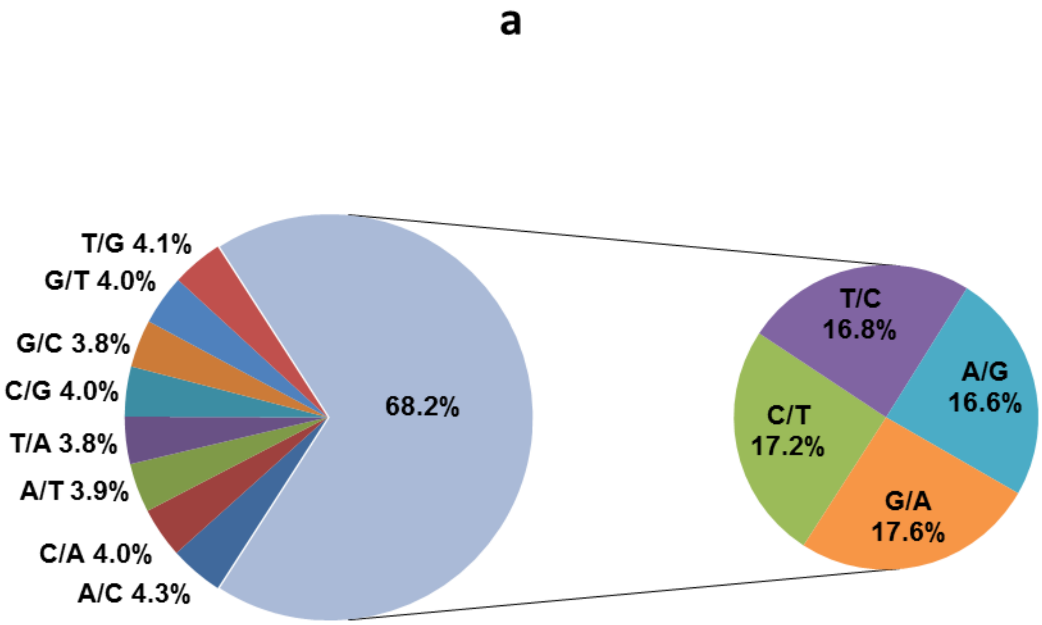
4.2 Types of SNPs and ASE-SNPs

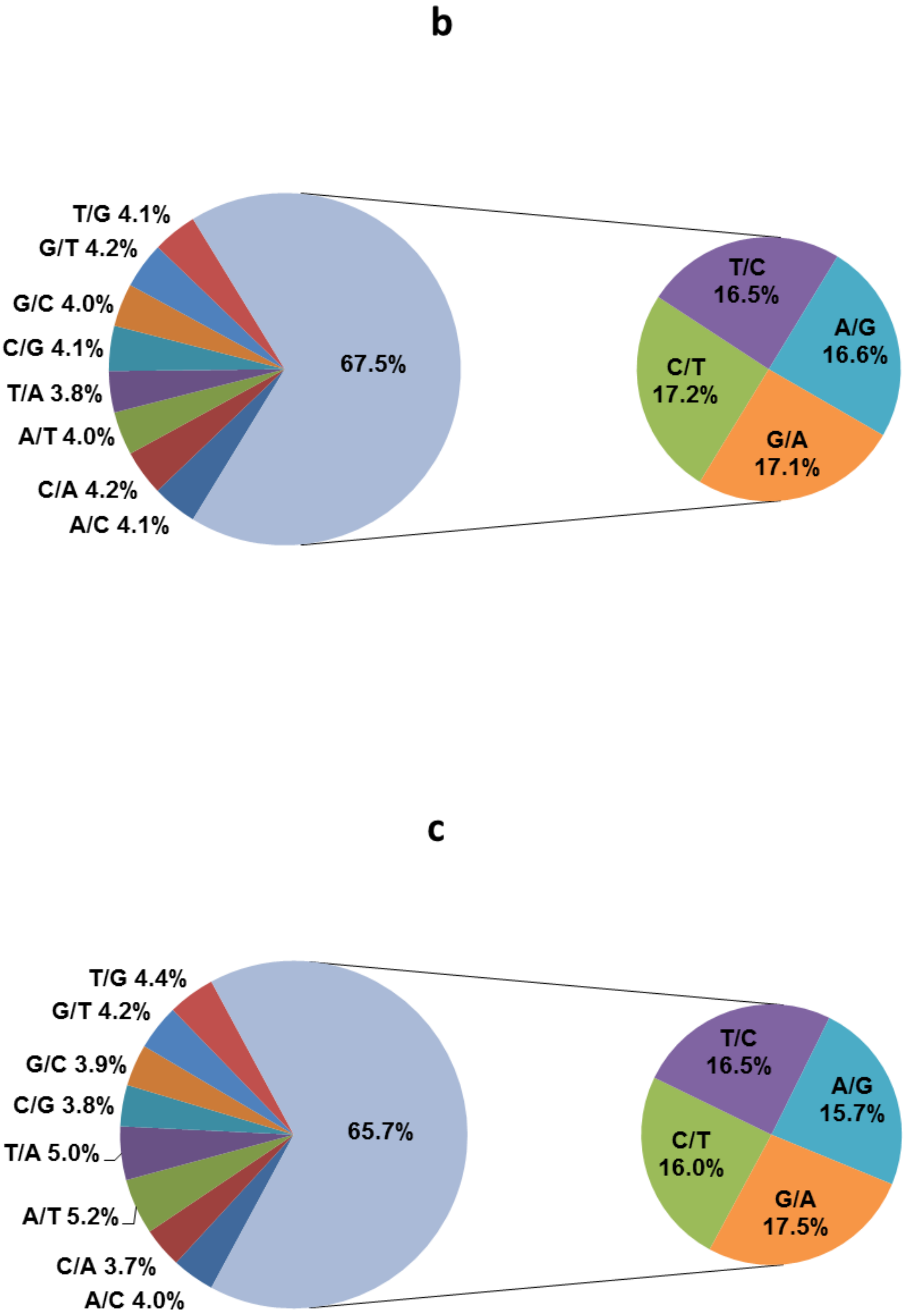
There were a total of 12 types of SNP identified. Four of them were transition SNPs, and eight of them are transversion SNPs. A transition is a point mutation that changes a purine nucleotide to another purine (A ↔ G) or a pyrimidine nucleotide to another pyrimidine (C ↔ T) while a transversion refers to the change from purine to pyrimidine (A → C, A → T, G→ C, G→ T) or vice versa (T → G, G → T, G → C, C → G, T → A, A → T, C → A, A → C). The number and percentage of SNPs in each SNP type is presented in Table 4. The most commonly occurring SNP types of all identified SNPs were transition SNPs, accounting for 68.2% SNPs in the liver and 67.5% SNPs in the gill (Figure 1). The transition-transversion ratio (Ts/Tv) of the all the SNPs was approximately 2.1 for both tissues. Among ASE-SNPs, the percentage of the transition SNPs decreased from 68.2% and 67.5% to 65.7% and 62.3% in liver and gill, respectively. The Ts/Tv values for ASE-SNPs also decreased from 2.1 to 1.9 in the liver and from 2.1 to 1.6 in the gill.

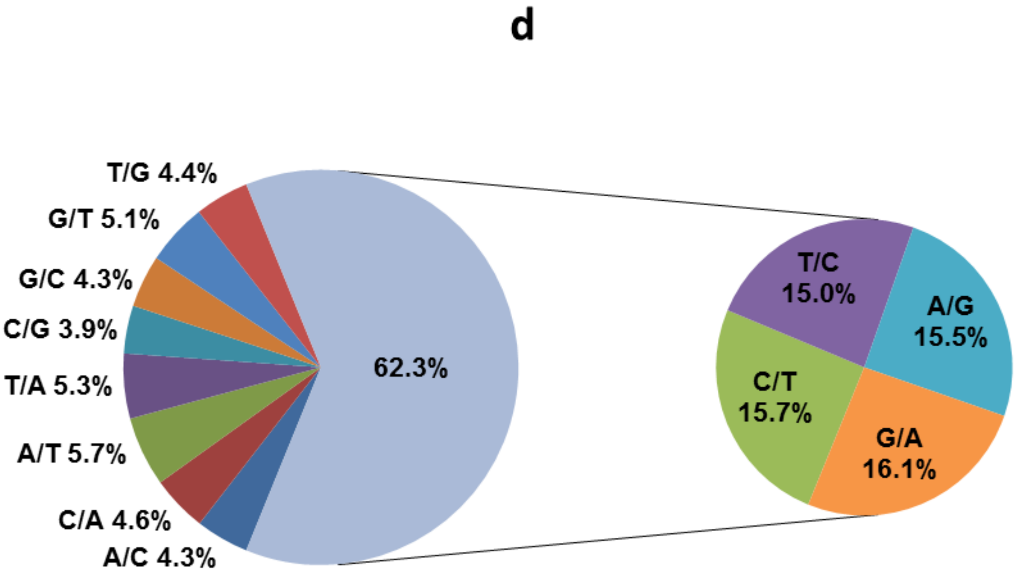
Table 4. SNP type proportion of initial SNPs and ASE-SNPs in the liver and gill tissues of hybrids catfish. There are 12 types of SNPs, T/C, C/T, A/G, G/A, T/G, G/T, G/C, C/G, T/A, A/T, C/A, A/C. The number and percentage of SNPs in each type were displayed in the table.

	Liver				Gill			
	Initial-SNPs		ASE-SNPs		Initial-SNPs		ASE-SNPs	
SNP types	SNP number	percentage	SNP number	percentage	SNP number	percentage	SNP number	percentage
A/C	2822	4.3%	215	4.0%	7361	4.1%	582	4.3%
C/A	2657	4.0%	203	3.7%	7416	4.2%	615	4.6%
A/T	2592	3.9%	282	5.2%	7092	4.0%	765	5.7%
T/A	2495	3.8%	270	5.0%	6751	3.8%	710	5.3%
C/G	2633	4.0%	208	3.8%	7361	4.1%	525	3.9%
G/C	2516	3.8%	211	3.9%	7034	4.0%	574	4.3%
G/T	2644	4.0%	230	4.2%	7548	4.2%	679	5.1%
T/G	2692	4.1%	238	4.4%	7264	4.1%	593	4.4%
C/T	11387	17.2%	867	16.0%	30561	17.2%	2098	15.7%
T/C	11119	16.8%	895	16.5%	29420	16.5%	2010	15.0%
A/G	11029	16.6%	851	15.7%	29535	16.6%	2081	15.5%
G/A	11665	17.6%	950	17.5%	30498	17.1%	2158	16.1%
Total	66251	100.0%	5420	100.0%	177841	100.0%	13390	100.0%

Figure1: SNP type proportion of initial SNPs and ASE-SNPs in the liver and gill tissues of hybrids catfish. There are 12 types of SNPs, T/C, C/T, A/G, G/A, T/G, G/T, G/C, C/G, T/A, A/T, C/A, A/C. Each type is marked with a different color and labeled as “reference genotype/ nonreference genotype”. **a:** Initial SNP type proportion in the liver; **b:** Initial SNP type proportion in the gill; **c:** ASE-SNP type proportion in the liver; **d:** ASE-SNP type proportion in the gill. The large light blue part of the pie chart represents the proportion of total transition SNPs over the total SNPs. This proportion is compared between initial SNPs and ASE-SNPs in liver and gill tissues separately.




b
T/G 4.1%
G/T 4.2%
G/C 4.0%
C/G 4.1%
T/A 3.8%
A/T 4.0%
C/A 4.2%
A/C 4.1%
67.5%
T/C 16.5%
A/G 16.6%
C/T 17.2%
G/A 17.1%
c
T/G 4.4%
G/T 4.2%
G/C 3.9%
C/G 3.8%
T/A 5.0%
A/T 5.2%
C/A 3.7%
A/C 4.0%
65.7%
T/C 16.5%
A/G 15.7%
C/T 16.0%
G/A 17.5%

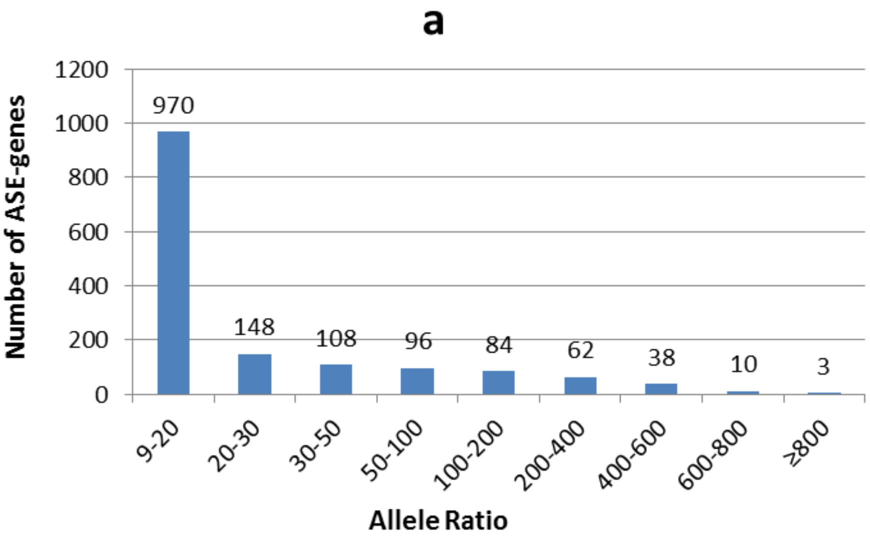


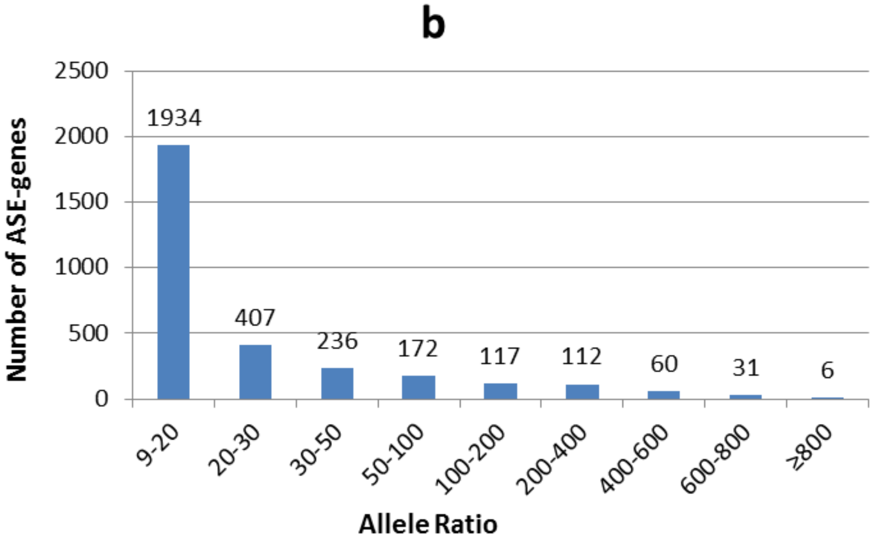
4.3 Identification of ASE-genes

BLAST analysis of 3,243 ASE-SNP containing contigs allowed assignment of 3,955 out of 5,420 ASE-SNPs to 1,519 genes in the liver. Similarly, 9,500 out of 13,390 ASE-SNPs in 6,732 contigs was associated with 3,075 genes in the gill. On average, there were around 3.6 and 4.4 ASE-SNPs on each of the ASE-genes in the liver and gill, respectively.

The allele ratios for ASE-genes ranged from 9.0 to 837.5 in liver and from 9.0 to 837 in gill. In liver, 970 (63.9%) out of 1,519 ASE-genes had an allele ratio between 9 and 20. Among these ASE-genes, 148 had an allele ratio within the range of 20-30 and 108 had an allele ratio from 30 to 50 (Figure 2). Although most of the ASE-genes had an allele ratio below 400, 51 (3.4%) ASE-genes had an allele ratio greater than 400. Gill showed similar pattern in the allele ratio distribution of ASE-genes. 1,934 (62.9%) out of 3,075 ASE-genes in gill had an allele ratio between 9 and 20. 407 and 236 ASE-genes had an allele ratio within 20-30 and 30-50 respectively.

Figure 2. Allele ratios distribution of ASE-Genes in the liver and gill of hybrid catfish. There are 499 and 674 ASE-genes identified in the liver and gill respectively. The allele ratio of a gene is represented by the highest allele ratio of the ASE-SNPs within the gene. **A:** Allele ratio distribution in the liver; **B:** Allele ratio distribution in the gill. About half of the ASE-genes have an allele ratio between 9 and 100. The number of ASE-genes decreases with the increase of allele-ratios.





4.4 Ontology Analysis of ASE-genes

BLAST analysis allowed assignment of 60,618 Contigs of the whole assembly to 6,710 zebrafish proteins. Similarly, 3,243 liver ASE-Contigs and 6,732 gill ASE-Contigs were assigned to 1,433 and 2,959 Zebrafish proteins, respectively. These protein IDs were later converted to ZFIN ID via Biomart as described previously. After conversion, 5,929 ZFIN IDs from the whole assembly were used as population set for Ontologizer. 1,433 and 2,959 ZFIN IDs from liver and gill were used as study sets respectively.

ASE-genes in the liver were subjected to 4,229 GO terms. Among them, 135 of them had an adjusted *p*-value smaller than 0.05. These enriched genes involved in a variety of metabolic processes including protein metabolism, sugar metabolism, lipid metabolism and energy metabolism. GO term “translation” (GO:0006412) and “ribosome” (GO:0005840) were the two most significant enriched terms. Genes under “translation” category were eukaryote translation initiation factors, eukaryote translation elongation factors, mitochondrial ribosomal proteins (mRPs), cytoplasmic ribosomal proteins (RPs) and tRNA synthases. Genes under “ribosome” includes several mRPs and many RPs. 11 representative GO terms were listed in the Table 5.

In the gill, ASE-genes were subjected to 5,801 GO terms, 12 of which were significant with adjusted *p*-value smaller than 0.05. The most significant GO term is “structural molecule activity” (GO:0005198). Gene products that helped to maintain cellular structure such as claudins, keratins, collagens, mRPs, RPs were subjected to this category. Similar to liver, GO term “translation” and “ribosome” were highly enriched in gill. Moreover, a group of immune related genes, “MHC protein complex” (GO:0042611) were also significantly enriched in gill ASE-genes. Six representative GO terms were listed in the Table 5.

Many of these enriched genes were correlated with liver functions and gill structures. The main function of fish liver was to help carbohydrates and fats digestion. There were ASE-genes in the liver found to be associated with steroid metabolism, lipid transportation and monosaccharide metabolism. Gill was the breathe organ for fish and the place to interact with outside environment. Gill was a “fragile” organ and it had very delicate structures. There were ASE-genes which products helped to maintain gill structures.

Table 5. Enriched GO terms of ASE-genes in the liver and gill. In the liver, 135 out of 4,229 GO terms had an adjusted *p*-value smaller than 0.05. 11 representative GO terms were listed in the table. In the gill, 12 of 5,801 GO terms were significantly enriched with adjusted *p*-value smaller than 0.05. 6 representative GO terms were listed in the table.

	GO ID	**GO Name**	**Category**	**Adjusted P-value**	**Population Count**	**Study Count**
Liver	GO:0006412	**translation**	Biological Process	1.84E-22	207	100
	GO:0005840	**ribosome**	Cellular Component	7.52E-13	123	70
	GO:0008152	**metabolic process**	Biological Process	1.71E-05	2827	674
	GO:0046034	**ATP metabolic process**	Biological Process	2.02E-05	33	19
	GO:0070469	**respiratory chain**	Cellular Component	8.69E-04	27	16
	GO:0006413	**translational initiation**	Biological Process	5.21E-03	39	19
	GO:0006457	**protein folding**	Biological Process	8.24E-03	69	29
	GO:0019538	**protein metabolic process**	Biological Process	2.76E-02	1155	287
	GO:0010876	**lipid localization**	Biological Process	3.54E-02	48	19
	GO:0008202	**steroid metabolic process**	Biological Process	3.62E-02	30	13
	GO:0005996	**monosaccharide metabolic process**	Biological Process	4.70E-02	46	21
Gill	GO:0005198	**structural molecule activity**	Molecular Function	2.61E-11	212	140
	GO:0006412	**translation**	Biological Process	2.43E-05	207	123
	GO:0005840	**ribosome**	Cellular Component	1.07E-04	123	81
	GO:0019538	**protein metabolic process**	Biological Process	4.61E-03	1155	509
	GO:0030529	**ribonucleoprotein complex**	Cellular Component	6.67E-03	261	143
	GO:0042611	**MHC protein complex**	Cellular Component	9.60E-03	20	18

4.5 Allele Ratio Distribution of RP Genes

In liver, a total of 54 catfish RP genes showed allelic specific expression. Among these allelic specifically expressed RP genes (ASE RP genes), RPL4 showed the highest allele ratio of 824.0 while RPS2 had an allele ratio as low as 11.7. There were 61 ASE RP genes identified in gill, with allele ratios ranged from 9.3 to 835.5. As shown in Table 6, a total of 62 ASE RP genes were found in our study, among which 53 were identified in both tissues, one was only allelic specifically expressed in liver and eight were only allelic specifically expressed in gill. The liver-specific ASE-RP gene was RPS24. The gill-specific ASE-RP genes were RPS10, RPS11, RPS18, RPL22, RPL28, RPL30, RPL34 and RPL35a. The allele ratios of ASE RP genes differed between tissues. Some of the ASE RP gene showing low allele ratio in liver had a high allele ratio in gill and vice versa. For example, RPS27 gene had an allele ratio of 13.5 in the liver and 737.0 in the gill. RPS15a gene had an allele ratio of 812.0 in the liver and 100.4 in the gill.

4.6 Parent-of-origins of ASE-ribosomal protein alleles

Studies had shown that the cell proliferation and growth were closely related to ribosome. As channel catfish generally grows faster than blue catfish, an interesting question to ask is which allele is preferentially expressed in F1 hybrid catfish. Therefore, I conducted an analysis to investigate the allele origins of the ASE RP genes. With existing catfish genomic information, in liver, the parent-of-origin could be determined for 27 ASE RP genes. Of these 27 genes, 13 were of channel catfish origins and 14 were of blue catfish origins. Similarly, 30 ASE RP genes could be identified with their parental origins in gill, 16 out of which were of channel catfish origin and the other 14 were of blue catfish origin (Table 6). Although there seemed no preferential expression from one parent, each RP gene appeared to be almost exclusively expressed from only one parent, i.e. in both liver and gill tissues, RPS7 was of channel catfish origin and RPL9 was of blue catfish origin. This indicated ribosomes in the hybrid catfish were in “hybrid forms”. For example, ribosomes were compromised of RPs from both species, with some from channel catfish and others from blue catfish. There was only one exception: RPL11 was of blue catfish origin in the liver while it was of channel catfish origin in the gill.

Table 6. ASE RP Genes Identified in Two Tissues. A total of 62 ASE RP genes were found in our study, among which 53 were identified in both tissues, one was only allelic specifically expressed in liver and eight were only allelic specifically expressed in gill.

	Tissues		**Allele Ratio**		**Parent-of-origin**	
Gene	**Liver**	**Gill**	**Liver**	**Gill**	**Liver**	**Gill**
RPSA	√	√	22.5	20.1	Blue	Blue
RPS2	√	√	11.7	202.0	\	\
RPS3	√	√	797.0	628.0	Channel	Channel
RPS3a	√	√	424.5	373.3	Blue	Blue
RPS5	√	√	165.7	782.0	\	\
RPS6	√	√	212.7	159.0	\	\
RPS7	√	√	810.0	538.3	Channel	Channel
RPS8	√	√	466.0	437.3	Blue	Blue
RPS9	√	√	624.0	679.5	\	\
RPS10	\	√	\	61.7	\	\
RPS11	\	√	\	811.5	\	\
RPS13	√	√	275.5	37.1	\	\
RPS14	√	√	526.0	400.0	\	\
RPS15	√	√	606.0	82.3	Blue	Blue
RPS15a	√	√	812.0	100.4	Blue	Blue
RPS16	√	√	17.2	264.3	\	\
RPS17	√	√	217.7	533.7	Channel	Channel
RPS18	\	√	\	258.3	\	\
RPS20	√	√	457.0	165.4	Blue	Blue

RPS23	√	√	539.5	460.0	Blue	Blue
RPS24	√	\	555.3	\	\	\
RPS25	√	√	784.0	403.8	\	\
RPS26	√	√	94.7	515.3	\	\
RPS27a	√	√	20.1	459.7	\	\
RPS30	√	√	48.0	429.5	\	Blue
RPL3	√	√	441.5	835.5	\	\
RPL4	√	√	824.0	698.5	Channel	Channel
RPL5	√	√	344.5	130.5	Channel	Channel
RPL6	√	√	708.5	409.0	Channel	Channel
RPL7	√	√	554.0	301.5	\	Channel
RPL7a	√	√	274.8	427.0	\	\
RPL8	√	√	448.0	217.8	Blue	Blue
RPL9	√	√	518.0	553.3	Channel	Channel
RPL10	√	√	331.0	252.5	Channel	Channel
RPL10a	√	√	311.2	483.7	Channel	Channel
RPL11	√	√	60.3	585.0	Blue	Channel
RPL12	√	√	176.6	203.5	Blue	Blue
RPL13	√	√	29.9	185.8	Channel	Channel
RPL13a	√	√	337.4	327.5	\	\
RPL14	√	√	196.0	256.3	Blue	Blue
RPL15	√	√	195.8	196.3	\	\
RPL17	√	√	417.7	780.0	\	\
RPL18	√	√	300.2	323.2	\	\
RPL18a	√	√	152.0	545.5	\	\
RPL19	√	√	459.0	258.5	\	\
RPL22	\	√	\	363.0	\	Channel
RPL23	√	√	241.9	128.5	Blue	\
RPL23a	√	√	219.2	129.1	\	\
RPL26	√	√	161.3	241.3	Blue	Blue
RPL27	√	√	13.5	737.0	Blue	Blue
RPL27a	√	√	568.0	325.0	Channel	\
RPL28	\	√	\	9.5	\	Channel
RPL30	\	√	\	415.5	\	\
RPL31	√	√	21.4	21.6	Channel	Channel
RPL34	\	√	\	650.5	\	\
RPL35	√	√	13.4	387.3	\	\
RPL35a	\	√	\	287.5	\	Blue
RPL36	√	√	87.9	29.5	\	\
RPL36a	√	√	133.0	154.1	\	\
RPL37a	√	√	34.3	9.3	\	\

RPP0	√	√	118.4	115.7	Channel	Channel
RPP1	√	√	502.0	214.1	\	\

4.7 Ribosomal proteins are highly expressed in catfish

The expression level of RP genes was characterized in terms of RPKM. RPKM value was calculated for each contig by mapping the short reads to the catfish full-length cDNA database. The previously generated catfish full-length cDNA database contained a total of 26,738 genes, among which 179 (0.67%) were RP genes. Although the RP genes accounted for such a low percentage of the total genes, they contribute as high as 16.1% and 19.7% of the RPKM value of all the contigs in the liver and gill RNA-Seq, respectively. In order to investigate whether the high expression of RP genes was a common phenomenon in catfish species or a specific pattern in F1 hybrid catfish, RNA-Seq data of blue and channel catfish from other studies were applied following the same procedure. As shown in Table 7, RP genes accounted for 25.3% of the total RPKM value in the RNA-Seq of channel catfish gill, which was obviously higher than the RPKM percentage in hybrid catfish. Also, when looking at the whole fry’s level, we found that the percentage of RPKM value contributed by RP genes was smaller in hybrid (10.2%) catfish fries than blue (10.9%) and channel (10.4%) catfish fries (unpublished RNA-Seq data). The less expression of RP genes in hybrids indicating less RP transcripts were needed. Thus, hybrid ribosomes probably worked more efficient than homozygous counterparts.

Table 7. RPKM value of different tissues in different catfish. The expression level of RP genes was characterized in terms of RPKM. The percentage of RPKM value contributed by RP genes was smaller in hybrid (10.2%) catfish fries than blue (10.9%) and channel (10.4%) catfish fries.

Tissue	Liver	Gill		Whole Fry		
Species	F1 Hybrid	F1 Hybrid	Channel	F1 Hybrid	Channel	Blue
RPKM value of total genes	683,213	513,512	492,015	338,690	341,685	354,291
RPKM value of RP genes	109,997	101,396	124,536	34,564	35,573	38,512
Percentage	16.1%	19.7%	25.3%	10.2%	10.4%	10.9%

4.8 ASE of ribosomal RNAs

A total of 18 contigs were considered as rRNA candidates. They were identified by BLAST the channel catfish and zebrafish rRNA queries against the RNA-Seq assembly. Further BLAST analysis of these candidate contig against the Non-redundant database excluded six candidate contigs with poor or fair rRNAs identities. After this step, only contigs associated with 18S and 28S rRNA were kept. For the rest 14 candidate contigs, nine were found to contain ASE-SNPs, three contain initial SNPs and two contains no SNPs.

The nine ASE-SNP containing contigs were later aligned to the channel catfish reference genome for their locations. I found that there were two copies of 18S rRNAs and multiple copies of 28S rRNAs. One copy of 18S rRNA showed ASE while the other copy showed bi-allelic expression. All of the 28S rRNA copies showed ASE. Parent-of-origins identification followed previous methods. However, no parent-of-origins were identified for these alleles. The expression status and allele ratios of rRNA contigs were displayed in Table 8.

Table 8. Expression Status of rRNAs. 5S and 5.8S rRNAs were not identified in the current RNA-Seq data. One copy of 18S rRNA showed ASE while the other copy showed biallelic expression. All of the 28S rRNA copies showed ASE.

rRNA	Expression	Allele Ratio	
		Liver	Gill
5S	not identified	\	\
5.8S	not identified	\	\
18S	Bi-allelic	\	\
	ASE	503.3	586.5
28S	ASE	636.5	546.5

4.9 ASE of mitochondrial RPs

A considerable set of mitochondrial ribosomal protein genes (mRPs) were also significantly enriched in the ASE-genes in the liver and gill. A total of 24 mRPs were found to be allelic specifically expressed in the two tissues. Eight ASE-RPs were from 28S small mitochondrial ribosomal subunit and 16 were from the 39S large mitochondrial ribosomal subunit. Ten mRPs were found to be allelic specifically expressed in both tissues. There were five mRPs specifically expressed in the liver and nine specifically expressed in the gill. The allele ratios of mRPs were found to be lower than those of RPs, ranging from 9.5 to 55.0.

The parent-of-origins of the alleles were identified using the species-specific SNPs and genome references. In the liver, seven of mRPs were found to be of blue catfish origins and two were of channel catfish origins. In the gill, six of mRPs were found to be of blue catfish origins and two were of channel catfish origins. It seemed that there were blue catfish origin mRPs than channel catfish ones. The allele ratios and parent-of-origins of mRPs were displayed in Table 9.

Table 9. ASE mRP Genes Identified in Two Tissues. A total of 24 ASE RP genes were found in our study, among which ten were identified in both tissues, five was only allelic specifically expressed in liver and nine were only allelic specifically expressed in gill.

	Tissues		**Allele Ratio**		**Parent-of-origin**	
Gene	**Liver**	**Gill**	**Liver**	**Gill**	**Liver**	**Gill**
mRPS5	√	\	15.5	\	\	\
mRPS9	\	√	\	13.0	\	Blue
mRPS17	\	√	\	11.5	\	\
mRPS21	√	\	13.2	\	\	\
mRPS23	√	√	13.0	23.0	Blue	\
mRPS27	\	√	\	21.0	\	Channel
mRPS28	√	\	9.5	\	\	\
mRPS29	√	√	10.0	34.0	Blue	\
mRPL2	\	√	\	20.3	\	\
mRPL4	√	√	10.7	16.3	Blue	Blue
mRPL12	√	√	17.0	21.0	Blue	\
mRPL13	\	√	\	77.0	\	\
mRPL18	√	√	15.5	18.7	Channel	\
mRPL19	√	√	14.4	12.1	\	\
mRPL24	√	\	14.0	\	Blue	\
mRPL27	√	√	21.5	31.7	Blue	Blue
mRPL28	√	√	13.0	9.8	\	\
mRPL35	\	√	\	12.0	\	Blue
mRPL41	\	√	\	9.7	\	\
mRPL43	√	√	23.5	10.5	Channel	\
mRPL45	\	√	\	55.0	\	Channel
mRPL46	\	√	\	13.3	\	Blue
mRPL47	√	\	23.0	\	\	\
mRPL55	√	√	18.0	23.3	Blue	Blue

4.10 Identification of thermal-induced ASE-SNPs and thermal-induced ASE-genes

In the liver, 4,991 SNPs out of 66,251 SNPs passed the Fisher's test with significant differences between the allele numbers of control group and heat group at each SNP site. A total of 1,944 (2.9%) SNPs identified were classified as thermal-induced ASE-SNPs in the liver and they are located on 1,294 contigs (Table 10). Similarly, in the gill, 4,185 out of 177,841 SNPs passed the Fisher's exact test and 2,066 (1.2%) SNPs were classified as thermal-induced ASE-SNPs in the gill. The thermal-induced ASE-SNPs in gill are located on 1,432 contigs (Table 10).

BLAST analysis of 1,294 thermal-induced ASE-SNP containing contigs allowed assignment of 1,525 out of 1,944 thermal-induced ASE-SNPs to 864 genes in the liver. Similarly, 1,546 out of 2,066 thermal-induced ASE-SNPs in 1,053 contigs was associated with 909 genes in the gill (Table 10).

Table 10. Summary of SNPs, thermal induced ASE-SNPs and thermal induced ASE-genes from the hybrid catfish transcriptome.

	Liver	**Gill**
Number of initial SNPs	66,251	177,841
Number of Contigs Containing initial SNPs	16,210	33,860
Number of SNPs after Fisher's exact test	4,991	4,185
Number of Contigs Containing SNPs after Fisher's exact test	2,672	2,503
Number of thermal induced ASE -SNPs	1,944 (2.9%)	2,066 (1.2%)
Number of Contigs Containing thermal induced ASE -SNPs	1,294	1,432
Number of thermal induced ASE -SNPs with thermal induced ASE -genes	1,525	1,546
Number of Contigs with thermal induced ASE -genes	998	1,053
Number of thermal induced ASE -genes	864	909

4.11 Analysis of thermal-induced ASE-genes

A total of 1,294 liver thermal-induced ASE-Contigs and 1,432 gill thermal-induced ASE-Contigs were assigned to 863 and 922 Zebrafish proteins, respectively. These protein IDs were later converted to ZFIN ID via Biomart as described previously. After conversion, 1,433 and 2,959 ZFIN IDs from liver and gill were used as study sets respectively. 5,929 ZFIN IDs from the whole assembly were used as population set for Ontologizer.

Thermal-induced ASE-genes in the liver were subjected to 3,624 GO terms. Among them, 16 of them had an adjusted *p*-value smaller than 0.1. A significantly set of enriched thermal-induced ASE-gene products were located as extracellular region (GO:0005576). Several enriched GO terms were function as "binding" including iron ion binding (GO:0005506), cofactor binding (GO:0048037), tetrapyrrole binding (GO:0046906) and quaternary ammonium group binding (GO:0050997). There were also some GO terms were involved in metabolic processes: organic acid metabolic process (GO:0006082), single-organism metabolic process (GO:0044710), xenobiotic metabolic process (GO:0006805), cellular amino acid metabolic process (GO:0006520), drug metabolic process (GO:0017144) and cellular ketone metabolic process (GO:0042180) (Table 11).

In the gill, thermal-induced ASE-genes were subjected to 3,692 GO terms and four of them had an adjusted *p*-value smaller than 0.1. Genes involved in nucleoside binding (GO:0001882) and cytoskeletal part (GO:0044430) were highly enriched. Examples of these genes were myosin, heat shock protein, keratin, ATPase and actin (Table 11).

Table 11. Enriched GO terms of thermal-induced ASE-genes in the liver and gill. In the liver, 16 out of 3,624 GO terms had an adjusted *p*-value smaller than 0.1. Six representative GO terms were listed in the table. In the gill, four of 3,692 GO terms were significantly enriched with adjusted *p*-value smaller than 0.1. Three representative GO terms were listed in the table.

	GO ID	**GO Name**	**Category**	**Adjusted P-value**	**Population Count**	**Study Count**
Liver	GO:0005506	**iron ion binding**	Molecular Function	1.56E-03	65	23
	GO:0005576	**extracellular region**	Cellular Component	1.56E-03	221	57
	GO:0044710	**single-organism metabolic process**	Biological Process	5.10E-03	940	167
	GO:0006520	**cellular amino acid metabolic process**	Biological Process	6.67E-02	101	27
	GO:0003824	**catalytic activity**	Molecular Function	8.12E-02	1986	299
	GO:0008593	**regulation of Notch signaling pathway**	Biological Process	8.26E-02	3	3
Gill	GO:0001882	**nucleoside binding**	Molecular Function	5.30E-02	688	117
	GO:0044430	**cytoskeletal part**	Cellular Component	6.15E-02	157	33
	GO:0016459	**myosin complex**	Cellular Component	9.07E-02	33	13

Chapter 5: Discussion

5.1 Transition SNP percentage Decreased from Initial SNPs to ASE-SNP

In liver, a much larger number of transition SNPs (68.2%) were identified in F1 hybrid catfish than transversion SNPs (31.8%). Similar pattern was also found in gill with 67.5% transition SNPs and 32.5% transversion SNPs. This result was consistent with previous ASE studies of F1 hybrid rice in which transition SNPs account for 68% and 74% of total SNPs identified (Chodavarapu, Feng et al. 2012; Zhai, Feng et al. 2013). In fact, it is generally believed that the bias in favor of transitions over transversions is universal, and it is possibly caused by the underlying chemistry of mutation. Purines can be altered to resemble each other, and so does pyrimidines. However, a purine cannot be altered to resemble a pyrimidine, nor vice versa. Recently, by systematically comparing transcriptome and methylome, Chodavarapu *et al.* found that methylated cytosines mutate more than three times more frequently than nonmethylated cytosines, and they mostly mutate to thymines (Chodavarapu, Feng et al. 2012). Theses may explain why such a high percentage of transition SNPs occurred.

More interstingly, we found that the percentage of transition SNPs decreased in ASE-SNPs. As shown in Table 3 and Figure 1, the percentage of transition SNPs decreased from 68.2% to 65.7% in liver, and from 67.5% to 62.3% in gill. A previous study demonstrated that the methylated cytosine-guanine (CpG) dinucleotides exhibit the high transition frequencies while the transition rate at other cytosine residues is significantly lower (Keller, Bensasson et al. 2007). This indicates that fewer methylated CpG dinucleotides might be included in the ASE-SNPs. In human, the Ti/Tv ratio falls between ~2.0-2.1 for genome-wide datasets and 3.0-3.3 for exonic variations (DePristo, Banks et al. 2011). More specifically, Freudenberg-Hua, Y. *et al.* observed that the Ti/Tv ratio was smaller in noncoding region (1.99) than coding region (3.02) in an European population (Freudenberg-Hua, Freudenberg et al. 2003). In my study, the Ti/Tv values decreased from initial SNPs to ASE-SNPs: from 2.1 to 1.9 in the liver and from 2.1 to 1.6 in the gill. As RNA-Seq datasets mainly contains sequence information of mRNAs and non-coding RNAs, our findings of decreased Ti/Tv ratio indicates that a certain portion of ASE-SNPs are distributed in the non-coding region. Previous study found that non-coding RNAs, including 3' UTRs and long non-coding RNAs (lncRNAs), were more accessible for interaction with the rest of the cellular factors, thus paly an role in regulation of gene expression (Niazi and Valadkhan 2012). If a considerable portion of ASE-SNOs falls in non-coding region, they would be more responsible for gene expression regulation than directly generating variation.

5.2 RPs Expression and Growth and Fitness

The translation process is catalyzed by the translational machine, ribosomes. Eukaryotic ribosomes, also known as 80S ribosomes, have two unequal subunits, designated as small subunit (40S) and large subunit (60S) according to their sedimentation coefficients. The 40S subunit contains the decoding center, which functions as monitor of the complementarity of transfer RNA (tRNA) and messenger RNA (mRNA) in protein translations. The main function of 60S subunit is to catalyze peptide formation. Mammalian ribosomes are well characterized; they are composed of 79 proteins and four RNAs. The 60S subunit is composed of three ribosomal RNAs (rRNAs) and 47 RPs where the 40S subunit is composed of 18S rRNA and 32 RPs (Wool 1979). In channel catfish, 32 RP genes of the 40S subunit and 47 RP genes of the 60S subunit have been cloned and sequenced in previous studies (Karsi, Patterson et al. 2002; Patterson, Karsi et al. 2003). Ribosome is often considered as “house-building” machinery because of its protein synthesis function. As cell growth requires large numbers of ribosomes for protein accumulation, thus, ribosome biogenesis is thought to be closely related to the cell’s capacity to grow (Lempiäinen and Shore 2009). In rapidly growing yeast cells, 60% of total transcription activity is devoted to rRNA, and 50% of RNA polymerase II activity occurs on RP genes (Warner 1999).

Studies found ribosomes could differ in the stoichiometry of RPs, thus tuning their functions. In *E. coli,* the alterations in stoichiometry of RPs have long been observed to be associated with growth. The amount of RPs S6, S21 and L21 were found to differ significantly between cells grown from rich or minimal media (Deusser 1972). Milne *et al.* further confirmed the association of the amount of these three RPs and *E. coli* growth rate in different nutrient conditions (Milne, Mak et al. 1975). In slime mold *Dictiostelium discoideum*, the qualitative and quantitative differences of 12 unique ribosomal proteins between the vegetative amoebae and spores have been observed, indicating cell differentiation in a eukaryotic system was accompanied by ribosome heterogeneity (Ramagopal and Ennis 1981). The “ribosomal filter hypothesis” was later proposed: the ribosome modulates translation by selectively translating specific mRNAs. The hypothesis emphasizes that mRNA sequences compete for binding to rRNA or ribosomal proteins, and this differential binding may affect translation rates (Mauro and Edelman 2002). Potentially, this hypothesis could benefit the “energy-use efficiency theory” of heterosis: hybrid offspring have greater energy efficiency via selective protein synthesis and metabolism (Goff 2011). Goff proposed that the gain of multi-genetic heterosis in hybrids lies in energy efficiency during protein processing. He argued that allelic variants within a gene might encode unstable or inefficient proteins, thus cost more energy. A hybrid may have alleles that code both efficient protein and inefficient protein at the same locus. If the hybrid could preferentially transcribe or translate efficient proteins, better energy efficiency and superior phenotypical performance would be achieved. In our study, RP genes are highly enriched among ASE genes in both tissues. These allelic variations within RP genes may potentially attribute to the "ribosome filter hypothesis" and "energy hypothesis". Hybrid catfish with ribosomes assembled of favorable alleles could selectively translate specific mRNA that potentially benefit for certain traits. For example, as blue catfish is more resistant for enteric septicemia of catfish (ESC) disease than channel catfish, if the hybrid ribosomes preferentially translate ESC-resistance blue catfish transcripts, the hybrid could gain more disease resistance. If hybrid catfish could express favorable alleles that encode efficient protein, they could save more energy and benefit for growth.

There is one study also identified an ASE-RP. In this maize ASE study, researchers identified a RP gene exhibited tissue-specific ASE pattern (Springer and Stupar 2007). Although there is little ASE researches describe RP genes, several studies report the differential expression of RP genes between the parents and hybrids on both transcriptional and translational levels. In pacific oyster, three RP transcripts were found to display nonadditive expression between hybrid and inbred larvae (Hedgecock, Lin et al. 2007). Later, by comparing the mRNA expression pattern of slow-growth and fast-growth larvae, 17 RP genes were identified to be responsible for growth difference in bivalve larval oysters (Meyer and Manahan 2010). In wheat, the analysis of cDNA amount found that 35 RP genes were differentially expressed between hybrid and its parents (Yao, Ni et al. 2005). On the proteomic level, many studies observed the differential expression or nonadditive pattern of RPs between hybrids and parents in plants including wheat, maize and sunflower (Song, Ni et al. 2007; Hoecker, Lamkemeyer et al. 2008; Marcon, Lamkemeyer et al. 2013; Mohayeji, Capriotti et al. 2014). These extensive studies provide evidence for the association of stoichiometry RPs expression and growth and fitness, however, the underlying mechanism remains unknown. Our identification of ASE-RP genes may provide new perspective towards the answer.

5.3 Less RP Transcripts were Expressed in Hybrid

Interestingly, we found that RP genes were highly expressed in different tissues of hybrid catfish, channel catfish and blue catfish, composing about 10% in two-month-old fries and ranging from 16% to 25% in one-year-old fingerlings. This indicates that highly expression of RP genes was probably a general phenomenon of the catfish species. Since the expression of RP genes were highly tissue specific, only RNA-Seq data from the same tissue between different species could be compared. As shown in Table 3, in gill, RP genes contributed 19.75% of the total RPKM value in the F1 hybrid catfish, which was much lower than the percentage in channel catfish gill (25.31%). For the whole fry of the same age, the RPKM percentage of RP genes in hybrid (10.2%) exhibited a lower level than that of channel catfish (10.4%) and blue catfish (10.6%). The lower expression of RP genes in hybrid catfish indicates that fewer ribosomal proteins were needed in hybrid catfish to carry out the protein synthesis activities. So maybe ribosomes may work more efficient in the hybrids than in channel and blue catfish. Such high expression of RP genes was not unique, early expressed sequence tags analysis in human observed that the expression RP gene accounted for 5%~20% of total transcription in different tissues (Bortoluzzi, d'Alessi et al. 2001).

5.4 Different RP Genes Had Different Parent-of-origins

The parent-of-origins of these RP genes were listed in table 6, with some RP genes were of channel catfish origin while some were of blue catfish origin. This indicates that the ribosomes in hybrid catfish were assembled with RPs of different origins, which means, the ribosomes were in "hybrid forms". Combined with these two observations of RP genes with lower expression and different parent-of-origins, we propose the hypothesis that hybrid ribosomes probably work more efficiently than homozygous counterparts. As the genomes of channel catfish and blue catfish share high similarity, the RP genes with allelic variations may just have a minor difference. However, these minor alterations in mRNA sequence have the potential to cause bigger difference in protein functions via amino acid substitution and post-translational modifications. One example is that, the mammalian RPS6 is phosphorylated on five serine residuals, when researchers substituted all five residuals with alanines, mouse embryos displayed an increased rate of protein synthesis and accelerated cell division with smaller cell sizes (Ruvinsky, Sharon et al. 2005).

My study is the very first of its kind in catfish to determine if ASE exists in the interspecific hybrid system. It provides a new avenue of research to determine the potential causes of heterosis at the transcriptional level and genome scale. Such research may provide insights into molecular mechanisms for heterosis, and therefore, may have broad implications for genetic improvement programs. In a future study, we plan to investigate the translation efficiency of ribosomes in hybrid catfish and inbred catfish. With the assistance of techniques such as ribosomal profiling, we hope to validate our findings and to investigate the expression patterns on the translational level.

Reference

Bortoluzzi S, d'Alessi F, et al. (2001) Differential expression of genes coding for ribosomal proteins in different human tissues. Bioinformatics 17(12): 1152-1157

Chodavarapu RK, Feng S, et al. (2012) Transcriptome and methylome interactions in rice hybrids. Proceedings of the National Academy of Sciences 109(30): 12040-12045

DePristo MA, Banks E, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43(5): 491-498

Deusser E (1972) Heterogeneity of ribosomal populations in *Escherichia coli* cells grown in different media. Molecular and General Genetics MGG 119(3): 249-258

Freudenberg-Hua Y, Freudenberg J, et al. (2003) Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. Genome research 13(10): 2271-2276

Goff SA (2011) A unifying theory for general multigenic heterosis: energy efficiency, protein metabolism, and implications for molecular breeding. New Phytologist 189(4): 923-937

Hedgecock D, Lin J-Z, et al. (2007) Transcriptomic analysis of growth heterosis in larval Pacific oysters (*Crassostrea gigas*). Proceedings of the National Academy of Sciences 104(7): 2313-2318

Hoecker N, Lamkemeyer T, et al. (2008) Analysis of nonadditive protein accumulation in young primary roots of a maize (*Zea mays L.*) F1 ‐ hybrid compared to its parental inbred lines. Proteomics 8(18): 3882-3894

Karsi A, Patterson A, et al. (2002) Translational machinery of channel catfish: I. A transcriptomic approach to the analysis of 32 40S ribosomal protein genes and their expression. Gene 291(1): 177-186

Keller I, Bensasson D, et al. (2007) Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS genetics 3(2): e22

Lempiäinen H and Shore D (2009) Growth control and ribosome biogenesis. Current opinion in cell biology 21(6): 855-863

Marcon C, Lamkemeyer T, et al. (2013) Heterosis-associated proteome analyses of maize (*Zea mays L*.) seminal roots by quantitative label-free LC–MS. Journal of proteomics 93: 295-302

Mauro VP and Edelman GM (2002) The ribosome filter hypothesis. Proceedings of the National Academy of Sciences 99(19): 12031-12036

Meyer E and Manahan D (2010) Gene expression profiling of genetically determined growth variation in bivalve larvae (*Crassostrea gigas*). The Journal of experimental biology 213(5): 749-758

Milne AN, Mak W, et al. (1975) Variation of ribosomal proteins with bacterial growth rate. Journal of bacteriology 122(1): 89-92

Mohayeji M, Capriotti AL, et al. (2014) Heterosis profile of sunflower leaves: A label free proteomics approach. Journal of proteomics 99: 101-110

Niazi F and Valadkhan S (2012) Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3′ UTRs. Rna 18(4): 825-843

Patterson A, Karsi A, et al. (2003) Translational machinery of channel catfish: II. Complementary DNA and expression of the complete set of 47 60S ribosomal proteins. Gene 305(2): 151-160

Ramagopal S and Ennis HL (1981) Regulation of synthesis of cell-specific ribosomal proteins during differentiation of *Dictyostelium discoideum*. Proceedings of the National Academy of Sciences 78(5): 3083-3087

Ruvinsky I, Sharon N, et al. (2005) Ribosomal protein S6 phosphorylation is a determinant of cell size and glucose homeostasis. Genes & development 19(18): 2199-2211

Song X, Ni Z, et al. (2007) Wheat (*Triticum aestivum* L.) root proteome and differentially expressed root proteins between hybrid and parents. Proteomics 7(19): 3538-3557

Springer NM and Stupar RM (2007) Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. The Plant Cell Online 19(8): 2391-2402

Warner JR (1999) The economics of ribosome biosynthesis in yeast. Trends in biochemical sciences 24(11): 437-440

Wool IG (1979) The structure and function of eukaryotic ribosomes. Annual review of biochemistry 48(1): 719-754

Yao Y, Ni Z, et al. (2005) Identification of differentially expressed genes in leaf and root between wheat hybrid and its parental inbreds using PCR-based cDNA subtraction. Plant molecular biology 58(3): 367-384

Zhai R, Feng Y, et al. (2013) Identification of transcriptome SNPs for assessing allele-specific gene expression in a super-hybrid rice Xieyou9308. PloS one 8(4): e60668

Appeddix Table1 Shared ASE-genes and the allele ratios in the liver and the gill.

Gene Name	Liver allele ratio	Gill allele ratio
15 kDa selenoprotein	149.0	26.6
26S protease regulatory subunit 8	12.5	18.0
26S protease regulatory subunit S10B	19.5	29.8
26S proteasome non-ATPase regulatory subunit 11	9.6	26.0
26S proteasome non-ATPase regulatory subunit 14	24.5	17.1
26S proteasome non-ATPase regulatory subunit 8	12.1	10.6
28S ribosomal protein S29, mitochondrial	10.0	34.0
2-oxoisovalerate dehydrogenase subunit alpha, mitochondrial	11.0	44.5
39S ribosomal protein L12, mitochondrial	17.0	21.0
39S ribosomal protein L18, mitochondrial	15.5	18.7
39S ribosomal protein L19, mitochondrial	14.4	12.1
39S ribosomal protein L27, mitochondrial	21.5	31.7
39S ribosomal protein L28, mitochondrial	13.0	9.8
39S ribosomal protein L4, mitochondrial	10.7	16.3
39S ribosomal protein L43, mitochondrial	23.5	10.5
39S ribosomal protein L55, mitochondrial	18.0	23.3
3-ketoacyl-CoA thiolase B, peroxisomal	12.6	13.5
3-ketoacyl-CoA thiolase, mitochondrial	11.0	12.7
3-mercaptopyruvate sulfurtransferase	10.6	14.5
40S ribosomal protein S13	64.0	37.1
40S ribosomal protein S14	526.0	400.0
40S ribosomal protein S15	606.0	82.3
40S ribosomal protein S15a	119.7	100.4
40S ribosomal protein S16	17.2	264.3
40S ribosomal protein S17	622.5	533.7

40S ribosomal protein S2	11.7	202.0
40S ribosomal protein S20	457.0	165.4
40S ribosomal protein S23	297.0	460.0
40S ribosomal protein S25	784.0	403.8
40S ribosomal protein S26	94.7	515.3
40S ribosomal protein S27a	20.1	459.7
40S ribosomal protein S3	354.7	628.0
40S ribosomal protein S30	48.0	429.5
40S ribosomal protein S3a	180.6	373.3
40S ribosomal protein S5	165.7	782.0
40S ribosomal protein S6	212.7	159.0
40S ribosomal protein S7	16.0	538.3
40S ribosomal protein S8	101.0	437.3
40S ribosomal protein SA	22.5	20.1
5-beta-cholestane-3-alpha,7-alpha-diol 12-alpha-hydroxylase	39.7	15.0
60 kDa heat shock protein, mitochondrial	38.0	89.0
60S acidic ribosomal protein P0	20.7	115.7
60S acidic ribosomal protein P1	13.8	214.1
60S ribosomal protein L10	19.1	252.5
60S ribosomal protein L10a	311.2	483.7
60S ribosomal protein L11	31.7	585.0
60S ribosomal protein L12	167.6	203.5
60S ribosomal protein L13	29.9	185.8
60S ribosomal protein L13a	38.4	327.5
60S ribosomal protein L14	196.0	256.3
60S ribosomal protein L15	136.5	196.3
60S ribosomal protein L17	417.7	780.0
60S ribosomal protein L18	300.2	323.2
60S ribosomal protein L18a	22.7	545.5
60S ribosomal protein L19	459.0	258.5
60S ribosomal protein L23	91.2	128.5

60S ribosomal protein L23a	219.2	129.1
60S ribosomal protein L26	161.3	241.3
60S ribosomal protein L27	13.5	737.0
60S ribosomal protein L27a	568.0	325.0
60S ribosomal protein L3	441.5	835.5
60S ribosomal protein L31	21.4	21.6
60S ribosomal protein L35	13.4	387.3
60S ribosomal protein L36	30.3	29.5
60S ribosomal protein L36a	133.0	154.1
60S ribosomal protein L37a	34.3	9.3
60S ribosomal protein L4	261.8	698.5
60S ribosomal protein L5	269.5	130.5
60S ribosomal protein L6	29.0	409.0
60S ribosomal protein L7	554.0	301.5
60S ribosomal protein L7a	16.3	427.0
60S ribosomal protein L8	86.0	217.8
60S ribosomal protein L9	518.0	553.3
60S ribosome subunit biogenesis protein NIP7	19.0	20.5
78 kDa glucose-regulated protein	22.9	29.7
7-alpha-hydroxycholest-4-en-3-one 12-alpha-hydroxylase	75.5	12.0
acidic leucine-rich nuclear phosphoprotein 32 family member E	10.7	22.0
actin, cytoplasmic 1	21.8	383.5
actin-related protein 3	14.0	76.5
actin-related protein 6	9.0	11.5
acyl-CoA dehydrogenase-like	9.2	10.0
acyl-CoA synthetase long-chain family member 1	9.8	13.5
acyl-protein thioesterase 2	12.0	15.8
ADAMTS-like protein 4	11.0	10.0
Adapter protein CIKS	12.0	20.0
adenosine kinase	22.5	10.4
adenosylhomocysteinase	200.0	73.8

adenylate kinase 2, mitochondrial	11.5	10.8
Adipocyte enhancer-binding protein 1	9.5	12.5
ADP/ATP translocase 2	37.4	324.3
ADP-ribosylation factor 5	13.0	91.0
ADP-ribosylation factor-like protein 6-interacting protein 1	13.0	27.2
ADP-ribosylation factor-like protein 8B-A	9.0	10.3
ADP-ribosylation factor-like protein 9	15.5	16.3
aflatoxin B1 aldehyde reductase member 2	11.5	11.7
alcohol dehydrogenase 8a	123.4	11.0
Alcohol dehydrogenase class-3	13.3	26.3
aldehyde dehydrogenase 2b	17.7	11.4
aldehyde dehydrogenase family 9 member A1-A	15.0	9.2
alpha globin-like	267.0	70.1
alpha-2-macroglobulin receptor-associated protein	42.0	12.0
alpha-actinin-1	160.5	11.5
alpha-aminoadipic semialdehyde dehydrogenase	12.3	15.8
aminoacyl tRNA synthase complex-interacting multifunctional protein 2	13.5	18.4
Aminoacylase-1	55.0	12.0
anamorsin	21.0	19.5
ancient ubiquitous protein 1	17.0	12.5
Angiopoietin-related protein 1	13.5	10.0
Angiopoietin-related protein 4	11.5	9.7
annexin 11a	10.0	46.5
annexin A1	10.7	108.5
annexin A3b	21.7	89.4
anoctamin-10	10.0	9.5
AP-2 complex subunit mu-1-A	68.0	16.9
AP-3 complex subunit sigma-1	15.8	24.0
Apolipoprotein Eb	13.3	765.0
archain 1 like	11.5	66.0
asparagine synthetase	23.3	14.7

aspartate aminotransferase 2a	9.0	137.0
aspartate aminotransferase, cytoplasmic	11.0	14.9
ATP synthase F0 subunit 6	213.5	39.4
ATP synthase subunit b, mitochondrial	20.2	200.5
ATP synthase subunit delta, mitochondrial	9.5	223.5
ATP synthase subunit epsilon, mitochondrial	22.2	46.1
ATP synthase subunit gamma, mitochondrial	13.0	244.8
ATP synthase, H+ transporting, mitochondrial F0 complex, subunit C3	16.4	613.5
ATP synthase-coupling factor 6, mitochondrial	67.4	91.2
ATP-binding cassette sub-family E member 1	21.5	31.5
ATP-binding cassette sub-family F member 1	11.0	18.3
ATP-dependent RNA helicase DDX18	10.0	35.0
ba1 globin, like	32.8	741.0
B-cadherin (Fragment)	23.3	276.0
B-cell receptor-associated protein 31	18.0	14.2
bcl2-associated X protein, a	12.0	29.5
Beta-2-microglobulin	57.0	705.5
beta-hexosaminidase subunit beta	14.5	15.0
bifunctional aminoacyl-tRNA synthetase	11.7	18.4
brain creatine kinase	44.0	16.8
brain protein 44	16.6	114.5
Brain protein 44-like protein	180.7	96.0
brain protein I3-like	9.3	9.2
BTB/POZ domain-containing protein 10	9.0	9.3
bystin	16.0	25.7
cadherin 1, epithelial	12.0	309.0
Cadherin-1	19.5	89.7
calmodulin 2, beta (phosphorylase kinase, delta)	47.5	10.0
calpain, small subunit 1 a	9.2	247.0
calreticulin	129.5	19.0
calumenin-B	52.5	25.7

carbonic anhydrase	11.8	753.5
carbonic anhydrase 9	15.5	18.0
carbonic anhydrase II	9.0	34.1
cardiac muscle alpha actin 1	10.3	257.0
carnitine O-acetyltransferase	22.6	16.5
caspase-6	26.0	36.9
catalase	12.4	9.5
catechol-O-methyltransferase domain containing 1	56.2	9.7
cathepsin B, a	31.0	271.7
cathepsin S, b.1	13.6	322.5
CC chemokine SCYA107	13.8	50.5
CC chemokine SCYA110	10.5	15.0
C-C motif chemokine 19-like	19.4	27.4
C-C motif chemokine 26	11.1	13.2
C-C motif chemokine 3	27.3	78.5
CD59 glycoprotein	28.6	22.5
CD59 glycoprotein-like	67.3	218.4
CD63 antigen	125.5	396.5
CD82 antigen, b	44.5	73.6
CD9 antigen, b	73.5	615.5
CDK5 regulatory subunit-associated protein 3	9.8	9.7
Centrin-1	16.3	10.0
chemokine (C-X-C motif) receptor 7b	11.3	13.5
CHK1 checkpoint-like protein	45.4	294.0
Clathrin light chain B	10.5	16.4
clathrin, heavy polypeptide b (Hc)	12.5	10.3
claudin b	14.0	477.3
cleavage and polyadenylation specificity factor subunit 2	9.0	14.6
cleavage and polyadenylation specificity factor subunit 3	13.5	16.5
coatomer subunit alpha	16.0	21.0
coatomer subunit epsilon	18.8	15.0

coatomer subunit gamma-2	23.8	27.2
Cohesin subunit SA-2	12.0	10.5
coiled-coil domain-containing protein 111	10.0	16.0
coiled-coil domain-containing protein 25	11.4	15.8
Coiled-coil domain-containing protein 86	24.5	12.5
coiled-coil-helix-coiled-coil-helix domain-containing protein 1	17.7	19.5
coiled-coil-helix-coiled-coil-helix domain-containing protein 10, mitochondrial	13.8	30.2
coiled-coil-helix-coiled-coil-helix domain-containing protein 3, mitochondrial	9.0	9.8
Cold shock domain-containing protein E1	9.6	66.3
Complement C1q-like protein 3	23.5	33.0
Complement C4-B	123.4	12.0
complement component 1, q subcomponent, B chain	16.0	35.5
Complement factor H	502.0	9.7
conserved oligomeric Golgi complex subunit 2	63.0	10.5
Cordon-bleu protein-like 1	9.5	9.5
cordon-bleu protein-like 1-like	12.8	20.5
Cornifelin	11.3	35.0
coronin-1A	16.5	126.5
C-X-C motif chemokine 14	18.2	16.4
cyclin-G1	24.8	48.7
cystathionase (cystathionine gamma-lyase)	11.7	30.7
cystathionine-beta-synthase a	22.4	11.5
cysteine-rich with EGF-like domain protein 2	35.8	14.8
cystinosin	13.5	19.5
cytochrome b-245, alpha polypeptide	9.3	18.0
cytochrome b-c1 complex subunit 2, mitochondrial	35.3	29.3
Cytochrome b-c1 complex subunit 9	15.8	71.3
cytochrome c oxidase subunit 4 isoform 1, mitochondrial	350.3	656.5
Cytochrome c oxidase subunit 5A, mitochondrial	9.1	38.0
cytochrome c oxidase subunit I	538.5	55.1
cytochrome c oxidase subunit III	340.0	427.7

cytochrome c-1	255.0	325.5
cytochrome P450, family 2, subfamily AD, polypeptide 2	14.5	11.0
cytochrome P450, family 20, subfamily A, polypeptide 1	9.3	11.3
cytochrome P450, family 3, subfamily A, polypeptide 65	22.7	9.5
cytokeratin-like	13.0	487.5
cytoplasmic aconitate hydratase	9.0	12.0
cytosolic non-specific dipeptidase	9.0	97.0
cytosolic sulfotransferase 3	19.9	10.0
death ligand 3	10.5	9.4
Death-associated protein kinase 3	11.0	16.0
Deleted in malignant brain tumors 1 protein	22.5	48.0
DENN/MADD domain containing 2D	9.0	9.1
deoxyribonuclease I-like 3	12.3	37.7
diamine N-acetyltransferase 1	65.5	13.0
dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex, mitochondrial	10.8	50.5
dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex, mitochondrial	18.8	13.4
dipeptidyl peptidase 1	15.5	25.5
disabled homolog 2	9.5	11.5
dnaJ homolog subfamily B member 12	14.5	14.3
dnaJ homolog subfamily C member 11	12.5	16.0
dnaJ homolog subfamily C member 2	21.4	62.5
DnaJ subfamily A member 2	9.1	134.5
dolichol phosphate-mannose biosynthesis regulatory protein	28.2	14.3
dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 1	37.0	38.0
domain-containing protein 1	24.0	33.5
E3 SUMO-protein ligase NSE2	9.5	10.7
E3 ubiquitin/ISG15 ligase TRIM25	14.3	60.0
E3 ubiquitin-protein ligase MARCH7	11.0	9.5
E3 UFM1-protein ligase 1	10.3	10.3
EH domain-containing protein 1	10.0	9.0

EH domain-containing protein 3	11.0	9.0
elongation factor 1-alpha	203.0	417.5
elongation factor 1-beta	16.6	749.0
elongation factor 1-gamma	22.5	382.5
Elongation factor 2	17.0	790.5
elongation factor-1, delta, b	19.7	24.0
Endonuclease domain-containing 1 protein	14.5	22.0
endophilin-B1	10.0	22.0
Endoplasmic reticulum lectin 1	9.3	24.0
endoplasmic reticulum resident protein 44	43.0	11.0
endoplasmic reticulum-Golgi intermediate compartment protein 2	15.5	12.0
endothelial PAS domain-containing protein 1	10.0	69.3
epidermal retinal dehydrogenase 2	11.0	23.5
epoxide hydrolase 1	12.7	26.5
eukaryotic translation elongation factor 1 alpha 1	699.0	146.0
eukaryotic translation elongation factor 2b	12.7	764.0
eukaryotic translation initiation factor 1A, X-linked, b	9.5	24.0
eukaryotic translation initiation factor 3 subunit 8	10.9	72.0
Eukaryotic translation initiation factor 3 subunit B	254.0	58.5
eukaryotic translation initiation factor 3 subunit D	26.4	20.7
eukaryotic translation initiation factor 3 subunit E-A	17.2	31.3
eukaryotic translation initiation factor 3 subunit G	32.5	292.5
eukaryotic translation initiation factor 3 subunit H-B	11.1	254.0
eukaryotic translation initiation factor 3 subunit I	15.8	110.5
eukaryotic translation initiation factor 3 subunit L	31.5	44.3
eukaryotic translation initiation factor 3 subunit M	46.9	29.7
eukaryotic translation initiation factor 4, gamma 2b	15.0	15.9
eukaryotic translation initiation factor 4E family member 1c	13.7	9.1
eukaryotic translation initiation factor 4E-binding protein 2	11.0	14.8
eukaryotic translation initiation factor 4E-binding protein 3	26.5	11.7
eukaryotic translation initiation factor 5	14.0	10.0

F11 receptor	12.0	33.6
far upstream element-binding protein 3	12.0	10.0
farnesyl pyrophosphate synthase	90.0	18.3
Fatty acid-binding protein, heart	14.0	10.6
F-box only protein 9	11.7	9.8
F-box protein 44	12.0	10.7
ferritin heavy chain	35.3	127.4
Ferritin, middle subunit	58.7	703.0
finTRIM family protein	20.5	12.6
finTRIM family, member 14	9.7	89.0
finTRIM family, member 67	12.2	74.5
flavin reductase	16.4	15.0
follistatin-like 1b	14.0	19.5
fructose-bisphosphate aldolase B	16.7	17.7
FUN14 domain-containing protein 2	14.0	20.7
FXYD domain containing ion transport regulator 5b	25.2	11.5
G1 to S phase transition 1	18.0	13.8
galactoside-binding soluble lectin 9	9.3	19.1
Gamma-aminobutyric acid receptor-associated protein	144.0	43.0
gamma-glutamyl hydrolase	15.2	16.5
Gastrula zinc finger protein xFG20-1	10.0	21.0
GC-rich sequence DNA-binding factor	12.5	11.5
general transcription factor IIE, polypeptide 2, beta	10.0	23.6
glioma tumor suppressor candidate region gene 2 protein	11.8	14.0
glucose phosphate isomerase a	12.9	23.7
glucose phosphate isomerase b	15.2	20.7
glutathione S-transferase pi	15.4	520.0
glutathione S-transferase theta 1b	357.0	22.5
glyceraldehyde-3-phosphate dehydrogenase	9.5	424.0
glyoxalase domain-containing protein 5	12.9	51.0
GMP reductase 2	39.5	15.5

Golgi SNAP receptor complex member 2	45.0	17.4
Grainyhead-like protein 1	14.5	10.4
grancalcin	11.0	19.0
G-rich sequence factor 1	14.0	10.3
growth and transformation-dependent protein	17.7	26.7
growth arrest and DNA-damage-inducible, beta a	180.3	93.0
growth hormone inducible-like protein	20.0	18.6
growth hormone receptor b	11.0	12.0
grpE protein homolog 1, mitochondrial	11.1	13.6
GTPase IMAP family member 4-like	14.0	18.5
GTPase IMAP family member 8-like	14.1	14.7
guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2, like	9.0	25.0
guanine nucleotide-binding protein G(i) subunit alpha-1	19.5	43.7
Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1	9.5	18.4
guanine nucleotide-binding protein subunit beta-2-like 1	26.5	296.3
guanine nucleotide-binding protein-like 3-like protein	16.0	19.3
H/ACA ribonucleoprotein complex subunit 2-like protein	15.5	16.4
H/ACA ribonucleoprotein complex subunit 3	13.0	47.5
H-2 class II histocompatibility antigen, A-D alpha chain	13.5	136.4
H-2 class II histocompatibility antigen, A-K alpha chain	11.8	485.0
heat shock cognate 70 kDa protein	800.5	586.0
heat shock cognate 71 kDa protein	276.5	457.0
heat shock protein 90-alpha 2	55.5	490.0
heat shock protein HSP 90-beta	111.2	833.0
Heme oxygenase	14.9	21.5
heme oxygenase 1	46.5	28.5
heme-binding protein 2	11.3	15.3
Hemicentin-1	13.4	23.0
hemoglobin subunit alpha	12.5	39.9
Hemoglobin subunit beta	38.6	115.3
hemoglobin subunit beta-2	238.8	592.0

Hepatitis B virus X-interacting protein	9.3	14.8
hepatocyte growth factor-regulated tyrosine kinase substrate	11.5	16.0
HERV-H LTR-associating protein 2	68.0	9.0
heterogeneous nuclear ribonucleoprotein A1	13.0	151.4
hexaprenyldihydroxybenzoate methyltransferase, mitochondrial	15.5	15.0
high affinity copper uptake protein 1	10.2	22.0
high-mobility group box 2b	9.3	67.4
Histone H3.3	15.8	370.5
HLA class I histocompatibility antigen, B-47 alpha chain	14.5	11.0
homocysteine-responsive endoplasmic reticulum-resident ubiquitin-like domain member 1 protein	11.0	11.5
hsp90 co-chaperone Cdc37	41.5	79.7
hydroxyacyl-coenzyme A dehydrogenase, mitochondrial	45.7	28.6
hydroxysteroid dehydrogenase-like protein 2	12.4	18.8
hypothetical protein BRAFLDRAFT_76460	142.0	43.7
hypothetical protein LOC100000596	264.3	280.0
hypothetical protein LOC100002844	22.3	12.7
hypothetical protein LOC100037361	290.5	543.0
hypothetical protein LOC100124608	36.0	19.0
hypothetical protein LOC100136852	15.5	70.5
hypothetical protein LOC100147469	636.5	109.6
hypothetical protein LOC100535052, partial	14.0	14.0
hypothetical protein LOC100537730	18.8	32.4
hypothetical protein LOC100538251	11.5	10.5
hypothetical protein LOC556341	21.5	79.6
hypothetical protein LOC568716	10.7	182.5
hypothetical protein LOC792544	9.0	10.5
hypothetical protein RUMLAC_02319	37.0	77.0
hypothetical protein TTHERM_02141640	11.1	22.8
hypothetical protein_XP_002569566.1	161.3	115.5
hypoxia-inducible factor 3-alpha	18.0	38.3
Ig kappa chain C region	18.0	82.0

Ig kappa chain V-III region MOPC 63	9.0	18.0
importin-5	9.6	26.7
importin-7	17.0	11.7
influenza virus NS1A-binding protein	10.8	22.3
inosine-5'-monophosphate dehydrogenase 2	13.2	15.7
insulin-degrading enzyme	17.5	15.1
insulin-like growth factor binding protein 1a	30.3	10.0
insulin-like growth factor binding protein 7	11.0	19.5
integral membrane protein 1	13.7	32.4
Intelectin-1a	176.0	14.5
Inter-alpha-trypsin inhibitor heavy chain H3	179.0	19.5
interferon gamma inducible protein 30	11.1	377.5
Interferon-induced protein 44	10.0	357.5
Interferon-induced transmembrane protein 3	38.5	17.2
Interferon-induced very large GTPase 1	16.0	69.7
Interferon-inducible GTPase 5	13.7	37.5
interleukin-10 receptor subunit beta	21.0	10.6
invariant chain-like protein 1	9.8	211.6
isoleucyl-tRNA synthetase, cytoplasmic	19.0	52.0
isopentenyl-diphosphate Delta-isomerase 1	12.2	12.5
keratinocyte-associated protein 2	183.5	9.5
LAG1 longevity assurance	12.3	10.0
LDLR chaperone MESD	16.5	16.5
lectin, galactoside-binding, soluble, 9 (galectin 9)-like 1	67.0	25.0
legumain	16.3	14.6
leptin receptor gene-related protein	12.0	12.6
leucine-rich repeat-containing protein 8D	13.5	27.0
Leucyl-tRNA synthetase, cytoplasmic	12.0	24.0
Leukocyte elastase inhibitor	12.5	101.0
LIM and SH3 domain protein 1	57.5	82.0
low molecular weight phosphotyrosine protein phosphatase	15.3	12.1

LSM7 homolog, U6 small nuclear RNA associated	12.5	11.5
LYR motif-containing protein 2	11.0	9.0
LYR motif-containing protein 5A	31.7	17.5
lysophospholipid acyltransferase 7	15.5	9.5
lysosomal acid lipase/cholesteryl ester hydrolase	11.9	30.0
lysosomal Pro-X carboxypeptidase	29.5	16.0
Lysosome-associated membrane glycoprotein 1	12.5	28.8
lysozyme g-like 1	9.5	19.9
lysyl-tRNA synthetase	11.4	10.6
major histocompatibility complex class I UDA	60.0	339.3
major histocompatibility complex class I UXA2	9.3	746.5
major histocompatibility complex class I ZE like	16.5	642.5
Mannan-binding lectin serine protease 1	79.3	22.5
mannosyl-oligosaccharide glucosidase	21.0	55.5
matrix metalloproteinase-9	11.0	45.0
Mature T-cell proliferation 1 neighbor protein	50.7	39.7
Membrane-spanning 4-domains subfamily A member 4A	12.7	13.9
Membrane-spanning 4-domains subfamily A member 8A	82.5	30.1
membrane-spanning 4-domains, subfamily A, member 17A.1	10.0	29.3
membrane-spanning 4-domains, subfamily A, member 4-like	14.7	56.0
Metallothionein	21.8	41.3
methionine adenosyltransferase II, alpha	35.6	405.5
methionyl-tRNA synthetase, cytoplasmic	9.2	836.0
methylmalonic aciduria (cobalamin deficiency) cblD type, with homocystinuria	9.3	15.0
methyltransferase Mb3374	46.5	9.5
methyltransferase-like protein 10	18.7	11.0
methyltransferase-like protein 9	16.5	11.3
MHC class I alpha chain	13.0	720.0
MHC class II integral membrane protein alpha chain 1	52.5	13.0
Microfibril-associated glycoprotein 4	805.0	120.7
microfibrillar-associated protein 4	456.5	62.2

mid1-interacting protein 1-B	11.6	14.1
mitochondrial 2-oxoglutarate/malate carrier protein	20.5	13.7
mitochondrial carrier	11.7	57.5
mitochondrial import inner membrane translocase subunit tim16	43.3	32.0
Mitochondrial import inner membrane translocase subunit TIM44	12.0	20.0
mitochondrial import inner membrane translocase subunit Tim8 B	34.5	9.6
mitochondrial import receptor subunit TOM20	22.4	33.4
mitochondrial inner membrane protein OXA1L	12.8	18.8
mitochondrial trifunctional protein, alpha subunit	10.6	90.5
monocarboxylate transporter 10	10.0	10.0
M-phase phosphoprotein 6	18.5	12.7
Multidrug resistance protein 1	9.5	37.0
myb-binding protein 1A-like protein	43.0	11.0
myeloid cell leukemia sequence 1-like	9.4	58.5
Myosin light chain kinase, smooth muscle	31.0	48.0
Myosin-9	10.0	226.0
N-acetylglucosamine-1-phosphotransferase subunit gamma	9.6	9.5
NACHT, LRR and PYD domains-containing protein 12	10.8	397.5
NACHT, LRR and PYD domains-containing protein 3	11.0	148.0
NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1	11.5	35.0
NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 5	21.2	40.0
NADH dehydrogenase [ubiquinone] 1 subunit C1, mitochondrial	56.0	23.5
NADH dehydrogenase [ubiquinone] flavoprotein 1, mitochondrial	22.0	25.7
NADH dehydrogenase [ubiquinone] flavoprotein 3, mitochondrial	19.0	22.5
NADH dehydrogenase subunit 2	149.0	112.9
NADH dehydrogenase subunit 4	167.3	207.8
Nedd4 family interacting protein 1	16.7	49.7
NEDD4 family-interacting protein 1-like	10.2	12.3
NEDD4 family-interacting protein 2	12.1	16.0
neural proliferation, differentiation and control, 1	10.0	13.0
Neuroblast differentiation-associated protein AHNAK	18.5	369.0

neutrophil cytosol factor 1	14.5	12.8
Nidogen-1	17.7	9.8
nitric oxide synthase-interacting protein	11.0	11.8
nonspecific cytotoxic cell receptor protein 1	9.4	54.8
Nuclear factor 7, ovary	164.5	24.5
nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha a	12.6	25.8
Nuclear prelamin A recognition factor	20.8	18.3
nuclear transcription factor Y, beta	15.0	30.0
nucleolar protein 56	12.1	84.5
nucleolar RNA helicase 2	10.5	290.3
nucleolin	11.8	12.1
nucleoplasmin-3	30.0	28.0
obg-like ATPase 1	34.0	24.5
omega-amidase NIT2	27.0	12.5
optineurin	21.0	20.0
ORF2-encoded protein	10.8	125.0
oxysterol-binding protein 1	9.0	18.8
oxysterol-binding protein-related protein 2	23.0	11.8
P2Y purinoceptor 2	9.8	16.0
pancreatic progenitor cell differentiation and proliferation factor B	14.3	66.0
peptidyl-prolyl cis-trans isomerase A	12.0	25.5
Perforin-1	11.2	10.5
perilipin-2	11.7	29.5
peroxiredoxin-2	47.0	9.7
peroxisomal multifunctional enzyme type 2	10.2	18.5
peroxisomal trans-2-enoyl-CoA reductase	15.0	12.5
PHD finger-like domain-containing protein 5A	16.7	51.0
phenylalanyl-tRNA synthetase beta chain	9.6	14.5
Phosphatidylinositol-binding clathrin assembly protein	11.5	9.0
phosphatidylserine synthase 1	14.3	22.0
phosphoenolpyruvate carboxykinase [GTP], mitochondrial	59.0	9.1

phosphoethanolamine methyltransferase	76.0	15.0
phosphoglycerate kinase 1	13.4	22.8
phospholipid transfer protein	16.4	9.4
Poly(ADP-ribose) glycohydrolase	11.0	37.0
polyadenylate-binding protein 4	72.7	10.5
polymerase (RNA) II (DNA directed) polypeptide F	32.0	17.7
polymerase (RNA) II (DNA directed) polypeptide J	13.0	9.0
Polypeptide N-acetylgalactosaminyltransferase 5	13.0	10.5
prefoldin subunit 2	16.3	21.3
prefoldin subunit 3	26.8	51.3
Prefoldin subunit 4	9.2	10.5
pre-mRNA 3'-end-processing factor FIP1	13.3	10.7
pre-mRNA branch site protein p14	19.1	23.2
prestin	15.0	25.5
Probable Bax inhibitor 1	396.0	277.0
probable ribosome biogenesis protein NEP1	10.5	61.5
probable ribosome biogenesis protein RLP24	16.8	77.5
probable rRNA-processing protein EBP2	30.0	32.0
profilin-2	14.6	14.3
programmed cell death 4a	14.5	25.0
programmed cell death protein 10-B	13.0	16.0
prohibitin 2	24.3	14.9
proliferating cell nuclear antigen	12.1	38.3
prostaglandin reductase 1	13.0	20.5
proteasomal ubiquitin receptor ADRM1	15.8	11.4
proteasome (prosome, macropain) 26S subunit, ATPase, 1	10.1	11.6
proteasome 26S subunit, ATPase, 4	11.8	14.0
proteasome subunit alpha type-3	10.6	16.1
proteasome subunit alpha type-7-like	35.8	20.1
proteasome subunit beta type-3	36.4	30.9
proteasome subunit beta type-4	23.0	16.4

proteasome subunit beta type-5	16.3	11.8
proteasome subunit beta type-6	22.4	17.2
protein BUD31	34.5	26.8
Protein C16orf88	15.0	22.0
protein canopy	9.2	23.0
protein CDV3	25.7	35.8
protein disulfide-isomerase	16.0	16.2
protein disulfide-isomerase A4	10.3	10.5
Protein FADD	9.0	11.6
Protein FAM195B	10.5	10.2
Protein LYRIC	11.5	22.5
protein MIS12	10.0	13.5
protein NDRG2	17.8	18.0
Protein NLRC3	503.3	394.3
protein phosphatase 1, catalytic subunit, alpha-like	10.0	21.6
protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), alpha	19.0	23.6
Protein prune	10.5	11.5
protein QIL1	15.8	30.7
Protein SCAF11	9.0	32.0
protein SCO2 homolog, mitochondrial	13.0	10.5
protein SDA1	13.0	19.5
protein Tob1	12.0	15.6
protein transport protein Sec61 subunit alpha-like 2	11.2	38.3
protein-kinase, interferon-inducible double stranded RNA dependent inhibitor	14.5	10.3
Pumilio	10.5	10.3
Pumilio domain-containing protein KIAA0020	12.4	26.3
purine nucleoside phosphorylase 5a	17.6	20.5
putative all-trans-retinol 13,14-reductase	29.5	26.0
putative oxidoreductase GLYR1	9.0	16.0
Putative uncharacterized protein ART2	68.6	326.5
pyridoxal (pyridoxine, vitamin B6) kinase a	22.5	86.5

pyruvate dehydrogenase E1 alpha 1	16.0	30.1
rab3 GTPase-activating protein catalytic subunit	9.0	9.8
RAB5A, member RAS oncogene family, b	12.0	9.3
Rano class II histocompatibility antigen, D-1 beta chain	17.5	820.5
ras homolog gene family, member Ad	12.0	26.5
ras-related C3 botulinum toxin substrate 1	11.0	108.5
ras-related protein Rab-1A	120.0	12.0
Ras-related protein Rab-35	12.5	19.5
ras-related protein Rab-7a	10.5	86.8
RecQ-mediated genome instability protein 2	29.0	30.0
retinol-binding protein 4	87.2	78.0
reverse transcriptase	10.0	106.2
rho GDP-dissociation inhibitor 1	18.5	18.3
Rho-class glutathione S-transferase	44.4	27.4
Ribonuclease inhibitor	25.0	153.0
ribonuclease kappa-A	9.3	13.4
ribonuclease like 2	20.3	12.7
ribosome biogenesis protein bop1	9.3	20.0
ribosome biogenesis protein NSA2	12.9	309.5
ribosome biogenesis regulatory protein	18.5	9.5
Ribosome-binding protein 1	21.3	11.0
RING finger protein 141	11.0	10.0
RING-box protein 2	9.0	11.1
RLA class II histocompatibility antigen, DP alpha-1 chain (Fragment)	9.4	791.5
RNA binding protein with multiple splicing 2	9.5	17.0
RNA polymerase II subunit A C-terminal domain phosphatase SSU72	9.3	12.0
RNA-binding protein	9.7	27.0
RNA-binding protein FUS	10.5	21.0
rRNA promoter binding protein	60.1	421.3
ruvB-like 1	10.3	17.9
sarcosine dehydrogenase, mitochondrial	17.5	9.5

sel1 repeat-containing protein 1	11.5	9.7
selenophosphate synthetase 2	10.5	101.0
selenoprotein H	9.8	43.5
selenoprotein Pa	31.8	13.0
selenoprotein S	25.0	13.7
selenoprotein W	49.0	410.0
selenoprotein W, 2b	10.0	327.3
septin-7	11.3	70.7
serglycin	9.5	567.0
serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 6	21.0	321.7
serine/arginine repetitive matrix 2	9.5	14.8
Serine/threonine-protein kinase/endoribonuclease IRE1	10.3	17.5
serine/threonine-protein phosphatase 2A activator	18.7	12.5
Serine/threonine-protein phosphatase PP1-gamma catalytic subunit	9.5	15.3
Serpin B6	10.0	181.0
Serum amyloid P-component	9.0	17.3
SH3 domain-binding glutamic acid-rich-like protein	11.0	60.3
SH3 domain-containing kinase-binding protein 1	10.5	70.0
si:dkeyp-110c7.1	15.0	22.0
si:rp71-45k5.4	9.0	9.3
signal peptidase complex subunit 1	30.3	19.6
signal peptidase complex subunit 3	198.5	12.1
signal recognition particle 9	12.1	9.8
signal transducer and activator of transcription 3	23.0	42.5
small glutamine-rich tetratricopeptide repeat-containing protein alpha	9.8	26.0
small nuclear ribonucleoprotein polypeptide F-like	17.0	66.7
small nuclear ribonucleoprotein polypeptides B and B1	28.5	60.8
solute carrier family 16 (monocarboxylic acid transporters), member 8	20.0	11.5
solute carrier family 25 member 3	186.0	215.6
Solute carrier family 35 member B1	11.2	14.8
sorting and assembly machinery component 50	12.0	59.5

spectrin beta chain, brain 1	9.8	10.8
spermidine/spermine N1-acetyltransferase	133.5	164.7
S-phase kinase-associated protein 1	86.5	43.0
spindle and kinetochore-associated protein 2	9.5	9.9
splicing factor 3A subunit 1	10.0	10.7
splicing factor 3B subunit 2	11.0	35.5
splicing factor, proline- and glutamine-rich	9.5	12.4
staphylococcal nuclease domain-containing protein 1	42.5	15.6
steroid receptor RNA activator 1	9.3	9.3
Sterol 26-hydroxylase, mitochondrial	96.0	23.5
sterol regulatory element-binding protein 2	14.1	9.7
sugar transporter SWEET1	11.5	10.0
sulfotransferase family 2, cytosolic sulfotransferase 2	9.5	10.4
superoxide dismutase [Cu-Zn]	54.5	16.5
superoxide dismutase [Mn], mitochondrial	42.0	64.7
suppressor of cytokine signaling 3a	19.0	12.1
surfeit gene 4, like	24.0	19.5
survival of motor neuron-related-splicing factor 30	16.0	18.8
synaptojanin-2-binding protein	19.5	21.3
synaptosomal-associated protein 23	15.5	13.7
synaptotagmin binding, cytoplasmic RNA interacting protein, like	13.5	10.9
TATA-box-binding protein	10.0	9.3
T-cell leukemia translocation-altered gene protein	14.4	22.0
T-complex protein 1 subunit alpha	236.3	646.0
T-complex protein 1 subunit beta	19.9	288.0
T-complex protein 1 subunit delta	9.7	221.5
T-complex protein 1 subunit eta	63.3	64.8
T-complex protein 1 subunit gamma	38.3	41.9
T-complex protein 1 subunit theta	9.8	102.0
telomeric repeat-binding factor 2	18.0	22.0
tetraspanin-9	14.1	16.0

tetratricopeptide repeat protein 1	12.5	17.3
tetratricopeptide repeat protein 32	9.0	14.0
thimet oligopeptidase	13.0	11.3
thioredoxin	20.8	16.4
thioredoxin domain containing 1	11.0	13.0
thioredoxin domain-containing protein 12	26.0	9.3
thioredoxin domain-containing protein 5	35.0	19.8
thioredoxin domain-containing protein 9	13.0	27.3
thioredoxin, mitochondrial	10.8	48.0
TNF receptor	11.3	24.0
toll-interacting protein	11.0	30.0
trafficking protein particle complex 2-like	11.1	33.7
trafficking protein particle complex subunit 4	13.5	11.4
trafficking protein particle complex subunit 5	13.0	16.0
TRAF-interacting protein with FHA domain-containing protein A	9.3	34.0
transcription initiation factor TFIID subunit 7	22.5	21.0
transferrin receptor 1a	13.5	10.5
transitional endoplasmic reticulum ATPase	63.5	9.4
transketolase-like protein 2	25.3	40.8
translationally-controlled tumor protein	108.9	211.0
translocating chain-associated membrane protein 1	10.7	12.6
translocon-associated protein subunit alpha	37.8	12.5
Translocon-associated protein subunit gamma	14.1	10.6
transmembrane emp24 domain-containing protein 9	48.0	19.6
Transmembrane protein 104	11.5	9.3
transmembrane protein 147	9.0	10.3
transmembrane protein 18	11.1	18.5
transmembrane protein 180	10.5	13.0
Transmembrane protein 214-A	13.0	12.0
Transmembrane protein 70, mitochondrial	13.0	13.0
Transmembrane protein 79	9.3	30.5

transmembrane protein 93	15.7	45.0
Transposable element Tc1 transposase	14.0	70.0
Transposable element Tcb1 transposase	18.5	69.0
Transposable element Tcb2 transposase	17.5	83.7
transposase	18.5	58.0
triosephosphate isomerase A	62.3	108.5
triosephosphate isomerase B	231.3	236.3
Tripartite motif-containing protein 29	60.0	74.5
Tripartite motif-containing protein 39	11.0	258.0
tRNA-dihydrouridine synthase 3-like	10.5	10.3
tsukushin	9.0	10.5
tubulin alpha-1B chain	11.7	393.5
tubulin alpha-1C chain	13.3	136.3
tubulin-folding cofactor B	12.3	13.5
tumor necrosis factor receptor superfamily member 1A	15.5	54.5
tumor protein D54	9.5	12.3
tumor protein p53-inducible nuclear protein 1	10.0	9.7
type I cytokeratin, enveloping layer	15.5	421.3
tyrosine-protein phosphatase non-receptor type 1	12.5	13.5
ubiquinol-cytochrome c reductase complex chaperone	10.0	26.5
ubiquitin carboxyl-terminal hydrolase 14	9.2	59.5
ubiquitin carboxyl-terminal hydrolase 16	9.0	15.0
ubiquitin-conjugating enzyme E2 K	9.0	14.0
ubiquitin-fold modifier 1	15.5	12.2
ubuquitin c	79.8	625.0
UDP glucuronosyltransferase 2 family, polypeptide A6	86.7	10.7
UDP-glucose pyrophosphorylase 2	88.8	11.7
Uncharacterized gene 87 protein	11.0	12.1
Uncharacterized protein C18orf19	11.5	9.7
uncharacterized protein c1orf43-like protein	9.0	19.0
Uncharacterized protein C20orf4	12.3	9.8

Uncharacterized protein C7orf50	9.5	23.5
Uncharacterized protein_CBN81934.1	33.0	9.3
unnamed protein product_CAG06282.1	9.0	9.7
unnamed protein product_CAG11942.1	9.0	12.0
upf0389 protein fam162b	17.3	15.5
UPF0451 protein C17orf61	15.3	14.3
UPF0556 protein C19orf10	29.5	30.0
UPF0729 protein C18orf32	12.2	11.8
up-regulated during skeletal muscle growth protein 5	21.3	14.5
Urokinase plasminogen activator surface receptor	25.5	357.3
vacuolar protein sorting-associated protein 29	12.0	11.7
vacuole membrane protein 1	10.4	12.7
vertebrate transmembrane 4 superfamily-like	11.8	9.4
very long-chain specific acyl-CoA dehydrogenase, mitochondrial	18.3	25.3
vesicle transport protein SFT2A	11.4	20.3
Vesicle-trafficking protein SEC22b-B	35.7	19.3
Vigilin	101.0	41.0
villin 2	16.0	97.0
Vitelline membrane outer layer protein 1	44.3	357.5
voltage-dependent anion-selective channel protein 2	10.8	577.5
WAS/WASL-interacting protein family member 2	9.5	34.7
WD repeat-containing protein 85	13.0	10.5
WD repeat-containing protein mio	9.5	15.0
WW domain binding protein 2	9.8	142.7
xaa-Pro aminopeptidase 1	9.0	15.0
xaa-Pro dipeptidase	10.3	34.5
X-box-binding protein 1	20.9	27.2
zinc finger CCHC domain-containing protein 9	16.7	10.6
zinc transporter 6	9.5	9.0