Assembly and Annotation of the Channel Catfish Transcriptome and Assessment of Pervasive Expression

by

Chen Jiang

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
December 12, 2015

Key words: transcriptome; pervasive transcription; gene; long non-coding RNAs;
induced expression; correlated expression

Approved by

Zhanjiang Liu, Chair, Professor of School of Fisheries, Aquaculture, and Aquatic Sciences
Rex Dunham, Professor of School of Fisheries, Aquaculture, and Aquatic Sciences
Nannan Liu, Professor of Entomology and Plant Pathology
Charles Y. Chen, Associate Professor of Crop, Soil and Environmental Sciences

Abstract

For a long period of time, it was widely believed that only a small fraction of the genome is transcribed, and researchers were mainly focused on studying expression of protein-coding genes. In the last a few years, pervasive transcription became highly apparent from mounting evidence, therefore increasing attention has been given to the entire transcriptome including both coding and non-coding regions. However, the ability to assess the extent of pervasive transcription and expression depends on continuous and thorough analysis of transcriptome over space, time and various conditions.

Catfish is the primary aquaculture species in the United States. However, its transcriptome has not been well characterized. In recent years, a number of RNA-Seq studies have been conducted, most of which have had a focus of expression profiling under specific stress conditions such as after disease infection, under high temperature exposure, or under hypoxic conditions. With these studies, a large number of RNA-Seq reads became available, totaling approximately six billion reads. These datasets make it possible to assemble a reference transcriptome for channel catfish. At the same time, these RNA-Seq were conducted using various tissues, allowing analysis of transcriptome level of expression profiling among various tissues. Likewise, systematic analysis of stress-induced expression is possible using these RNA-Seq datasets obtained after various stress treatment. The objectives of this study were to 1) assemble a reference transcriptome using all channel catfish RNA-Seq datasets; 2) annotate the protein-coding genes from the transcriptome; 3) identify a set of full length transcripts from the transcriptome; 4) analyze expression patterns of protein-coding gene along the channel catfish genome; 5) identify long non-coding RNAs and determine their expression patterns; 6) assess the extent of pervasive transcription in channel catfish; and 7) identify correlated expression of protein-coding genes and long non-coding RNAs.

The reference transcriptome was assembled using both the *de novo* and the genome-guided assembly approaches. A total of 27,448 protein-coding genes were identified, of which 25,489 were homologous genes to known genes in other species, and 1,959 were unknown genes. Of the 27,448 protein-coding genes, 800 genes were not included in the catfish genome. Of all the protein-coding genes, full length transcripts were reconstructed for 20,371 genes. In addition to the protein-coding genes, a total of 36,266 long non-coding RNAs were also assembled and identified. Through mapping of all the short reads, coding or non-coding, to the catfish genome, 79.7% of the channel catfish genome was found to be transcribed.

Mapping of the short reads to the genome allowed analysis of tissue-specific and stress-induced protein-coding genes as well as lncRNAs. A total of 1,455 genes and 2,599 lncRNAs were observed to be expressed in a tissue-specific manner, while 8,560 genes and 748 lncRNAs were differentially expressed after stress treatments such as disease infections, high temperature and short-term deprivation. Notably, the expression of 45 co-induced co-localized genes and lncRNAs sets were identified in this study, suggesting coordinated regulation of the protein-coding genes and lncRNAs.

My dissertation work accomplished the set goals. The reference transcriptome will be a great resource for functional research and digital gene expression analysis in catfish. The full length transcripts will provide further assistance for improvement of genome annotation and constructions of intense phylogenetic analysis or structural analysis of orthologies. The identified set of tissue-specific genes and lncRNAs enabled greater understanding of organismal development, complexity at the system level. The identification of the lncRNAs followed with the initial characterization of expression profiles along with the protein−coding genes could contribute to the future understanding of the function and mechanisms of lncRNAs.

Acknowledgments

First and foremost, I would like to express my deep and sincere gratitude to my advisor Dr. Zhanjiang Liu for his constant guidance, support and patience throughout my degree program. His scientific attitude and philosophy have had a profound influence on me. Under his supervision, I learned how to do scientific research. His encouragement and support certainly help me in my determination to finish this dissertation on time. I am also grateful to Dr. Rex Dunham, Dr. Nannan Liu and Dr. Charles Y. Chen for being on my committee. Thanks Dr. Timothy Mcdonald for being my dissertation reader. Their suggestions and feedbacks are greatly improving my research. Sincere thanks will also go to Dr. Huseyin Kucuktas and Ludmilla Kaltenboeck for their technical assistance, and to Dr. Ruijia Wang, Dr. Shikai Liu, Dr. Chao Li, Dr. Jiaren Zhang, Dr. Yu Zhang, Dr. Luyang Sun, Dr. Jun Yao, Dr. Yun Li, Dr. Lin Song, Lisui Bao, Xin Geng, Qifan Zeng, Xiaozhu Wang, Ning Li, Yulin Jin and all the other colleagues in the laboratory for their help, collaboration, and friendship, especially for Dr. Ruijia Wang's assistance for the analysis of long non-coding RNA. I would like to also express my gratitude to the Chinese Scholarship Council for the scholarship that made my studies possible.

Finally and as always, I want to give a special acknowledgement to my parents and my boyfriend for their love, support and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

Transcriptome is the entire RNA of a species including mRNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs; mainly tRNAs and rRNAs), signal recognition particle (7SL/SRP) RNAs, microRNAs (miRNAs), small interfering RNAs (siRNAs), piwi RNAs (piRNAs) and *trans*-acting siRNAs (ta-siRNAs), natural *cis*-acting siRNAs and non-coding RNAs. The term is often used to also describe the entire RNA in a specific cell, of a specific developmental stage, under a specific physiological condition, or under a specific environment. As depicted in the central dogma of genetics, genome (the entire DNA of an organism) is transcribed into RNA, and RNA is translated into proteins for biological functions. This seemingly correct central dogma, however, is increasingly challenged by the complexities of the transcriptome. While the genome of an organism is relatively stable (with exceptions of certain cell types such as recombination of immunoglobulins in certain types of immune cells), transcriptome is dynamic. Each cell type expresses specific set of genes in a tissue-specific manner, and each gene is expressed at different levels, and such baseline expression is regulated through various factors including development, physiology, and the environment. The many types of non-coding RNAs themselves can regulate the composition and profiles of the transcriptome. Therefore, Studies on transcriptomes could provide us insights into the functional elements of the genome, reveal the molecular constituents of cells or tissues, and understand the mechanism of development, and responses to various environmental stresses such as diseases, high temperature and low oxygen.

Back a quarter of century ago when the International Human Genome Sequencing Consortium started the human genome project, it was estimated that the total length covered by the coding exons is only about 1.2% of the human euchromatic genome (Consortium, 2004).

However, it is reported the first time by ENCODE pilot project that 93% of the human genome is transcribed (Birney et al., 2007). As such, the human genome is pervasively transcribed. Even though different opinions exist as to the extent of pervasive transcription (Clark et al., 2011; van Bakel et al., 2010, 2011), increasing evidence supports the notion of pervasive transcription. The vast majority of the genome is indeed transcribed and the previously dismissed "dark matter" transcripts are being demonstrated being more than "transcriptional noise". The non-coding transcripts have important functions in gene regulation and genome evolution (Wade and Grainger, 2014). In addition to humans, pervasive transcription has been also observed in many other organisms. For example, over 85% of *Saccharomyces cerevisiae* genome is transcribed (David et al., 2006).

Of the pervasively transcribed genome, the fraction being translated into proteins is very small. Large numbers of non-coding RNAs have been discovered, especially from mammals (Hangauer et al., 2013; Rinn et al., 2007; Tripathi et al., 2010). Various types of non-coding RNAs have been identified including long non-coding RNAs (lncRNAs), microRNAs (miRNAs), short interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), etc (Kapranov 2007). Among these non-coding RNAs, lncRNAs are non-coding RNAs whose sizes are greater than 200 bases. Among these non-coding RNAs, long non-coding RNAs drew the most attention since it can be easily discovered with high confidence from existing RNA-Seq datasets and correlated with gene expression information from the same dataset using existing bioinformatics tools (Bawa et al., 2015). Long non-coding RNAs have been reported to play essential roles in regulating gene expression and affect various biological processes (Bawa et al., 2015; Hu et al., 2013; Huarte and Rinn, 2010; Hung et al., 2011; Rinn and Chang, 2012).

Channel catfish, *Ictalurus punctatus*, is the most important species of aquatic animal that commercially cultured in the United States. In recent years, the catfish industry has encountered unprecedented challenges due to outbreaks of diseases, international competition, and increased feed and energy costs. Such challenges caused big loss and threatened the sustainability of the catfish industry. Recently, large amount of efforts has been devoted to generation of transcripts using RNA-Seq technology under different disease and stress conditions in different tissues, trying to uncover the molecular mechanism underlying different conditions in order to provide insight into strategies for disease management and selection. In addition, analyses of biological characteristics were also conducted such as scales, barbels, sex determination, and sex differentiation. However, a comprehensive transcriptome of catfish has not been developed. Recently, the application of the next generation sequencing technologies has allowed rapid generation of RNA-Seq datasets that provide rich resources for constructing a comprehensive and relatively complete set of transcriptome. The main objectives of my dissertation are to develop a relatively complete and well-annotated transcriptome resource for channel catfish, and analyze expression of genes and long non-coding RNAs in relation to various environmental conditions. Toward this end, a total of 13 channel catfish RNA-Seq datasets containing approximately 4.8 billion reads, representing all the tissues, various developmental stages, and various environmental stress conditions, were collected to perform a comprehensive transcriptome assembly for transcriptome analysis and assessment of pervasive expression.

The long-term goal of this project is to construct a comprehensive and relatively complete transcriptome of channel catfish, to determine the extent to which catfish genome is transcribed, to establish tissue expression profiles, and to examine induced transcription under various stress conditions such as hypoxia, high temperature and disease infections. Understanding the control of gene expression is critical for our understanding of the relationship between genotype and phenotype.

For my dissertation project, I had the following specific objectives: 1) Assemble a comprehensive transcriptome using all 13 channel catfish RNA-Seq datasets including all the tissues, various developmental stages, and various environmental stress conditions; 2) Annotation for the protein-coding genes from the transcriptome; 3) Identification of the full length transcripts from the transcriptome; 4) Analysis of expression pattern of protein-coding gene along the channel catfish genome; 5) Assessment of pervasive transcription in channel catfish; 6) Identification of long non-coding RNAs and determine their expression patterns; 7) Identification of correlated expression of protein-coding genes with those of long non-coding RNAs to identify potential target genes of the long non-coding RNAs in their regulation of gene expression, as well as the positional correlations of these RNAs.

Reference

Bawa, P., Zackaria, S., Verma, M., Gupta, S., Srivatsan, R., Chaudhary, B., Srinivasan, S., 2015. Integrative analysis of normal long intergenic non-coding RNAs in prostate cancer. PloS one 10, e0122143.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., Morris, K.V., Morillon, A., Rozowsky, J.S., Gerstein, M.B., Wahlestedt, C., Hayashizaki, Y., Carninci, P., Gingeras, T.R., Mattick, J.S., 2011. The Reality of Pervasive Transcription. PLoS biology 9.

Consortium, I.H.G.S., 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931-945.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., Steinmetz, L.M., 2006. A high-resolution map of transcription in the yeast genome. Proceedings of the National Academy of Sciences of the United States of America 103, 5320-5325.

Hangauer, M.J., Vaughn, I.W., McManus, M.T., 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS genetics 9, e1003569.

Hu, G., Tang, Q., Sharma, S., Yu, F., Escobar, T.M., Muljo, S.A., Zhu, J., Zhao, K., 2013. Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. Nature immunology 14, 1190-1198.

Huarte, M., Rinn, J.L., 2010. Large non-coding RNAs: missing links in cancer? Human molecular genetics 19, R152-161.

Hung, T., Wang, Y.L., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., Wang, Y., Kong, B., Langerod, A., Borresen-Dale, A.L., Kim, S.K., van de Vijver, M., Sukumar, S., Whitfield, M.L., Kellis, M., Xiong, Y., Wong, D.J., Chang, H.Y., 2011. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nature genetics 43, 621-U196.

Rinn, J.L., Chang, H.Y., 2012. Genome regulation by long noncoding RNAs. Annual review of biochemistry 81, 145-166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., Chang, H.Y., 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129, 1311-1323.

Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., Blencowe, B.J., Prasanth, S.G., Prasanth, K.V., 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Molecular cell 39, 925-938.

van Bakel, H., Nislow, C., Blencowe, B.J., Hughes, T.R., 2010. Most "Dark Matter" Transcripts Are Associated With Known Genes. PLoS biology 8, e1000371.

van Bakel, H., Nislow, C., Blencowe, B.J., Hughes, T.R., 2011. Response to "The Reality of Pervasive Transcription". PLoS biology 9, 1423.

Wade, J.T., Grainger, D.C., 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nature reviews. Microbiology 12, 647-653.

Chapter II

Literature review

RNA-Seq

RNA sequencing (RNA-Seq) is a relatively new, high-throughput method for comprehensive transcriptome analysis that uses deep-sequencing technologies. It allows measurement of expression levels of tens of thousands of genes simultaneously and provide insight into functional pathways and regulations in biological processes (Khatoon et al., 2014). Although technically complex, RNA-Seq is just the application of the next generation sequencing technologies such as Illumina sequencing using RNA as templates. Cellular RNA is extracted, processed, and convert to cDNA, then sheared to form short segments for sequencing.

Since RNA-Seq is entirely based on the general principles of DNA sequencing, the methodology is applicable to any organism, subject to the availability of a sufficient amount of RNA. The vast majority of RNA (>90%) present in cells consists of ribosomal RNA (rRNA). In order to avoid wasting effort in re-sequencing the same ribosomal RNA millions of times, polyA-enriched step was usually recommended to selectively remove ribosomal RNA, so that 100-200 ng of polyA enriched RNA is used for double-stranded cDNA synthesis prior to sequencing (Wilhelm and Landry, 2009). However, when dealing with low quality RNA, rRNA depletion is the recommended approach to get rRNA removed (Vikman et al., 2014). The RNA is then fragmented and made into cDNA library that is subsequently sequenced. This can either be done as a paired end (PE), meaning that the fragment is sequenced from both ends, or as a single end (SE), meaning that it is only sequenced from one side. The resulted reads are typically 100-150 bp, although the initial read length a few years ago was only 36 bp, and the read length is now enhanced over time up to 400 bp. After sequencing, the resulting reads are either aligned to a reference genome or reference transcriptome, or assembled *de novo* without the genomic sequence to assess the expression profiles.

RNA-Seq technique has revolutionized the way researchers examine the transcriptome. It can not only detect the precise expression levels of thousands of genes simultaneously, but also identify novel transcripts, miRNAs, and fusion genes. Unlike hybridization-based approaches, RNA-Seq is not limited to detecting transcripts that correspond to existing genomic sequence, which makes attractive for non-model organisms without genomic sequences. In addition, RNA-Seq has very low background signal compared to microarray (Wang et al., 2009). RNA-Seq can reveal the fine structure of the transcriptome with a single nucleotide resolution, which can help identify allele specific expression, alternative splicing, and SNPs in the transcribed regions (Khatoon et al., 2014).

Pervasive transcription studies

The term "pervasive transcription" referred to widespread transcription of DNA into RNA along the genome, way more than the protein-coding genes to include almost all genomic regions including both the sense strand and the antisense strand. Analysis of pervasive transcription would allow the generation of assemblies of different RNAs including not only the traditional protein-coding genes, those with established functions like tRNAs, rRNAs, snRNAs, and snoRNAs (Jensen et al., 2013), but also those whose functions are unknown such as microRNA, lncRNA, and various types of RNA species. Over a decade ago, the international human genome sequencing consortium estimated that only about 1.2% of the human euchromatic genome codes for protein (Consortium, 2004). Studies at early times believed that only about 5-10 % of the human genome is stably transcribed in cell lines (Pertea, 2012). Beginning in the early 2000s, accumulating studies suggested that the vast majority of the genome is transcribed at some time point (Bertone et al., 2004; Cheng et al., 2005; Kapranov et al., 2002). In 2007, the pilot phase of the ENCODE Project reported that as much as 93% of the human genome is transcribed in at least one cell type (Birney et al., 2007). However, scientists did not have the same opinion at that time. Van Bakel et al. (2010) claimed that most of the genome is not appreciably transcribed, and the majority of intergenic and intronic transcripts observed might be caused by biological artifacts from improperly processed RNAs or technical background noise produced by high false-positive rate of tiling array technology (van Bakel et al., 2010). Their conclusion was soon being challenged by another groups of scientists, they claimed that the pervasive transcription is observed in multiple independent techniques including RT-PCR, RACE, and Northern blot analyses, and they pointed out that Bakel et al.'s RNA-seq data suffers from insufficient sequencing depth and poor assembly (Clark et al., 2011). Soon in a subsequent paper, van Bakel et al. did not dispute the fact that much of the genome is transcribed but claimed that they observed that the abundance of these "dark matter" transcripts was low, and the number of well-supported independent RNAs was still relatively small, however given various source of enough reads and sequencing depth, the whole genome may be covered with transcripts (Pertea, 2012; van Bakel et al., 2011). Furthermore, the sequence reads that were previously dismissed as noise were found out to be indicative of unassembled rare transcripts (Mercer et al., 2012).

The studies concerning pervasive transcription was only conducted in humans and a few model species. For instance, by quantifying RNA expression on both strands of the complete genome of *Saccharomyces cerevisiae* using a high-density oligonucleotide tiling array, a total of 85% of the yeast genome is expressed in rich media (David et al., 2006). Besides the prediction using the genome tiling array by the ENCODE project, RNA-Seq datasets also showed an estimation that 85.2% of the human genome is transcribed, using RNA-Seq reads mapping to genome as well as additional evidence from full structures of known genes, spliced ESTs and cDNAs (Hangauer et al., 2013). In addition, analysis of full-length cDNAs from mouse using many tissues and developmental stages showed at least 63% of the mouse genome is transcribed (Carninci et al., 2005; Clark et al., 2011; Katayama et al., 2005; Okazaki et al., 2002). There were also reports claimed that estimation of range of pervasive transcription for nematode worm is 70% and 85% for fruit fly (Dinger et al., 2009). In spite of the complexity and the importance, analysis of pervasive transcription in teleost fish has largely been lacking. No studies have been conducted with any of the aquaculture species including catfish.

Application of RNA-Seq for various biological studies

The literature on RNA-Seq analysis is rapidly growing. A quick search of the PubMed database indicated publication over 5,000 papers involving RNA-seq. Rather than reviewing such a large body of literature, I will focus on several types of applications of RNA-Seq. In general, RNA-Seq has been used in the following areas of research: 1) Transcriptional profiling of tissues, cell types, developmental stages, etc 2) Identification of differentially expressed genes with “treatment”. 3) Identification of allele specific expression 4) Identification of alternative splicing event 5) Identification of single nucleotide polymorphisms (SNPs) 6) Identification of long non-coding RNAs and 7) Identification of microRNAs (miRNAs).

RNA-Seq is such a sensitive techniques that made it possible to sequence only a small amount of material when the material is difficult to obtain, such as oocytes or the cells of the early embryo (Saliba et al., 2014; Tang et al., 2009). It is also a power tool to characterize the transcriptional complexity during various development stages. Zenoni et al, reported the transcriptional responses associated with berry development in *Vitis vinifera* ‘Corvina’ using RNA-Seq, resulting in 6,695 genes to be expressed in a stage-specific manner, suggesting differences in expression for genes in numerous functional categories and a significant transcriptional complexity (Zenoni et al., 2010). Another primary objective of RNA-Seq is to reveal differentially expressed genes under different conditions. These studies are always conducted by comparing between wild-type and mutant strains of the same tissues, or treated versus untreated tissues, cancer versus normal etc. (Oshlack et al., 2010). Slattery et al. recently compared the expression changes of 144 colon cancer patients who had record of recent cigarette smoking, recent alcohol consumption, diet, and recent aspirin/non-steroidal anti-inflammatory use, resulted that diet and lifestyle factors associated with oxidative stress can alter gene expression, while genes altered were unique to type of alcohol and type of antioxidant (Slattery et al., 2015). By comparison of hygienic and non-hygienic honeybee (*Apis mellifera L.*) hives, revealed a limited set of genes linked to different regulation patterns associated with an over-expression of cytochrome P450 genes (Boutin et al., 2015). RNA-Seq also enables allele-specific expression (ASE) studies. Identification of ASE from spleen transcriptome were proposed in response to *Streptococcus suis* 2 infection in two differentially susceptible pig breeds, revealed 882 and 1,096 statistically significant ASEs and got some of ASE validated using Sanger sequencing and quantified by pyrosequencing assay (Wu et al., 2015). In addition, RNA-Seq is also an efficient way to comprehensively identify alternative splicing and SNPs events from the expressed genes. Wen et al. performed genome-wide transcriptome analyses by pairwise comparison of gene expression in the breast tumor versus matched healthy tissue from each patient, uncovered 2,839 differential expressed genes and nine splicing factors that were involved in aberrant splicing in breast cancer (Wen et al., 2015). Yang et al. conducted transcriptome sequencing in Asian lotus to identify 357,689 putative SNPs and 177,540 alternative splicing events in the four cultivars and the alternative splicing were found to distribute in 64% of the expressed genes of lotus (Yang et al., 2015). RNA-Seq is also a powerful tool to discover those non-coding RNAs, such as long non-coding RNAs and micro RNAs, and examine their expression profiles. Schrauwen et al. conducted comprehensive transcriptome characterization in human inner ear, revealed a tissue-specific pattern of expression and uncovered spatial specificity of expression of set of RNAs including long-noncoding RNAs in the hearing/balance system (Schrauwen et al., 2015). RNA-Seq were also utilized to identify mouse miRNAs that were differentially regulated in adult and neonatal CD8+ T cells, and miR 29 and miR 130 were found out to be important regulators of memory CD8+ T cell formation (Wissink et al., 2015).

RNA-Seq has been also widely used in fish species (Qian et al., 2014). For instance, Cui et al. conducted transcriptome analysis on gill and swim bladder of *Takifugu rubripes*, revealed three immune-related pathways and 32 immune-related genes in gill and five pathways including 43 swim bladder-enriched genes in swim bladder (Cui et al., 2014). A genome-wide transcriptional analysis was performed in zebrafish embryos exposed to an environmentally relevant carcinogenic and endocrine disrupting compound Benzo[a]pyrene (BaP) in order to reveal a set of differential expressed genes and genes with differential exon usage, providing novel insights on the mechanisms of action of BaP-induced developmental toxicities (Fang et al., 2015). Allele-specific expression analysis was performed in a F1 interspecies hybridized from southern platyfish (*Xiphophorus maculates*) and monterrey platyfish (*Xiphophorus couchianus*), revealed 27 allele-specific expressed genes (Shen et al., 2012). Identification of gene-associated SNPs at a genome-wide scale was conducted using four strains of common carp (*Cyprinus carpio*), a total of 712,042 intra-strain SNPs and 53,893 inter-SNPs were identified, which provided a solid base for the future genetic studies and contributed to the development of a high throughput SNP genotyping platform (Xu et al., 2012). Alternative splicing event was also observed and reported in grass carp (*Ctenopharyngodon idella*) challenged with grass carp reovirus (GCRV). The splicing transcripts of IL-12p40 and IL-1R1 were firstly found to play diverse roles in the antiviral response of fishes and validated using PCR amplification and rapid-amplification of cDNA ends (RACE) technology, and the expression levels were confirmed by reverse transcription quantitative real-time PCR (RT-qPCR) (Wan and Su, 2015). RNA-Seq has also been utilized to screen the expression profiles in Atlanta salmon (*Salmo salar*), which discovered 244 unique mature microRNAs and 18 miRNAs were significantly differentially expressed when exposed to acid/Al water for 3 days (Kure et al., 2013).

RNA-Seq utilized in channel catfish

Channel catfish, *Ictalurus punctatus*, is the most important species of aquatic animal that commercially cultured in the United States. In recent years, the catfish industry has encountered unprecedented challenges which could cause big loss and threated the sustainability of the catfish industry. Recently, large amount of efforts has been devoted to study the molecular mechanism by RNA-Seq under different disease or stress conditions in order to provide insights into strategies for selection. Transcriptome analysis of pooled RNA samples from multiple individuals was carried out by our lab to generate genome-scale gene-associated SNPs in catfish (Liu et al., 2011). Over two million putative SNPs were identified from channel catfish, among them, 342,104 intra-specific SNPs were identified for channel catfish and 420,727 inter-specific SNPs were identified between channel catfish and blue catfish. The SNPs identified in this project provided resources for genetic studies and the development of a high-density SNP array. Later Liu et al. took the advantage of the doubled haploid channel catfish and comprehensively characterized a draft transcriptome. A significant non-redundant set of 370,798 transcripts including 14,240 full length transcripts were identified, the results contributed significantly towards assembly and annotation of the channel catfish genome (Liu et al., 2012). Later, channel catfish were challenged with its two major diseases *Flavobacterium columnare* and *Edwardsiella ictaluri* in another two studies. For the analysis of expression profile following *E. ictaluri*, comparison of gene expression between challenged and control samples revealed 1,633 differentially expressed genes at 3 h, 24 h, and 3 day after exposed to *E. ictaluri*. Gene pathway analysis of the differentially expressed gene set indicated the centrality of actin cytoskeletal polymerization/remodelling and junctional regulation in pathogen entry and subsequent inflammatory responses (Li et al., 2012). Transcriptomic profiling of host responses to *F. columnare* were later conducted following an experimental challenge at three time points (4 h, 24 h, and 48 h) in channel catfish gill after bath immersion infection. Enrichment and pathway analyses of the differentially expressed genes revealed upregulation of a RBL with putative roles in bacterial attachment and aggregation, and suppression of NF- κB signaling pathway (Sun et al., 2012). The two diseases studies revealed initial molecular mechanisms of pathogen entry during infection, and provided insights into strategies for selection of resistant catfish brood stocks against various diseases. Furthermore, another RNA-seq were conducted by Peatman et al. to profile gill expression differences between channel catfish differing in their susceptibility to *F. columnare* both basally (before infection) and at three early time points post-infection (1 h, 2 h, and 8 h). Following the analyses of differentially expressed genes, the results showed that the immune and mucin profiles obtained suggested a basal polarization in the gill mucosa, with susceptible fish possessing a putative mucosecretory, toleragenic phenotype which may predispose them to *F. columnare* infection (Peatman et al., 2013). RNA-Seq of the testis tissue was conducted with a goal of profiling the genes expressed in this male specific organ, and identifying male-biased transcripts, which could contribute to elucidate sex determination mechanisms in channel catfish (Sun et al., 2013). Short-term feed deprivation is a common occurrence in aquaculture fish species due to season, production strategies, or disease. Therefore, an RNA-Seq-based transcriptome profiling of skin and gill homogenates from fed and 7 d fasted channel catfish fingerlings were conducted to better understand immune-nutritional regulation in teleost fish. The results revealed potential mechanistic similarities between gut and surface mucosa and underscore the complex interrelationships between nutrition, mucosal integrity, and immunity in teleost fish (Liu et al., 2013).

Long non-coding RNAs

For decades, most studies at the RNA level have focused on protein-coding mRNAs. Only in recent few years, the attention has been given to non-coding RNAs with the discovery of pervasive transcription. Long non-coding RNAs (lncRNA) are defined as RNAs larger than 200 base-pairs that do not have coding potential, which distinguishes lncRNAs from small regulatory RNAs such as miRNAs or piRNAs. LncRNAs can be further classified based on their gene loci, which includes antisense lncRNAs (overlap known protein-coding genes), intronic lncRNAs (encode within introns of protein-coding genes), bidirectional lncRNAs (initiate in a divergent fashion from a promoter of a protein-coding gene), and intergenic lncRNAs (encode completely within intergenic genomic space between protein-coding loci) (Rinn and Chang, 2012). First identified long non-coding RNA was *Xist*, which was exclusively expressed from the inactive X chromosome (Borsani et al., 1991). It only expressed on the inactive chromosome and not on the active one, X chromosomes lacking *Xist* will not be inactivated, while duplication of the Xist gene on another chromosome causes inactivation of that chromosome (Brown et al., 1992). Rinn et al. used a high-resolution tiled microarray to systematically identify and characterize another lncRNA HOTAIR, which is the first known lincRNA that can acts in *trans* to regulate the chromatin state of genes on distantly located chromosomes (Rinn et al., 2007; Woo and Kingston, 2007). With the development of the sequencing technology (RNA-Seq), genome-wide identification of lncRNAs has only recently become possible, large numbers of lncRNAs has been discovered in many organisms (Kung et al., 2013). Liao et al. reported the utilization of computational annotation of lncRNA functions based on public microarray expression profiles. The functions annotated to the lncRNAs mainly involve organ or tissue development, cellular transport or metabolic processes (Liao et al., 2011). The interaction between protein-coding genes and lncRNAs has also been reported in breast cancer patients and determined the correlation with the construction of co-expression networks (Banerjee et al., 2013). The lncRNAs have also been reported to be spatially correlated with transcription factors across the human genome among the 363 identified lncRNAs in the lung and foregut endoderm, which could play an important role in foregut and lung endoderm development by regulating multiple aspects of gene transcription (Herriges et al., 2014). In addition, co-localization of protein-coding genes and lncRNAs has also been observed in mouse brain, where brain-expressed lncRNAs were preferentially located adjacent to protein-coding genes that are also expressed in the brain and involved in transcriptional regulation or in nervous system development (Ponjavic et al., 2009). However, even with the explosion of discovery of large numbers of lncRNAs, little is known about how lncRNAs function and the biological significance of the lncRNAs. In teleost fish, analysis of lncRNAs have largely been lacking except in zebrafish. More than 550 distinct lincRNAs were identified in zebrafish using chromatin marks, poly(A)-site mapping and RNA-Seq data, where most of the lncRNA only had conserved genomic locations without detectable sequence conservation (Ulitsky et al., 2011). In order to identify lncRNAs with potential functions in vertebrate embryogenesis, RNA-Seq were performed in eight stages during early zebrafish development, resulted in a stringent set of 1,133 noncoding multi-exonic transcripts expressed during embryogenesis (Pauli et al., 2012). In addition, RNA-Seq was performed to identify tissue restricted lncRNA transcript signatures from five different tissues of adult zebrafish, and 442 predicted lncRNA transcripts were identified, out of which 419 were novel lncRNA transcripts, tissue-specific expression pattern were also discovered across the five major tissues investigated in 77 lncRNAs (Kaushik et al., 2013). Recently, a comprehensive online resource "zflncRNApedia" was constructed for zebrafish lncRNAs. The catalog of lncRNAs were collected from the three annotation sets, as well as manual curation of literature to compile a total of 2,267 lncRNA transcripts (Dhiman et al., 2015). So far, no lncRNA studies have been conducted in any other aquaculture species including catfish.

Reference

Banerjee, N., Chothani, S., Harris, L., Dimitrova, N., 2013. Identifying RNAseq-based coding-noncoding co-expression interactions in breast cancer, Genomic Signal Processing and Statistics (GENSIPS), 2013 IEEE International Workshop on. IEEE, pp. 11-14.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., Snyder, M., 2004. Global identification of human transcribed sequences with genome tiling arrays. Science 306, 2242-2246.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816.

Borsani, G., Tonlorenzi, R., Simmler, M.C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., Willard, H.F., Avner, P., Ballabio, A., 1991. Characterization of a murine gene expressed from the inactive X chromosome. Nature 351, 325-329.

Boutin, S., Alburaki, M., Mercier, P.L., Giovenazzo, P., Derome, N., 2015. Differential gene expression between hygienic and non-hygienic honeybee (*Apis mellifera L.*) hives. BMC genomics 16, 500.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafreniere, R.G., Xing, Y., Lawrence, J., Willard, H.F., 1992. The Human Xist Gene - Analysis of a 17 Kb Inactive X-Specific Rna That Contains Conserved Repeats and Is Highly Localized within the Nucleus. Cell 71, 527-542.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., 2005. The transcriptional landscape of the mammalian genome. Science 309, 1559-1563.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D.K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D.S., Gingeras, T.R., 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308, 1149-1154.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., Morris, K.V., Morillon, A., Rozowsky, J.S., Gerstein, M.B., Wahlestedt, C., Hayashizaki, Y., Carninci, P., Gingeras, T.R., Mattick, J.S., 2011. The Reality of Pervasive Transcription. PLoS biology 9.

Consortium, I.H.G.S., 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931-945.

Cui, J., Liu, S., Zhang, B., Wang, H., Sun, H., Song, S., Qiu, X., Liu, Y., Wang, X., Jiang, Z., Liu, Z., 2014. Transciptome analysis of the gill and swimbladder of *Takifugu rubripes* by RNA-Seq. PloS one 9, e85505.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., Steinmetz, L.M., 2006. A high-resolution map of transcription in the yeast genome. Proceedings of the National Academy of Sciences of the United States of America 103, 5320-5325.

Dhiman, H., Kapoor, S., Sivadas, A., Sivasubbu, S., Scaria, V., 2015. zflncRNApedia: A Comprehensive Online Resource for Zebrafish Long Non-Coding RNAs. PloS one 10, e0129997.

Dinger, M.E., Amaral, P.P., Mercer, T.R., Mattick, J.S., 2009. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Briefings in functional genomics & proteomics 8, 407-423.

Fang, X., Corrales, J., Thornton, C., Clerk, T., Scheffler, B.E., Willett, K.L., 2015. Transcriptomic Changes in Zebrafish Embryos and Larvae Following Benzo[a]pyrene Exposure. Toxicological sciences : an official journal of the Society of Toxicology 146, 395-411.

Hangauer, M.J., Vaughn, I.W., McManus, M.T., 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS genetics 9, e1003569.

Herriges, M.J., Swarr, D.T., Morley, M.P., Rathi, K.S., Peng, T., Stewart, K.M., Morrisey, E.E., 2014. Long noncoding RNAs are spatially correlated with transcription factors and regulate lung development. Genes & development 28, 1363-1379.

Jensen, T.H., Jacquier, A., Libri, D., 2013. Dealing with pervasive transcription. Molecular cell 52, 473-484.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P.A., Gingeras, T.R., 2002. Large-scale transcriptional activity in chromosomes 21 and 22. Science 296, 916-919.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K.C., Hallinan, J., Mattick, J., Hume, D.A., Lipovich, L., Batalov, S., Engstrom, P.G., Mizuno, Y., Faghihi, M.A., Sandelin, A., Chalk, A.M., Mottagui-Tabar, S., Liang, Z., Lenhard, B., Wahlestedt, C., 2005. Antisense transcription in the mammalian transcriptome. Science 309, 1564-1566.

Kaushik, K., Leonard, V.E., Kv, S., Lalwani, M.K., Jalali, S., Patowary, A., Joshi, A., Scaria, V., Sivasubbu, S., 2013. Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish. PloS one 8, e83616.

Khatoon, Z., Figler, B., Zhang, H., Cheng, F., 2014. Introduction to RNA-Seq and its applications to drug discovery and development. Drug development research 75, 324-330.

Kung, J.T.Y., Colognori, D., Lee, J.T., 2013. Long Noncoding RNAs: Past, Present, and Future. Genetics 193, 651-669.

Kure, E.H., Saebo, M., Stangeland, A.M., Hamfjord, J., Hytterod, S., Heggenes, J., Lydersen, E., 2013. Molecular responses to toxicological stressors: Profiling microRNAs in wild Atlantic salmon (*Salmo salar*) exposed to acidic aluminum-rich water. Aquat Toxicol 138, 98-104.

Li, C., Zhang, Y., Wang, R., Lu, J., Nandi, S., Mohanty, S., Terhune, J., Liu, Z., Peatman, E., 2012. RNA-seq analysis of mucosal immune responses reveals signatures of intestinal barrier disruption and pathogen entry following *Edwardsiella ictaluri* infection in channel catfish, *Ictalurus punctatus*. Fish & shellfish immunology 32, 816-827.

Liao, Q., Liu, C.N., Yuan, X.Y., Kang, S.L., Miao, R.Y., Xiao, H., Zhao, G.G., Luo, H.T., Bu, D.C., Zhao, H.T., Skogerbo, G., Wu, Z.D., Zhao, Y., 2011. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic acids research 39, 3864-3878.

Liu, L.S., Li, C., Su, B.F., Beck, B.H., Peatman, E., 2013. Short-Term Feed Deprivation Alters Immune Status of Surface Mucosa in Channel Catfish (*Ictalurus punctatus*). PloS one 8, e74581.

Liu, S., Zhang, Y., Zhou, Z., Waldbieser, G., Sun, F., Lu, J., Zhang, J., Jiang, Y., Zhang, H., Wang, X., Rajendran, K.V., Khoo, L., Kucuktas, H., Peatman, E., Liu, Z., 2012. Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. BMC genomics 13, 595.

Liu, S., Zhou, Z., Lu, J., Sun, F., Wang, S., Liu, H., Jiang, Y., Kucuktas, H., Kaltenboeck, L., Peatman, E., Liu, Z., 2011. Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC genomics 12, 53.

Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A., Mattick, J.S., Rinn, J.L., 2012. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol 30, 99-104.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563-573.

Oshlack, A., Robinson, M.D., Young, M.D., 2010. From RNA-seq reads to differential expression results. Genome biology 11, 220.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., Schier, A.F., 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22, 577-591.

Peatman, E., Li, C., Peterson, B.C., Straus, D.L., Farmer, B.D., Beck, B.H., 2013. Basal polarization of the mucosal compartment in *Flavobacterium columnare* susceptible and resistant channel catfish (*Ictalurus punctatus*). Molecular immunology 56, 317-327.

Pertea, M., 2012. The human transcriptome: an unfinished story. Genes 3, 344-360.

Ponjavic, J., Oliver, P.L., Lunter, G., Ponting, C.P., 2009. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS genetics 5, e1000617.

Qian, X., Ba, Y., Zhuang, Q.F., Zhong, G.F., 2014. RNA-Seq Technology and Its Application in Fish Transcriptomics. Omics-a Journal of Integrative Biology 18, 98-110.

Rinn, J.L., Chang, H.Y., 2012. Genome regulation by long noncoding RNAs. Annual review of biochemistry 81, 145-166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., Chang, H.Y., 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by Noncoding RNAs. Cell 129, 1311-1323.

Saliba, A.E., Westermann, A.J., Gorski, S.A., Vogel, J., 2014. Single-cell RNA-seq: advances and future challenges. Nucleic acids research 42, 8845-8860.

Schrauwen, I., Hasin-Brumshtein, Y., Corneveaux, J.J., Ohmen, J., White, C., Allen, A.N., Lusis, A.J., Van Camp, G., Huentelman, M.J., Friedman, R.A., 2015. A comprehensive catalogue of the coding and non-coding transcripts of the human inner ear. Hearing research.

Shen, Y.J., Catchen, J., Garcia, T., Amores, A., Beldorth, I., Wagner, J., Zhang, Z.P., Postlethwait, J., Warren, W., Schartl, M., Walter, R.B., 2012. Identification of transcriptome SNPs between *Xiphophorus* lines and species for assessing allele specific gene expression within F-1 interspecies hybrids. Comp Biochem Phys C 155, 102-108.

Slattery, M.L., Pellatt, D.F., Mullany, L.E., Wolff, R.K., 2015. Differential Gene Expression in Colon Tissue Associated With Diet, Lifestyle, and Related Oxidative Stress. PloS one 10, e0134406.

Sun, F., Peatman, E., Li, C., Liu, S., Jiang, Y., Zhou, Z., Liu, Z., 2012. Transcriptomic signatures of attachment, NF-kappaB suppression and IFN stimulation in the catfish gill following columnaris bacterial infection. Developmental and comparative immunology 38, 169-180.

Sun, F.Y., Liu, S.K., Gao, X.Y., Jiang, Y.L., Perera, D., Wang, X.L., Li, C., Sun, L.Y., Zhang, J.R., Kaltenboeck, L., Dunham, R., Liu, Z.J., 2013. Male-Biased Genes in Catfish as Revealed by RNA-Seq Analysis of the Testis Transcriptome. PloS one 8.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., Surani, M.A., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nature methods 6, 377-382.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., Bartel, D.P., 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147, 1537-1550.

van Bakel, H., Nislow, C., Blencowe, B.J., Hughes, T.R., 2010. Most "Dark Matter" Transcripts Are Associated With Known Genes. PLoS biology 8, e1000371.

van Bakel, H., Nislow, C., Blencowe, B.J., Hughes, T.R., 2011. Response to "The Reality of Pervasive Transcription". PLoS biology 9, 1423.

Vikman, P., Fadista, J., Oskolkov, N., 2014. RNA sequencing: current and prospective uses in metabolic research. J Mol Endocrinol 53, R93-R101.

Wan, Q.Y., Su, J.G., 2015. Transcriptome analysis provides insights into the regulatory function of alternative splicing in antiviral immunity in grass carp (*Ctenopharyngodon idella*). Scientific reports 5.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics 10, 57-63.

Wen, J., Toomer, K.H., Chen, Z., Cai, X., 2015. Genome-wide analysis of alternative transcripts in human breast cancer. Breast cancer research and treatment 151, 295-307.

Wilhelm, B.T., Landry, J.R., 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48, 249-257.

Wissink, E.M., Smith, N.L., Spektor, R., Rudd, B.D., Grimson, A., 2015. MicroRNAs and Their Targets Are Differentially Regulated in Adult and Neonatal Mouse CD8+ T Cells. Genetics, genetics. 115.179176.

Woo, C.J., Kingston, R.E., 2007. HOTAIR lifts noncoding RNAs to new levels. Cell 129, 1257-1259.

Wu, H., Gaur, U., Mekchay, S., Peng, X., Li, L., Sun, H., Song, Z., Dong, B., Li, M., Wimmers, K., Ponsuksili, S., Li, K., Mei, S., Liu, G., 2015. Genome-wide identification of allele-specific expression in response to *Streptococcus suis* 2 infection in two differentially susceptible pig breeds. Journal of applied genetics, 1-11.

Xu, J., Ji, P.F., Zhao, Z.X., Zhang, Y., Feng, J.X., Wang, J., Li, J.T., Zhang, X.F., Zhao, L., Liu, G.Z., Xu, P., Sun, X.W., 2012. Genome-Wide SNP Discovery from Transcriptome of Four Common Carp Strains. PloS one 7, e48140.

Yang, M., Xu, L., Liu, Y., Yang, P., 2015. RNA-Seq Uncovers SNPs and Alternative Splicing Events in Asian Lotus (*Nelumbo nucifera*). PloS one 10, e0125702.

Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., Bellin, D., Pezzotti, M., Delledonne, M., 2010. Characterization of Transcriptional Complexity during Berry Development in *Vitis vinifera* Using RNA-Seq. Plant physiology 152, 1787-1795.

Chapter III

ASSEMBLY AND ANNOTATION OF THE CHANNEL CATFISH TRANSCRIPTOME AND ASSESSMENT OF PERVASIVE EXPRESSION

Materials and methods

Construction of comprehensive set of channel catfish transcriptome

There are two main approaches for assembly of a transcriptome: genome-guided approach when a reference genome is available, and *de novo* assembly when the reference genome is absent. *De novo* transcriptome assembly is more challenging especially in higher eukaryotes due to the large number of genes, great variations in the expression levels, and the large numbers of alternatively spliced transcript variants. However, *de novo* assembly could reconstruct transcripts from regions missing in the genome assembly. Collectively, both *de novo* and genome-guided approaches have their own advantages, but neither set of assembly tools could achieve the optimal desired assembly with sensitivity, specificity, and the ability to assemble full-length transcripts on their own (Jain et al., 2013). Therefore, two steps of assembly were performed to constructed a comprehensive set of channel catfish transcriptome, including Trinity *de novo* assembly (Grabherr et al., 2011; Haas et al., 2013), and genome-guided TopHat-Cufflinks assembly (Kim et al., 2013; Robertson et al., 2010; Trapnell et al., 2009; Trapnell et al., 2010).

De novo assembly of channel catfish transcriptome

De novo assembly was conducted using the 100 bp short reads from all available channel catfish RNA-Seq including all published datasets downloaded from NCBI and all ongoing RNA-Seq projects that's been working on in our lab. Raw sequencing reads were pooled from a collection of 13 channel catfish RNA-seq libraries, containing approximately 4.8 billion reads derived from various tissues including brain, gill, head kidney, trunk kidney, intestine, liver, muscle, skin, spleen, stomach, heart, fin, pancreas, brain, adipose, gall bladder, ovary, testis, thymus, eye, swim bladder, and barbels.

Raw reads were trimmed by removing adaptor sequences, ambiguous nucleotides, and low quality sequences (quality scores < 30 or read length < 30 bp) using Trimmomatic software version 0.32 (Bolger et al., 2014) with the option of ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 for removing the adaptor sequences, LEADING:3 for removing leading low quality or N bases (below quality 3), TRAILING:3 for remove trailing low quality or N bases (below quality 3), SLIDINGWINDOW:4:30 for scanning the read with a 4-base wide sliding window and cutting when the average quality per base drops below 30, MINLEN:30 for dropping reads below the 30 bases long. The remaining high-quality sequences were used in the subsequent assembly. All trimmed reads were then mapped to all genomes of potential contamination sources downloaded from NCBI using deconseq software (Schmieder and Edwards, 2011) to get most clean channel catfish reads for transcriptome assembly. The potential contamination sources included bacteria database, protozoa database, fungi database, virus database, *Caenorhabditis elegans* genome, *Drosophila melanogaster* genome and *Arabidopsis thaliana* genome. Reads were also mapped to several teleost genome sequences as retain catfish data for assembly. Only those reads that were only mapped to the contamination databases but not to the teleost databases are excluded from assembly. The used teleost databases included half smooth tongue sole (*Cynoglossus semilaevis* Cse_v1.0), fugu (*Takifugu rubripes* FUGU5), tilapia (*Oreochromis niloticus* Orenil1.1), medaka (*Oryzias latipes* ASM31367v1), carp (*Cyprinus carpio*), zebrafish (*Danio rerio* GRCz10), European seabass (*Dicentrarchus labraxseabass* V1.0), platyfish (*Xiphophorus maculatus* 4.4.2), Atlantic salmon (*Salmo salar* ICSASG_v1), green spotted puffer (*Tetraodon nigroviridis* ASM18073v1), stickleback (*Gasterosteus aculeatus* ASM18067v1), Amazon molly (*Poecilia Formosa* 5.1.2), cave fish (*Astyanax mexicanus* 1.0.2), zebra mbuna (*Maylandia zebra* MetZeb1.1), coelacanth (*Latimeria chalumnae* latCha1), Atlantic cod (GadMor May2010), spotted gar (*Lepisosteus oculatus* LepOcu1).

Before assembly of these large sets of RNA-Seq datasets, *in silico* read normalization was utilized to lower the memory and compute requirements for assembly. The *in silico* read normalization was carried out using the parameters of –JM 200G for 200GB of system memory to use for k-mer counting by jellyfish, --max_cov 50 for targeted maximum coverage of 50X for reads, --pairs_together for process paired reads by averaging stats between pairs and retaining linking info, --PARALLEL_STATS in order to generate read stats in parallel for paired reads.

The assembly of the whole transcriptome was carried out using Trinity software of version r20140717 (Grabherr et al., 2011; Haas et al., 2013), which is a novel method for the efficient and robust *de novo* reconstruction of transcriptomes from RNA-seq data. It combines three independent software modules: Inchworm, Chrysalis, and Butterfly, applied sequentially to process large volumes of RNA-seq reads. Trinity partitions the sequence data into many individual *de Bruijn* graphs, each representing the transcriptional complexity at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes. The assembly was carried out using the parameters of JM 150G for 150GB of system memory to use for k-mer counting by jellyfish, PasaFly for utilization of PASA assembly algorithm in the context of butterfly transcript graphs in order to produce fewer isoforms than the most conservative parameterization of the default method, group_pairs_distance 450 was set according to the larger insert library since multiple paired-end libraries were used, pairings that exceed that distance will be treated as if they were unpaired by the Butterfly process. The CD-HIT-EST (Li and Godzik, 2006) and CAP3 (Huang and Madan, 1999) were then used to remove assembly redundancy and retain the longest possible contigs by setting global sequence identity in CD-HIT-EST to 1, the minimal overlap length and percent identity in CAP3 to 100 bp and 99%.The remaining contigs composes the final assembly of non-redundant *de novo* contigs. The assembled contigs were then mapped to channel catfish genome in order to assess the quality of the assembly. The mapping was carried out using BLAT software (Kent, 2002), with the parameter of –q=rna for query type of rna, -minIdentity=90, for setting the minimum sequence identity to 90%, the coverage was set to 50% and the top hit of each alignment were picked.

Functional annotation for the *de novo* assembled transcriptome

All the non-redundant contigs from final assembly were used as queries to search against NCBI non-redundant (NR) protein database and Uniprot/Swiss-Prot database using BLASTX program. The cutoff E-value was set at 1e-5 and only the top gene ids and descriptions were initially assigned to each contig. Duplicated gene ids and descriptions were removed to get unique protein hits. Since channel catfish genome is considered almost complete, therefore contigs that had a BLAST hit but cannot map to channel catfish genome were potentially coming from contaminants. These congits were further BLASTX to NR database again with top five gene ids and descriptions, only those contigs that had all top five descriptions from invertebrate were excluded from the final annotation.

Genome-guided TopHat-Cufflinks RABT assembly of channel catfish transcriptome

The trimmed reads from each study generated previously were mappeed to channel catfish genome by TopHat version 2.0.14 (Kim et al., 2013; Trapnell et al., 2009) which allows spliced alignment and utilized an 'exon-first' approach where reads were first mapped to the genome then the unmapped reads were split into shorter segments and aligned independently. Mapping was carried out with genome annotation general transfer format (GTF) table by default parameter, which allowed less than two mismatches and less than two gaps in the final read alignments. Cufflinks assembly version 2.1.1 (Trapnell et al., 2010) was then utilized for reference annotation based transcript (RABT) assembly of each study (Roberts et al., 2011), the RABT assembler builds upon a known reference annotation to better identify novel transcripts, reference transcripts is tiled with faux-reads to provide additional information in assembly. After assembly of separate studies, cuffmerge was utilized to merge together all cufflinks assemblies into a master transcriptome. After merging of the cufflinks assemblies, cuffcompare was performed to compare the assembled transcripts to channel catfish reference annotation, in order to assess the quality of the assembly.

Identification of putative full length transcripts

All contigs from both *de novo* Trinity assembly and genome-guided TopHat-Cufflinks assembly that have NR or Uniprot BLAST hit were selected to predict for putative full length transcripts. The complete coding sequences were predicted with the software Transdecoder (Haas et al., 2013) by finding the start and stop codon and aided by BLASTP against Uniprot and zebrafish RefSeq protein database. Only those ORFs that were at least 100 amino acids long were retained to avoid false positive ORF predictions. Full length contigs were only selected if 1) the contig contains a complete CDS, and 2) the ORF length ratio (catfish predicted protein length/reference protein length) was over 0.8. All full length contigs got from above from both *de novo* Trinity assembly and genome-guided TopHat-Cufflinks assembly were further removed redundancy based on annotation to get final unique full length transcripts.

Assessment of pervasive transcription in channel catfish

All reads from hybrid catfish RNA-Seq datasets were trimmed using the same method described for trimming channel catfish raw reads, along with all channel catfish trimmed reads from each study generated previously were used to map to channel catfish genome by Spliced Transcripts Alignment to a Reference (STAR) software (Dobin et al., 2013) for its high mapping accuracy and ultrafast speed (Engstrom et al., 2013), the mapping allowed 5% mismatch of the mapped length and restricted the minimum 90% of the bases matched to the genome. After the resulting mapping file was sorted using samtools sort function (Li et al., 2009), it is applied to samtools depth function (Li et al., 2009), which could calculate the sequence depth for each nucleotide in the genome based on the mapping file. If one position had its resulted depth of more than 0, then it had been covered at least once in the transcriptome. Then the number of all covered nucleotides divided by the channel catfish genome size could calculate the percentage of how much genome had been transcribed. The assessment was carried out using the accumulated mapping file using each RNA-Seq datasets in a sequential but accumulative fashion.

Genome-wide expression profiles of channel catfish

Taken advantage of the large amount of RNAs that’s been sequenced, several RNA-Seq datasets under normal conditions were selected to perform the genome-wide expression profiles, including the 11 pooled tissues (brain, gill, head kidney, intestine, liver, muscle, skin, spleen, stomach, heart, and trunk kidney) from 47 adults, 19 pooled tissues (head kidney, fin, pancreas, spleen, gill, brain, trunk kidney, adipose, liver, stomach, gall bladder, ovary, intestine, thymus, skin, eye, swim bladder, muscle, heart) from one single doubled haploid female channel catfish adult, gill tissues of control group of the columnaris disease challenge experiments (0h, 4h, 24h, and 48h), intestine tissue of the control group of the ESC disease challenge experiments, single testis tissue, gill and skin tissues of the control group of the short-term feed deprivation experiments, two controls groups from two sets of channel catfish gill of different susceptibilities challenged with columnaris disease (0h, 1h, 2h, and 8h), single skin tissue, single barbel tissue, 10 to 29 days channel catfish of whole body without head, 90 to 110 days channel catfish gonads (both testis and ovary), whole fish of 1 to 14 days, control group of liver tissue from ESC challenged backcross progenies, and control group of gill tissue from low oxygen challenged backcross progenies. All trimmed reads from each time point and each treatment of channel catfish and hybrid catfish RNA-Seq datasets were used to map to channel catfish genome by STAR alignment software (Dobin et al., 2013), allowed 5% mismatch of the mapped length and restricted the minimum 90% of the bases matched to the genome. Different time points and treatments of each study was calculated separately in order to use a weighted trimmed mean of the log expression ratios (trimmed mean of M values (TMM)) normalization method to normalize different RNA libraries so that data generated from different libraries was comparable and could be added up. Genome-wide expression profiles were computed by counting the library size normalized reads coverage across 50 kilobase bin tiling the channel catfish genome using BEDTools version 2.24.0 coverage function (Quinlan and Hall, 2010).

Protein-coding gene expression profiles of channel catfish

All trimmed reads from each time point and each treatment of channel catfish and hybrid catfish RNA-Seq datasets were used to map to channel catfish genome by STAR alignment software (Dobin et al., 2013), allowed 5% mismatch of the mapped length and restricted the minimum 90% of the bases matched to the genome. Raw read counts for each protein-coding genes that annotated from genome were extracted from alignment files using htseq software version 0.6.1 htseq-count function (Anders et al., 2015) with the recommended union mode. Only the uniquely mapped reads could be used in htseq-count software in order to avoid false positives expression level. Protein-coding gene expression profiles were assessed by TMM normalized RPKM (reads per kilobase per million) gene expression value calculated using EdgeR software (McCarthy et al., 2012; Robinson et al., 2010).

Identification of tissue-specific expressed genes in channel catfish

All trimmed reads from each time point and each treatment of channel catfish and hybrid catfish from 10 RNA-Seq datasets were used to map to channel catfish genome by STAR alignment software (Dobin et al., 2013), allowed 5% mismatch of the mapped length and restricted the minimum 90% of the bases matched to the genome. 11 RNA-Seq datasets contained eight different tissues, including barbel, gill (pooled RNA-Seq datasets of control groups from columnaris disease challenged gill tissue (0h, 4h, 24h, and 48h), two sets of channel catfish gill of different susceptibilities challenged with columnaris disease (0h, 1h, 2h, and 8h), gill tissue from low oxygen challenged backcross progenies, and gill tissue from F1 hybrids challenged with heat stress), intestine (control group of ESC disease challenged intestine), liver (pool RNA-Seq datasets of control groups from ESC challenged liver tissue of backcross progenies and liver tissue from

F1 hybrids challenged with heat stress), ovary (90 to 110 days), skin, and testis (10 to 29 days and 90 to 110 days). Raw read counts for each protein-coding genes that annotated from the genome were extracted from alignment files using htseq software version 0.6.1 htseq-count function (Anders et al., 2015) with the recommended union mode. Different time points, treatments of each study were normalized using TMM normalization method. The expression level of a gene in a particular tissue was compared to its expression level in all remaining seven tissues. For distinction of tissue-specific genes, the fold change in expression level was set to 32 fold with the FDR (False discovery rate) adjusted p value of less than 0.05, which means that genes with an expression level in one tissue that was 32 fold higher than the maximum value in any of the other seven tissues.

Identification of induced expressed protein-coding genes in channel catfish

All trimmed reads from each time point and each treatment of channel catfish and hybrid catfish from five disease or stress challenge RNA-Seq datasets were used to map to channel catfish genome by STAR alignment software (Dobin et al., 2013), allowed 5% mismatch of the mapped length and restricted the minimum 90% of the bases matched to the genome. The five RNA-Seq datasets included ESC disease challenge, two sets of columnaris disease challenge, heat stress challenge and feed deprivation challenge. Raw read counts for each protein-coding genes that annotated from genome were extracted from alignment files using htseq software version 0.6.1 htseq-count function (Anders et al., 2015) with the recommended union mode. Different time point, treatment of each study was normalized using TMM normalization method to calculate a normalized RPKM using edgeR software (Robinson et al., 2010). The fold change between different treatments and controls were determined based on the normalized RPKM of each sample.

Differentially induced genes were defined as at least two-fold change in expression and FDR (false discovery rate) corrected p-value < 0.05.

Identification of long non-coding RNAs (lncRNAs) in channel catfish

Cufflinks assembled contigs were scored with CPAT 1.2.1 (Wang et al., 2013a) to determine their coding potential, only those contigs with coding probability less than 0.38 were assigned to the pre-trimmed non-coding contigs. The six frames of ORF region of these contigs were predicted using EMBOSS getorf module (Rice et al., 2000). The mRNA and putative amino acid sequences were used as queries against channel catfish genome annotation, NCBI non-redundant (nr) protein database, the UniProtKB/SwissProt database and Pfam database (including Pfam-A and Pfam-B database) using BLASTX and HMMER respective (BLASTX for nr and Uniport database and HMMER (Finn et al., 2011) for Pfam database) with cut-off Expect value of 1e-4. Any contig with an E-value greater than 1e-4 in the contig set was removed. The ORF region of remain contigs were further predicted by NCBI ORFinder, a maximal ORF cutoff less than 100 aa were imposed to get final lncRNA datasets. Contigs were also excluded that have any overlap with the UTR regions of protein-coding genes retrieved from the UTRdb database (Grillo et al., 2010). Finally, the contigs that were overlap with channel catfish annotated genes were excluded, and the longest contig was remained if contigs were overlap with each other. After all the described filtering steps, the kept contigs were considered as the candidates for lncRNA.

LncRNA expression profiles of channel catfish

All trimmed reads from each time point and each treatment of channel catfish and hybrid catfish RNA-Seq datasets were used to map to channel catfish genome by STAR alignment software (Dobin et al., 2013), allowed 5% mismatch of the mapped length and restricted the minimum 90% of the bases matched to the genome. Raw read counts for each lncRNA were extracted from alignment files using htseq software version 0.6.1 htseq-count function (Anders et al., 2015) with the recommended union mode. Only the uniquely mapped reads could be used in htseq-count software in order to avoid false positives expression level. LncRNAs expression profiles were assessed by TMM normalized RPKM (reads per kilobase per million) gene expression value calculated using EdgeR software (McCarthy et al., 2012; Robinson et al., 2010).

Identification of tissue-specific expressed lncRNAs in channel catfish

All trimmed reads from each time point and each treatment of channel catfish and hybrid catfish from 10 RNA-Seq datasets were used to map to channel catfish genome by STAR alignment software (Dobin et al., 2013), allowed 5% mismatch of the mapped length and restricted the minimum 90% of the bases matched to the genome. 10 RNA-Seq datasets contained eight different tissues, including barbels, gill, intestine, liver, ovary, skin, and testis. Raw read counts for each lncRNs were extracted from alignment files using htseq software version 0.6.1 htseq-count function (Anders et al., 2015) with the recommended union mode. Different time point, treatment of each study was normalized using TMM normalization method. The expression level of a lncRNA in a particular tissue was compared to its expression level in all remaining seven tissues. For distinction of tissue-specific genes, the fold change in expression level was set to 32 fold with the FDR adjusted p value of less than 0.05.

Identification of induced expressed lncRNAs in channel catfish

All trimmed reads from each time point and each treatment of channel catfish and hybrid catfish from five disease or stress challenge RNA-Seq datasets were used to map to channel catfish genome by STAR alignment software (Dobin et al., 2013), allowed 5% mismatch of the mapped length and restricted the minimum 90% of the bases matched to the genome. The five RNA-Seq datasets used were the same as described in identification of induced expressed protein-coding genes. Raw read counts for each lncRNAs were extracted from alignment files using htseq software version 0.6.1 htseq-count function (Anders et al., 2015) with the recommended union mode. Different time point, treatment of each study was normalized using TMM normalization method to calculate a normalized RPKM using edgeR software (Robinson et al., 2010). The fold change between different treatments and controls were determined based on the normalized RPKM of each sample. Differentially induced lncRNAs were defined as at least two-fold change in expression and FDR (false discovery rate) corrected p-value < 0.05.

Identification of correlated co-expressed lncRNAs and genes

Correlation analysis was performed using normalized RPKM of all the significantly differential expressed genes and lncRNAs in each time point, tissue and treatment group. Correlation matrix was constructed between differential expressed genes and lncRNAs using R version 3.0.3. The significant correlated co-expressed lncRNAs and genes were defined as correlation coefficient greater than 0.9 or smaller than -0.9 and two-tail significant p-value smaller than 0.05.

Identification of induced co-localized lncRNAs and genes

The significantly differential expressed genes and lncRNAs were compared based on each treatment group, time point and tissue. Only both the lncRNAs and their closest neighbouring gene were significantly differentially expressed were considered as the co-localization expression sets.

Result

Datasets for the assembly of the catfish transcriptome

With the advantage of next-generation sequencing techniques, transcriptome sequencing has become a powerful tool for obtaining large amount of functional genomic data. However, separate RNA-experiment cannot give a complete and comprehensive transcriptome due to different specific experiment aims and limited sequencing depth. Therefore *de novo* assembly of pooled RNA-seq datasets is necessary for analyses of channel catfish transcriptome. In order to construct the comprehensive transcriptome, a total of 13 channel catfish RNA-Seq datasets containing approximately 4.8 billion reads, representing all the tissues including brain, gill, head kidney, trunk kidney, intestine, liver, muscle, skin, spleen, stomach, heart, fin, pancreas, brain, adipose, gall bladder, ovary, testis, thymus, eye, swim bladder, and barbel, various developmental stages, and various environmental stress conditions were collected in Table 1.

Table 1. Summary of channel catfish RNA-Seq datasets used in transcriptome assembly

	tissue	accession	Sequencing	raw reads (million)	trimmed reads (million)	Publication
1	11 pooled tissues (47 adult)	SRA025099	pair end	222.5	169.7	Liu et al., 2011
2	19 pooled tissues (1 adult)	SRA047025	pair end	315.7	252.1	Liu et al.,2012
3	Gill (1 year)	SRP012586	pair end	203.2	182.3	Sun et al.,2012
4	Intestine (1 year)	SRP009069	pair end	197.6	177.3	Li et al.,2012
5	Testis (adult)	SRP018265	pair end	294.6	227.1	Sun et al.,2013
6	gill, skin	SRP017689	pair end	209	150.9	Liu et al.,2013
7	gill	SRP017689	pair end	350	294	Peatman et al.,2013
8	Skin (6 month-1 year)		pair end	193.6	158.9	Gao et al.

9	Barbel (2 year)		pair end	431.9	363.2	Zhou et al.
10	whole fish without head (10-29 days)		pair end	544.9	386.4	Zhang et al.
11	Gonad (male female) (90-110 days)		pair end	410.7	273.4	Zeng et al.,
12	whole fish (1-14 days)		pair end	261.9	140.3	Li et al.
13	liver	SRP041359	single end	1098	561.9	Mark Arick II et al.

Trinity *de novo* assembly of the channel catfish transcriptome

RNA-seq was all conducted using Illumina sequencing. Raw sequencing reads were pooled from the above 13 channel catfish RNA-seq libraries containing approximately 3.6 billion 100-bp short paired end reads and 1 billion 50-bp single end reads. All the raw reads were trimmed to obtain clean, high quality sequences. After removing of the exogenous reads, a total of 3.3 billion trimmed reads, about 70% of the raw reads were retained for the assembly. These reads were carried forward for *de novo* assembly.

As shown in Table 2, assembly using Trinity resulted in 780,637 initial contigs (including coding and non-coding RNA) with a N50 contig length of 1,289 bp, an average contig length of 784 bp, and median contig length of 422 bp. Of the assembled contigs, over 157,015 had a length of over 1,000 bp. The CD-HIT-EST and CAP3 were then used to remove redundancy and retain the longest possible contigs. The remaining non-redundant 769,270 contigs composed the final assembly of non-redundant contigs. The assembled contigs were then mapped to channel catfish genome in order to assess the quality of the assembly. The mapping was carried out using identity cutoff of 90% and coverage of 50%. Vast majority of the contigs (742,956 contigs or 96.56% of all of the *de novo* assembled contigs) were mapped to the genome.

Table 2. Summary of channel catfish transcriptome *de novo* assembly results

Contigs	780,637
Large contigs (≥1000 bp)	157,015
Maximum length (bp)	59,902
Average length (bp)	784.34
Median length (bp)	422
N50 size (bp)	1,289
Non-redundant contigs (After CD-HIT-EST + CAP3)	769,270
Average non-redundant contigs length (bp) (After CD-HIT-EST + CAP3)	760.75
Contigs mapped to genome (identity 90%, coverage 50%)	742,956
Contigs mapped to genome (%)	96.56

Annotation for the *de novo* assembled transcriptome

All 769,270 non-redundant contigs from final *de novo* assembly were used as queries to search against NCBI non-redundant (NR) protein database and Uniprot/Swiss-Prot database using BLASTX program. The cutoff E-value was set at 1e-5 and only the top gene id and name were initially assigned to each contig. Duplicated gene ids and descriptions were removed to get unique protein hits. Collectively, of the 769,270 non-redundant contigs, 169,180 had hits to at least one database, and resulted in 25,888 unique protein hits. Among these unique protein hits, 5,718 were unknown, unnamed, hypothetical, or uncharacterized genes. Mapping of genes to channel catfish genome was carried out using the same parameter above, those contigs that cannot map to channel catfish genome were BLASTX to NR database again with top five results for further excluding the genes that were contaminations. After carefully eliminating the contaminates, 800 genes remained as novel genes that cannot map to channel catfish genome but existed in the channel catfish transcriptome, accounting for 3.09% of all genes. The final annotation for the *de novo* assembly was summarized below in Table 3.

Table 3. Summary of annotation for the channel catfish *de novo* assembled transcriptome.

Number of contigs with hits	169,180
Number of unique protein hits	25,888
Number of contigs with hits to unknown hypothetical gene matches	5,718
Number of unique protein mapped to genome (identity 90%, coverage 50%)	25,088
% of unique protein mapped to genome (identity 90%, coverage 50%)	96.91
% of unique protein not mapped to genome (identity 90%, coverage 50%)	3.09

Construction of comprehensive set of channel catfish transcriptome using annotated genome-guided TopHat-Cufflinks assembly

Even though *de novo* transcriptome assembly had lots of advantages especially for construction of novel transcripts, however, it cannot give a complete transcriptome since *de novo* is not sensitive enough to reconstruct all low abundant region and less ability to recover full length transcripts. Therefore, annotated genome-guided TopHat-Cufflinks assembly was also conducted to reconstruct a relatively complete transcriptome. The genome-guided TopHat-Cufflinks assembly was carried out using the same trimmed reads of same RNA-Seq datasets. The high-quality reads were able to assemble 197,161 contigs which was much less than the *de novo* assembled contigs. However, the N50 contig lenth was 5,790, which was much larger than the *de novo* N50 length, indicating much more full-length transcripts were reconstructed. The average contig length was 3,182 bp, and median contig length of 1,971 bp. Of the assembled contigs, most of the contigs (133,740 contigs) had a length of over 1,000 bp. In order to assess the quality of the assembly, genome-guided TopHat-Cufflinks assembly was compared against the channel catfish annotation using cuffcompare. Of the 26,661 channel catfish annotated genes, 25,987 genes were reconstructed, accounting for 97.47% of all the annotated genes. The summary of genome-guided assembly results were summarized in Table 4. The assembled genes combined from *de novo* and genome-guided assemblies were 27,448, where 13 genes were not assembled from the genome annotation.

Table 4. Summary of channel catfish transcriptome genome-guided TopHat-Cufflinks assembly results

Contigs	197,161
Large contigs (≥1000 bp)	133,740
Maximum length (bp)	62,178
Average length (bp)	3,182
N50 size (bp)	5,790
Median length	1,971
Number genome annotated genes transcribed	25,987
% of genome derived protein-coding gene transcribed	97.47

Identification of the full length transcripts from the transcriptome

RNA-Seq had shown its ability to be a cost-effective approach for identification and characterization of full-length transcripts without the help from laborious cloning (Liu et al., 2012). In the present study, we collected a large set of 13 channel catfish RNA-Seq datasets, containing various tissues, various developmental stages, and various environmental stress conditions for assembling a relatively complete transcriptome. We took advantage of Trinity *de novo* assembly in conjunction with genome-guided TopHat-Cufflinks assembly to identify full-length transcripts. To identify full-length transcripts, all contigs from both *de novo* Trinity assembly and genome-guided TopHat-Cufflinks assembly that had functional annotation of either NR or Uniprot were selected to predict for putative full length transcripts. The ORFs were predicted with the software Transdecoder and the ORF that had functional significance were selected by scanning all potential ORFs for homology to known proteins from zebrafish RefSeq and Uniprot database. As shown in Table 5, a total of 102,279 genome-guided TopHat-Cufflinks assembled contigs and 72,418 Trinity *de novo* assembled contigs were identified to have complete ORF, while a total of 24,946 genome-guided TopHat-Cufflinks assembled transcripts and 14,070 Trinity *de novo* assembled transcripts were identified to have complete ORF. Full length contigs were then identified only if the contigs contain a complete CDS and the ORF length ratio of catfish predicted protein length/reference protein length was over 0.8. The resulted full length transcripts were 19,374 from genome-guided TopHat-Cufflinks assembly and 10,733 from Trinity *de novo* assembly, the overall full length transcripts combined from both assemblies were 20,371, which had 20,244 transcripts mapped to channel catfish genome (Supplemental Table 1). The remaining 127 were full length transcripts that were not included in the catfish genome assembly.

Table 5. Summary of channel catfish full length transcripts identified from the transcriptome

	Cufflinks genome-guided	Trinity *de novo*
Number of contigs with complete ORF	102,279	72,418
Number of transcripts with complete ORF	24,946	14,070
Number of full length transcripts (ORF length ratio >=0.8)	19,374	10,733
Number of combined unique full length transcripts	20,371	
Number of combined unique full length transcripts mapped to the genome	20,244	
Number of combined unique full length transcripts not mapped to the genome	127	

Assessment of pervasive transcription in channel catfish

In the present study, we took advantage of the large numbers of RNA-Seq studies ever conducted using catfish (both channel catfish and hybrid catfish (female channel catfish crossed with male blue catfish, *Ictalurus furcatus*)) to assess the pervasiveness of genome transcription. In addition to the above channel catfish RNA-Seq datasets, three more RNA-Seq datasets were used, including F2 generation backcross progenies (F1 hybrid backcrossed with the susceptible channel catfish) of the interspecific hybrids challenged with ESC infection (Wang et al., 2013b), F1 hybrid catfish fingerlings (female channel catfish crossed with male blue catfish) challenged with heat stress (Liu et al., 2013b) and another sets of F2 generation backcross progenies challenged with low oxygen stress (data unpublished). As shown in Figure 1, 16 RNA-Seq datasets were used for pervasive transcription assessment. Among these datasets, the 13th dataset (two different channel catfish strains challenged with heat stress) had the largest size of about 1 billion reads, while the smallest size were the 4th dataset which only contained about 200 million reads, all 16 RNA-Seq datasets accumulated to a total of six billion reads. The first dataset was conducted using a pool of 11 tissues, which accounted for 35% of the channel catfish genome length. When the second dataset was used from a double haploid channel catfish with a pool of 19 tissues, along with all the reads from the first 11 tissues, the transcriptome accounted for 46% of the whole channel catfish genome length. After the first two datasets, several challenged datasets and specific tissues were added to the assessment, including columnaris disease challenged gill tissue (0h, 4h, 24h, and 48h) (3rd dataset), ESC disease challenged intestine tissue (4th dataset), testis tissue (5th dataset), short-term feed deprivation challenged gill and skin tissues (6th dataset), two sets of channel catfish gill of different susceptibilities challenged with columnaris disease (0h, 1h, 2h, and 8h) (7th dataset), skin tissue (8th dataset), barbel tissue (9th dataset), 10 to 29 days channel catfish of whole body without head (10th dataset), 90 to 110 days channel catfish gonads (both testis and ovary) (11th dataset), whole fish of 1 to 14 days (12th dataset), liver tissue from two different channel catfish strains challenged with heat stress (13th dataset), and three previously described backcross progenies (liver tissue from ESC challenged backcross progenies was 14th dataset, gill tissue from low oxygen challenged backcross progenies were 16th dataset) and liver tissue and gill tissue from F1 hybrids challenged with heat stress (15th dataset). With the increase of the added RNA-Seq datasets and reads input, the percentage of genome being transcribed also increased gradually, until the 12th dataset. The curve was increased sharply from 11th dataset of 70.9% to 12th dataset of 77.7%. However, after the 12th dataset, even though more datasets were added, the curve became plateau to the 16th dataset, where all the transcripts covered 79.7% of the channel catfish genome length, suggesting that the RNA-Seq probably represented the full transcription potential.

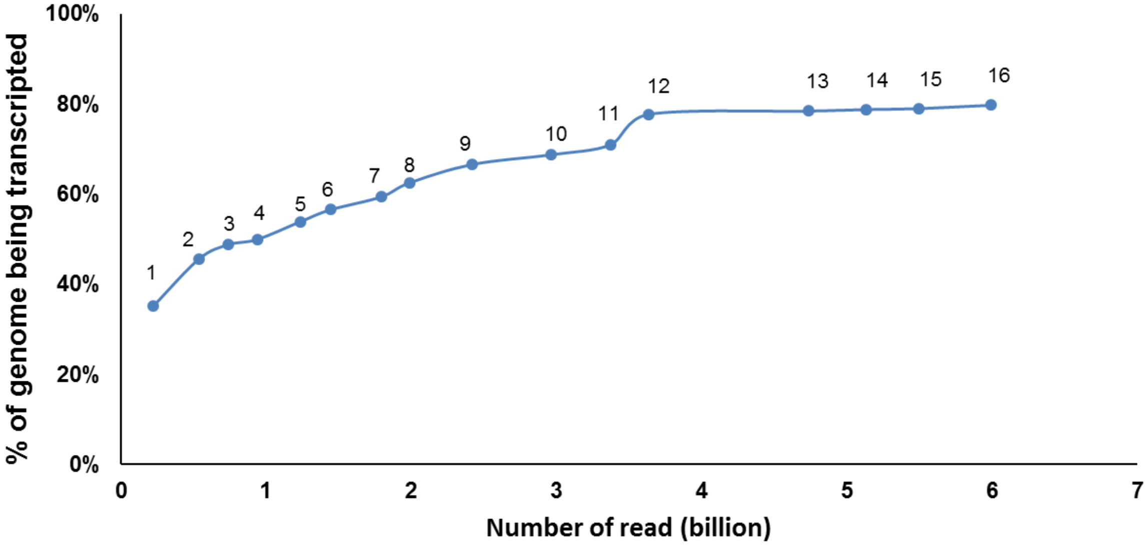

Figure 1. Assessment of pervasive transcription in channel catfish. The X-axis represented the number of reads used to assess the pervasive transcription in billion, while the Y-axis represented the how much genome had been transcribed in percentage.

Genome-wide expression profiles of channel catfish

Genome-wide expression profiles were examined in order to assess the pervasive expression levels. Taken advantage of the large amount of RNAs that's been sequenced, several RNA-Seq datasets under normal conditions were selected to perform the genome-wide expression profiles. The normalized RPKM expression values were computed by counting the library size normalized reads coverage across 50 kilobase bin tiling the channel catfish genome. The normalized RPKMs were depicted along each chromosome of channel catfish genome in Figure 2 with log transformation to $\log_2$ (normalized RPKM + 1) for better visualization. As shown in Figure 2, expression was detected from the vast majority of the 50 kb bins. However, some regions were highly expressed with the highest $\log_2$ (normalized RPKM + 1) of 8.85, which was normalized RPKM equaled to 488, in chromosome 1 from 7,450,000 bp to 7,500,000 bp, while some other regions were very lowly expressed with bars that could barely see, the lowest region was in chromosome 25 from 11,150, 000 bp to 11,200,000 bp, with a $\log_2$ (normalized RPKM + 1) of 0.0013, which was normalized RPKM equaled to 0.0009.

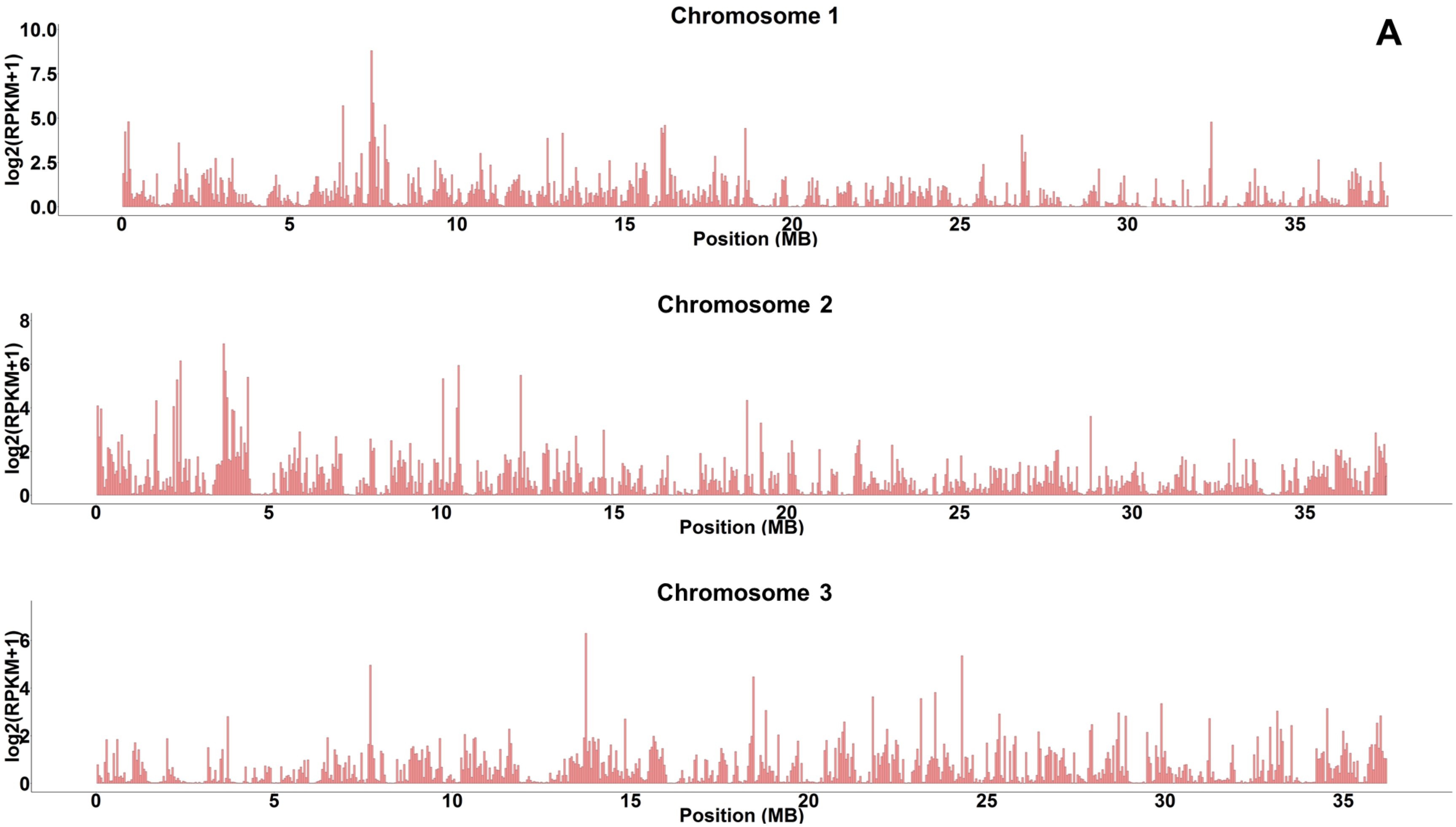

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

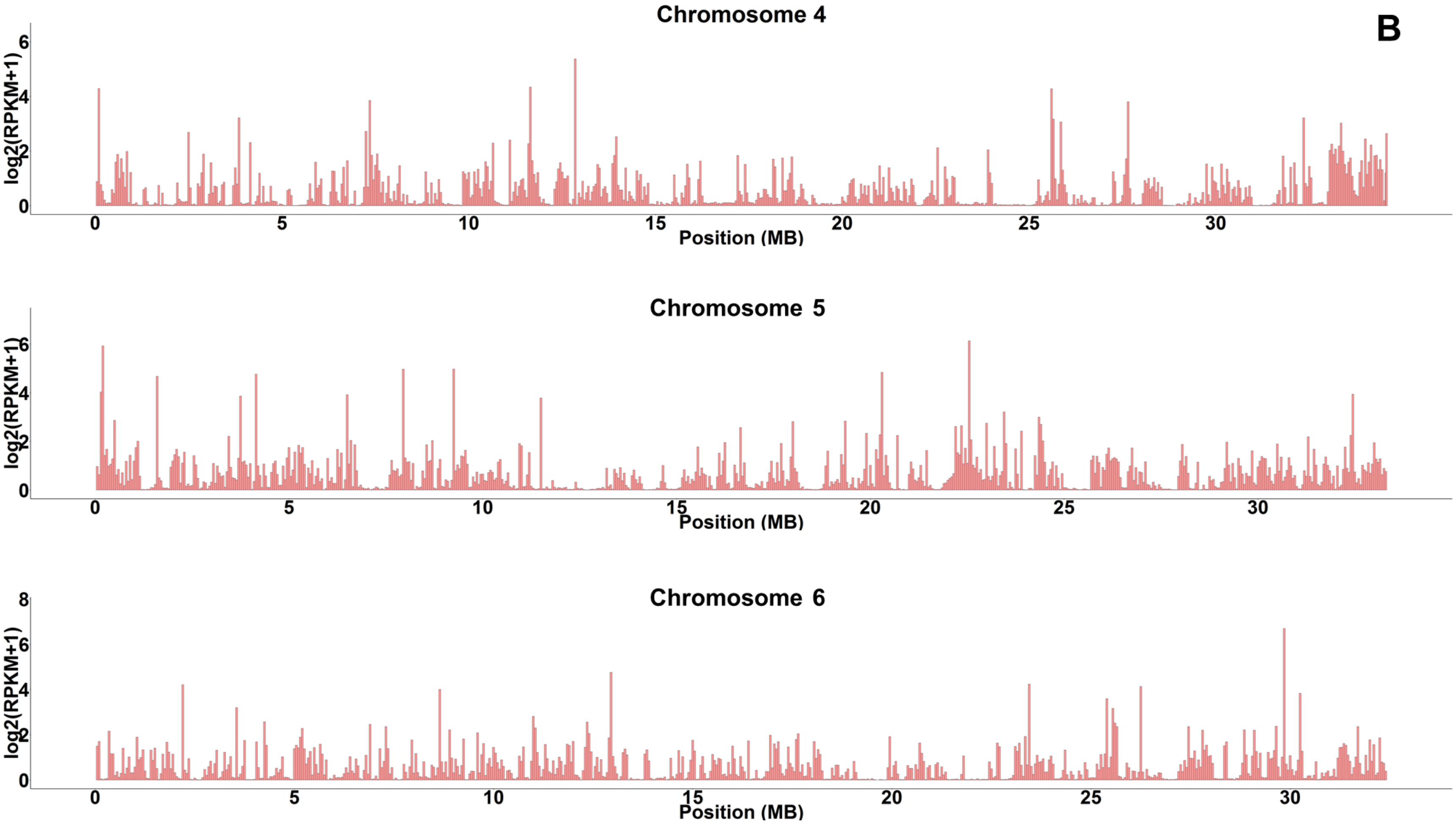

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

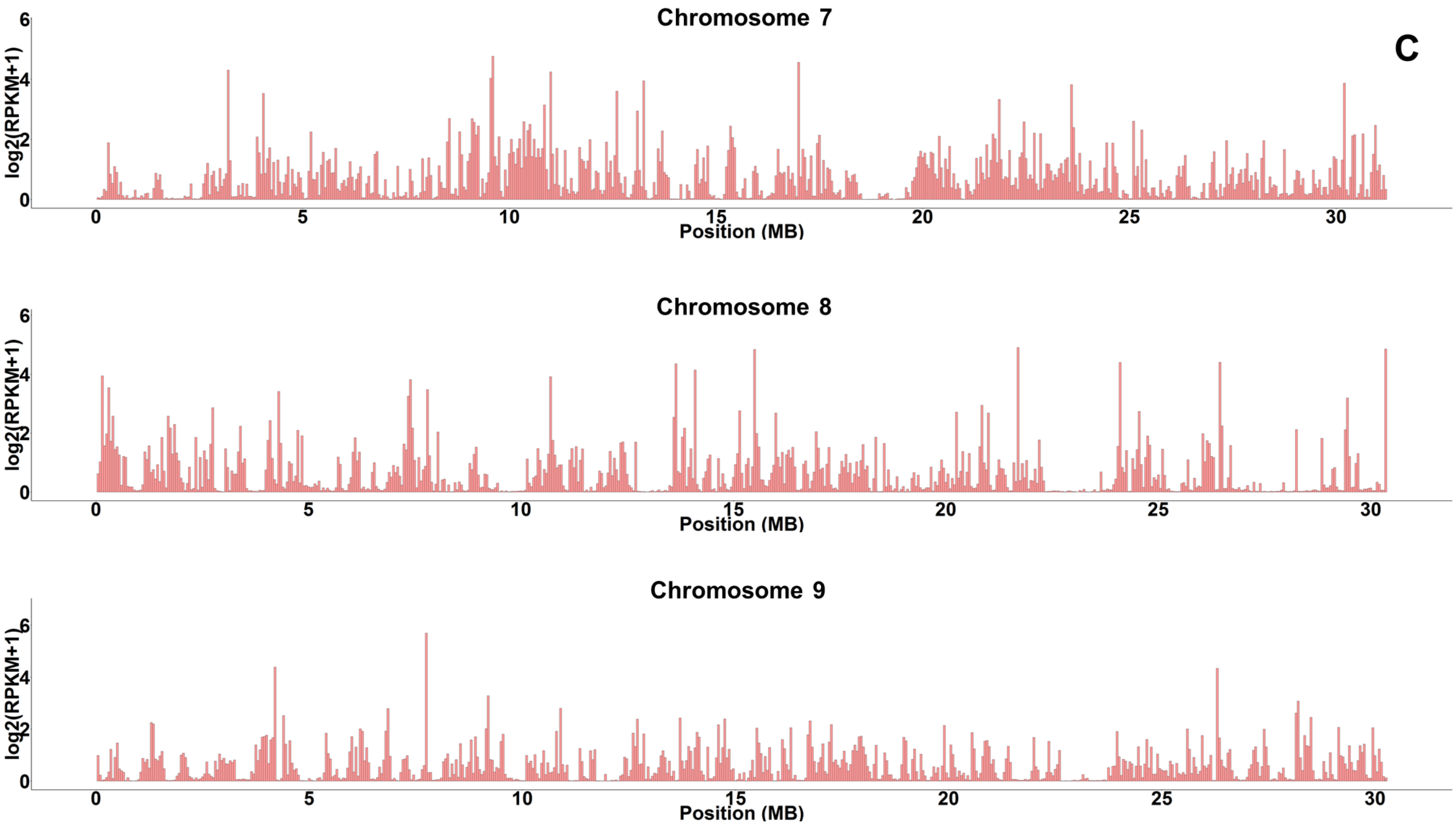

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

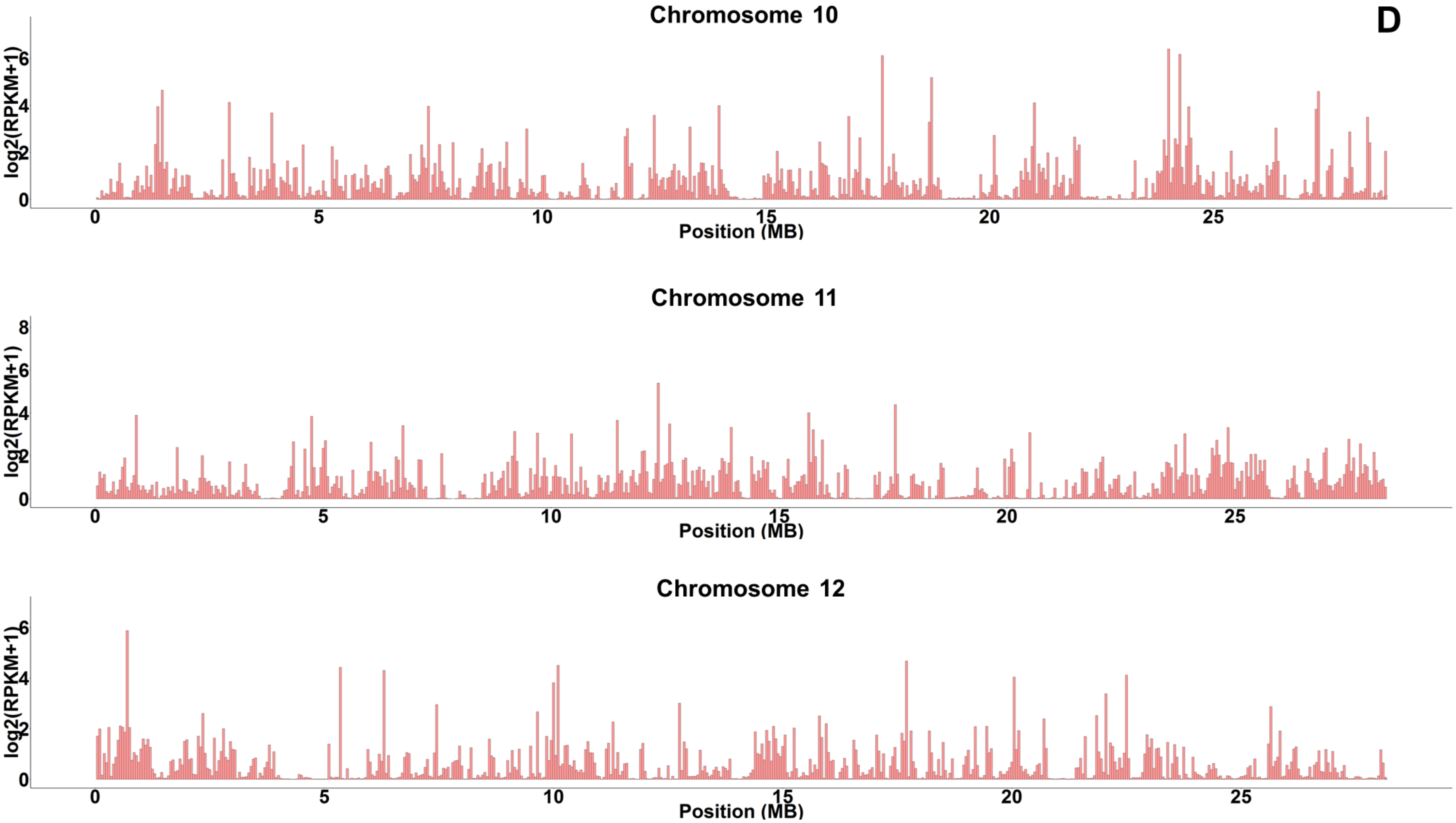

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

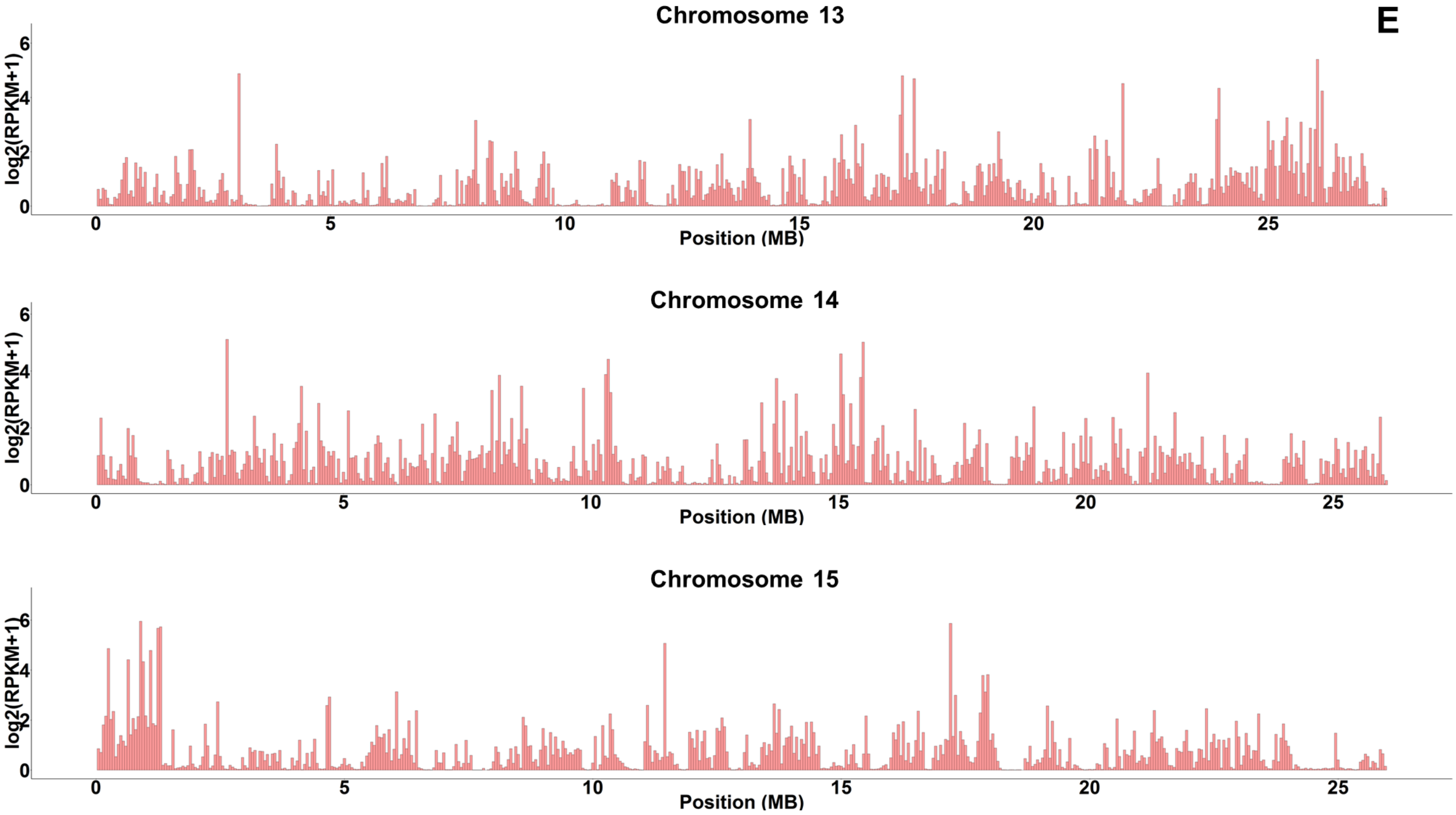

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

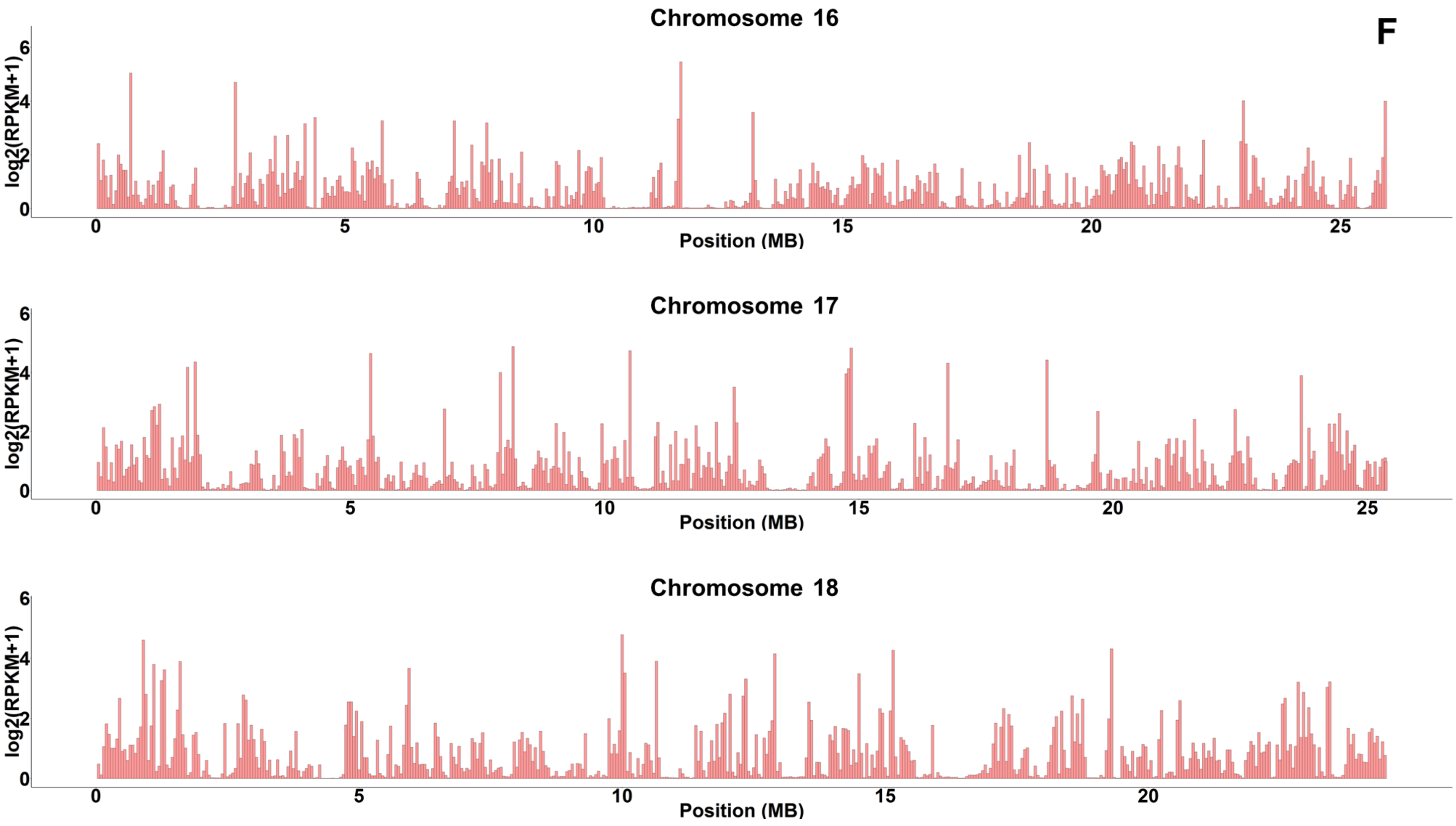

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

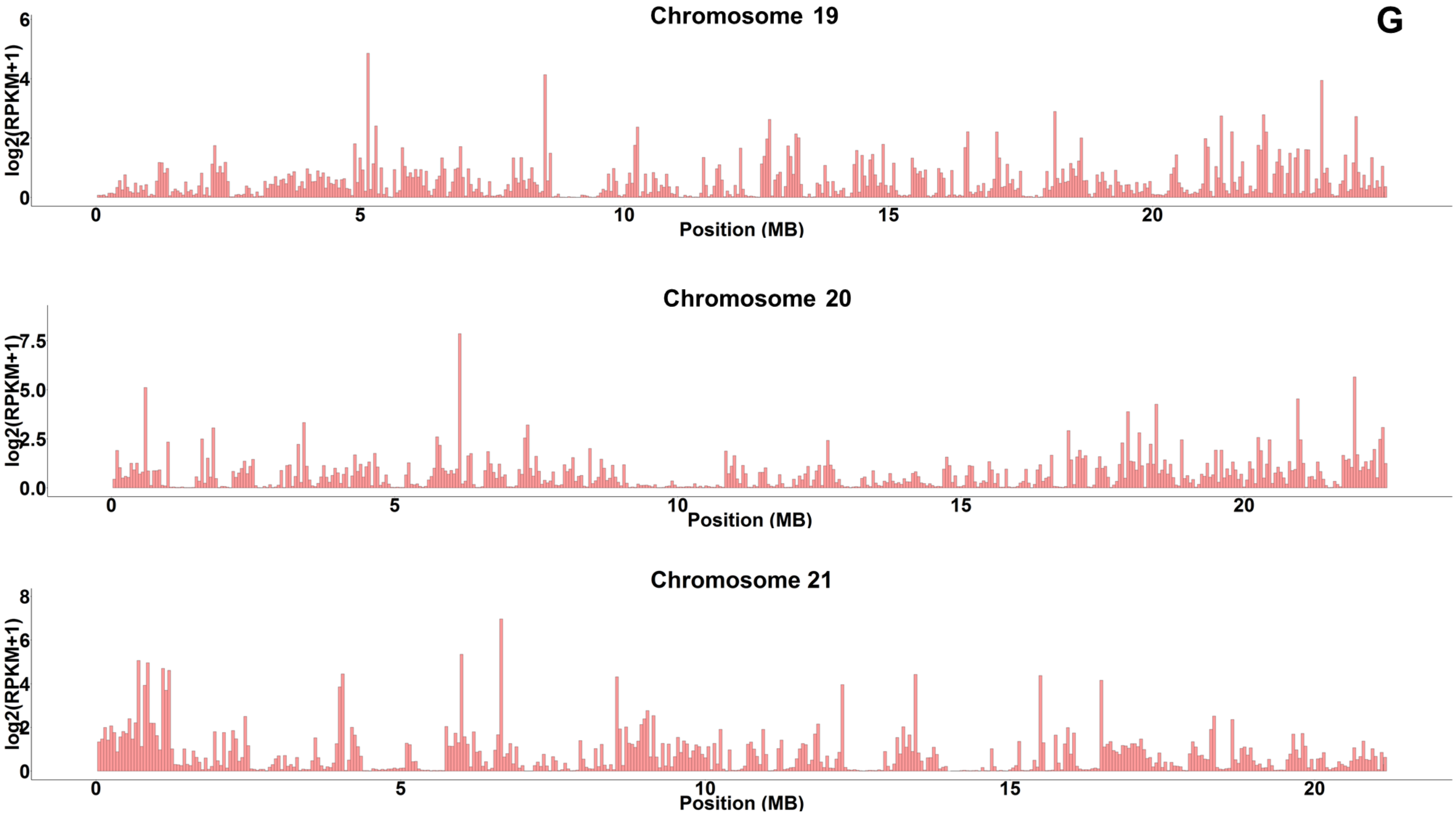

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

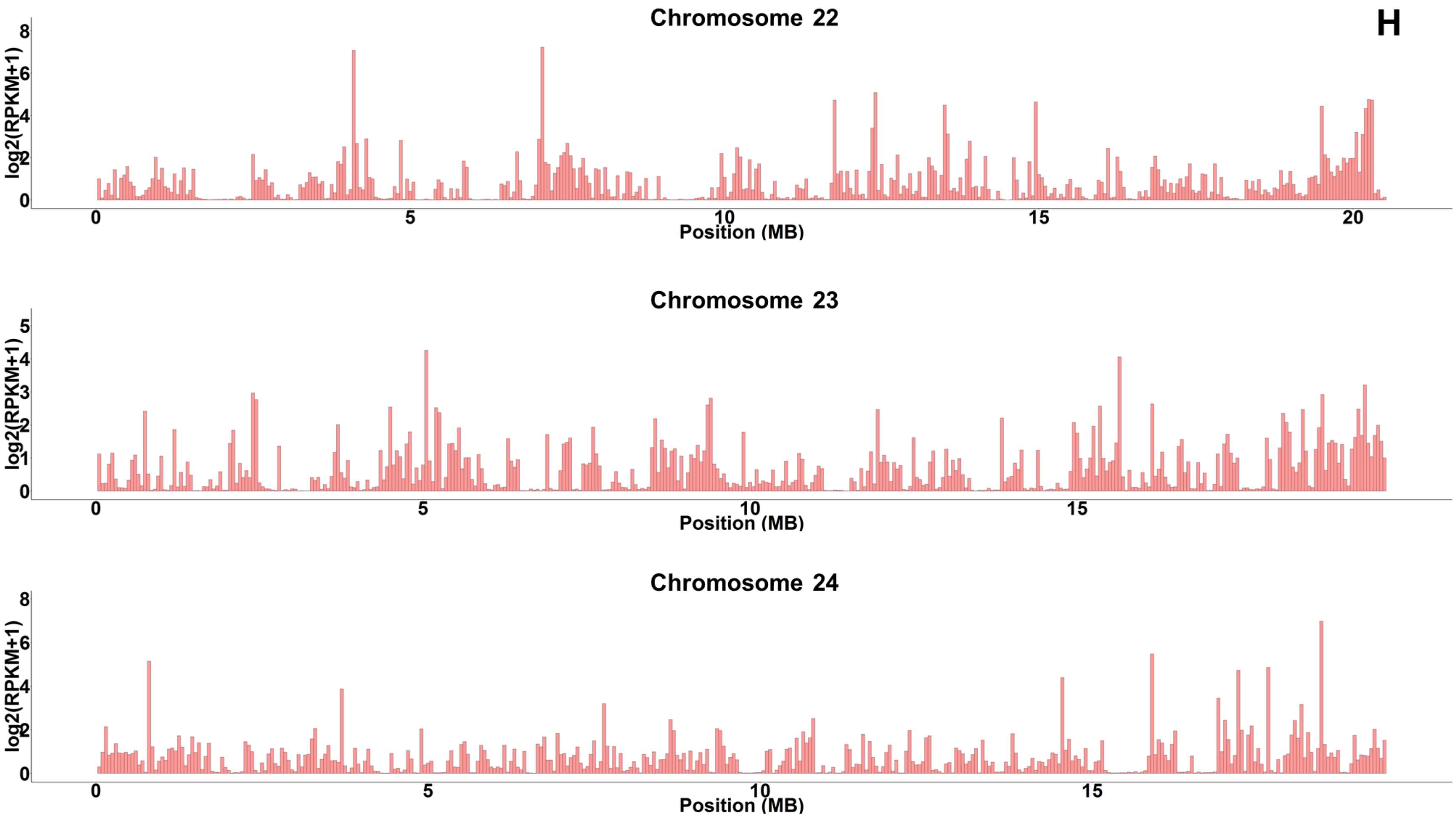

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value log_2 (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

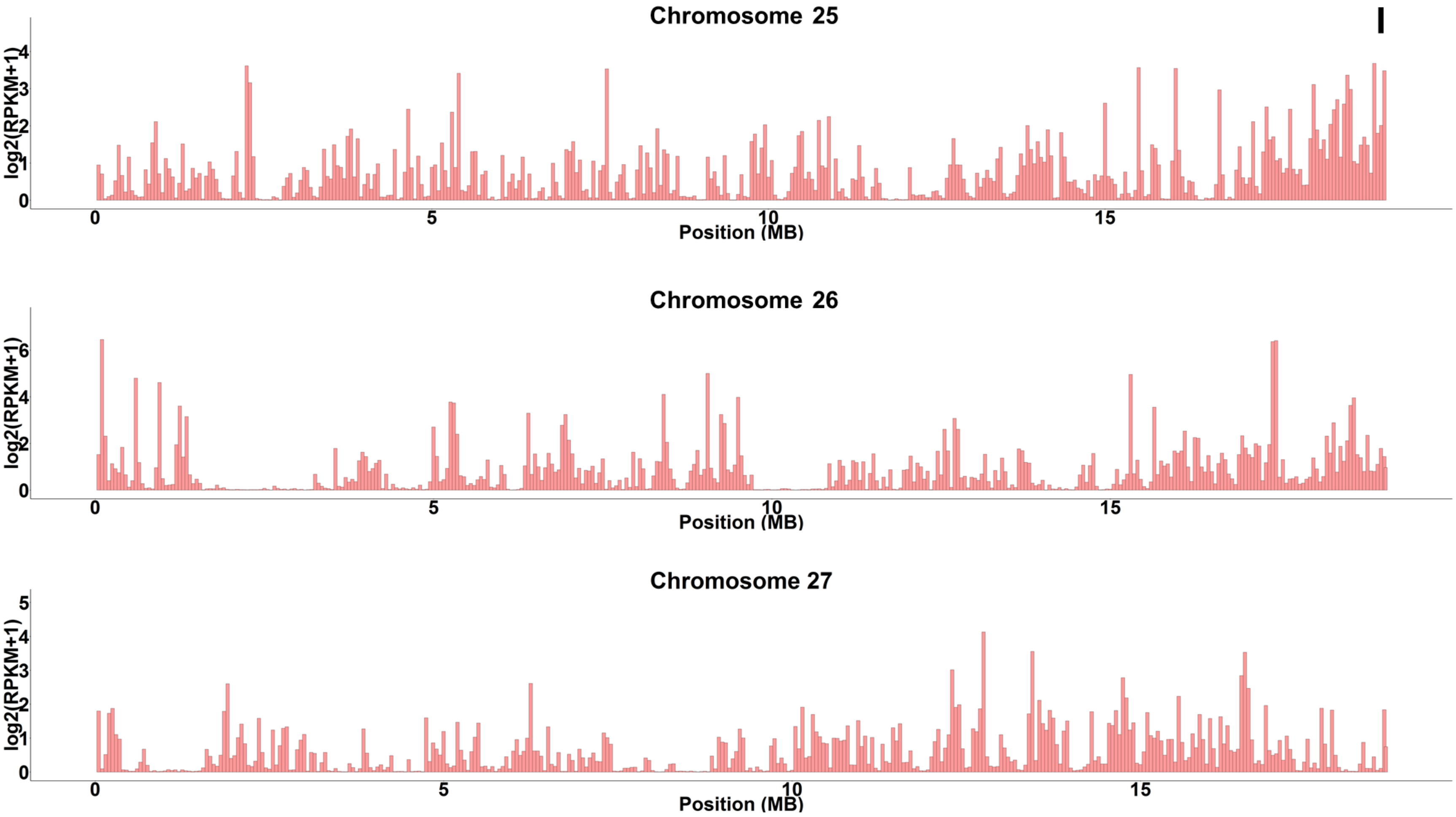

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

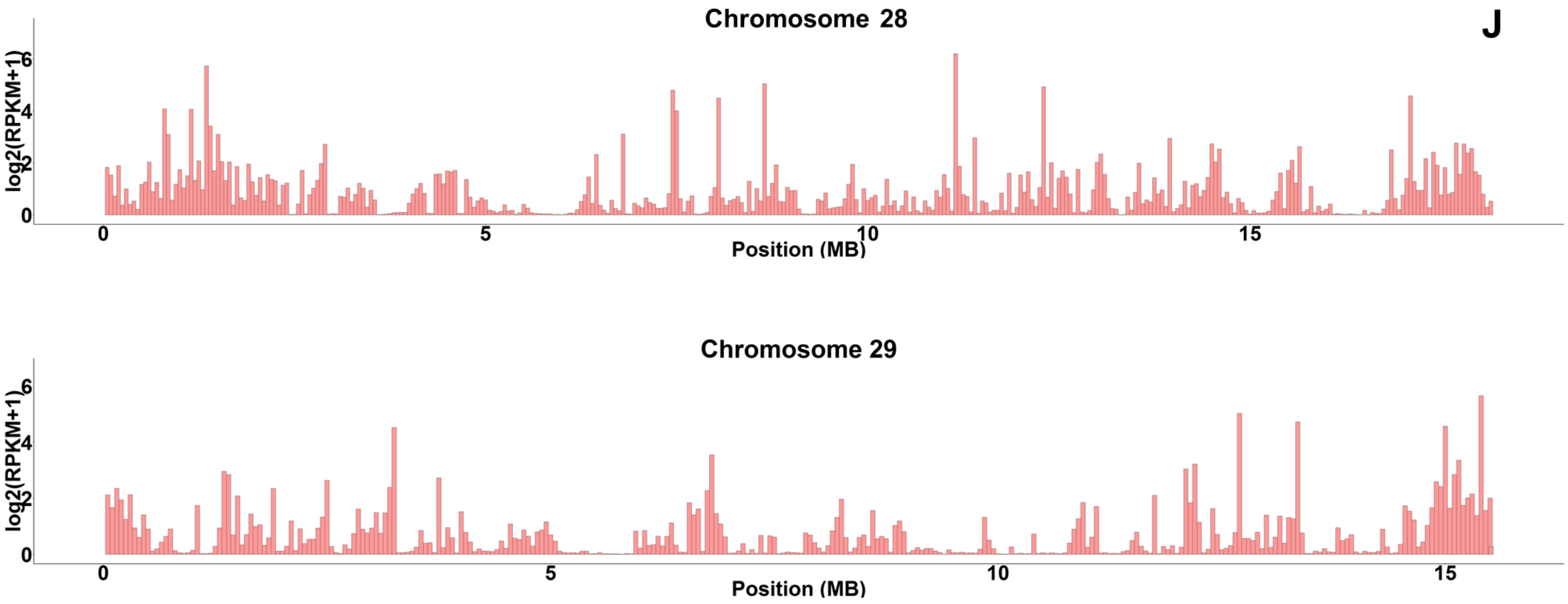

Figure 2. Genome-wide expression profiles of channel catfish. X-axis represented the position of the 50K bin along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

Protein-coding gene expression profiles of channel catfish

As dramatic range of genome-wide expression profiles were observed across channel catfish genome, the highly active region were expected to be mostly associated with the protein-coding genes. Therefore, protein-coding gene expression profiles were established to reveal the channel catfish genome transcription active regions. The normalized RPKM expression values were calculated based on raw read counts for each protein-coding genes that annotated from genome. The normalized RPKMs were depicted along each chromosome of channel catfish genome in Figure 3 with log transformation to $\log_2$ (normalized RPKM + 1) for better visualization. As shown in Figure 3, the expression profiles in protein-coding genes were much more dramatic than that across genome-wide 50 kilobase bin. The highest peak was also shown in chromosome 1, in the same highest region of genome-wide 50 kilobase bin expression profiles, from 7,493,092 bp to 7,494,395 bp, with a $\log_2$ (normalized RPKM + 1) of 14.22, which was normalized RPKM equaled to 17929. The highest expressed gene in this region was Apolipoprotein C-I, which was an inhibitor of lipoprotein binding to the low density lipoprotein (LDL) receptor, LDL receptor-related protein, and very low density lipoprotein (VLDL) receptor. However, protein-coding genes were not all highly expressed, 341 protein-coding genes were barely expressed with a $\log_2$ (normalized RPKM + 1) of 0, which was normalized RPKM equaled to 0 either. These genes were either not transcribed or not being properly sequenced or expressed extremely low that were normalized to 0. In addition, of these 341 protein-coding genes with a 0 normalized RPKM expression value, most of them (211) were not anchored in chromosomes, which means that they were detected from the short scaffolds, the short length of scaffolds also makes it hard for mapping to detect the real expression values of these genes.

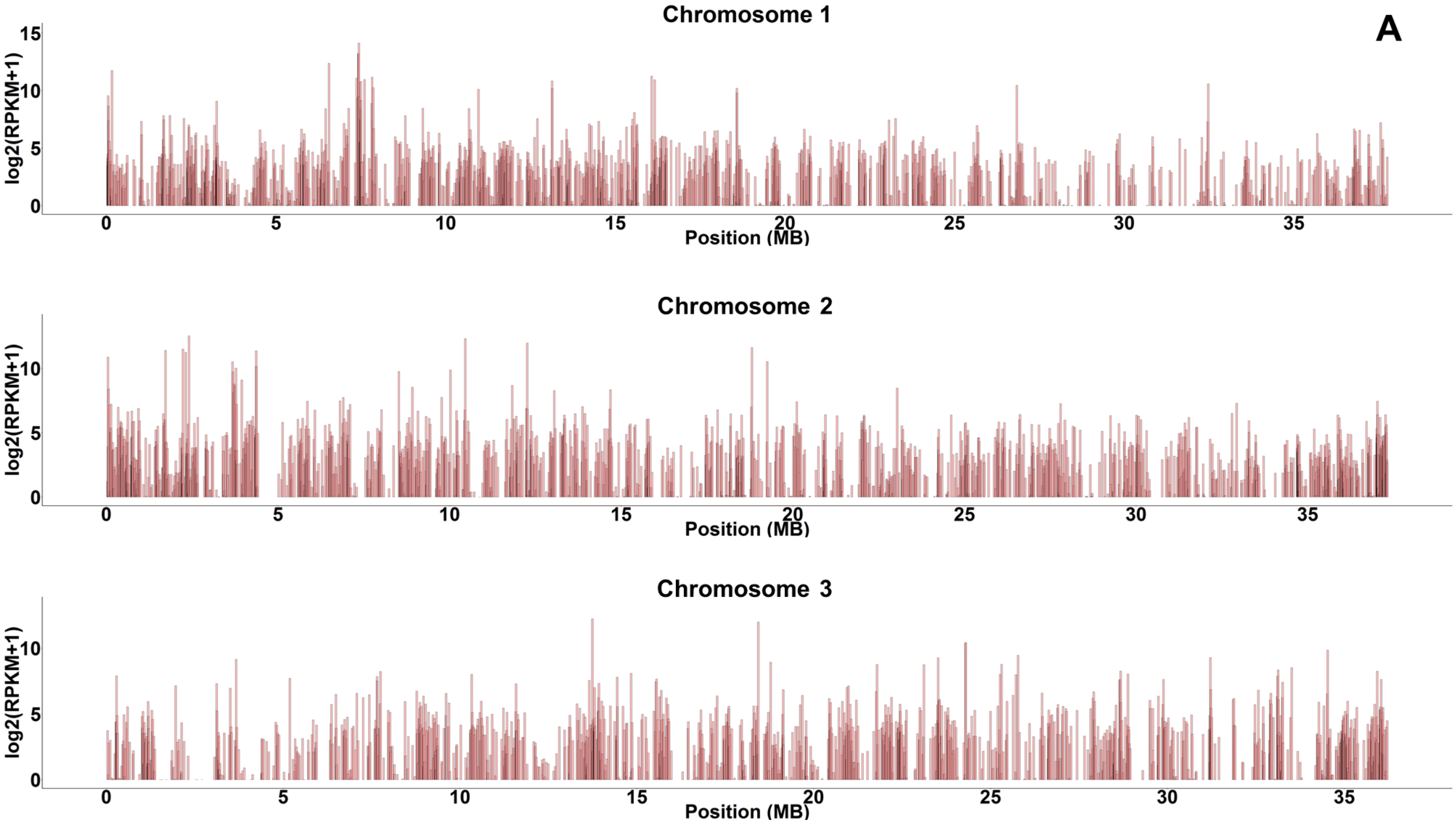

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

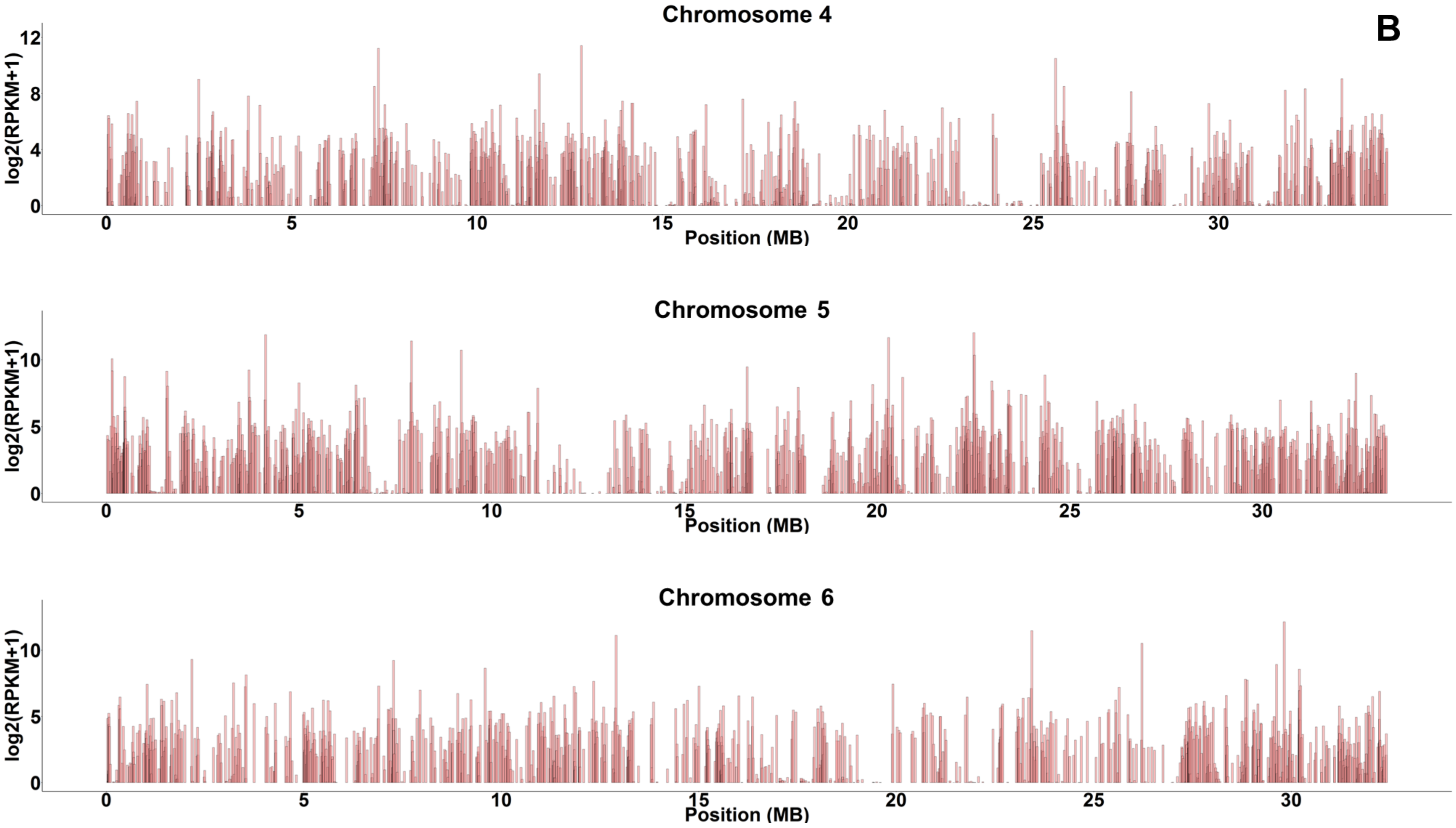

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

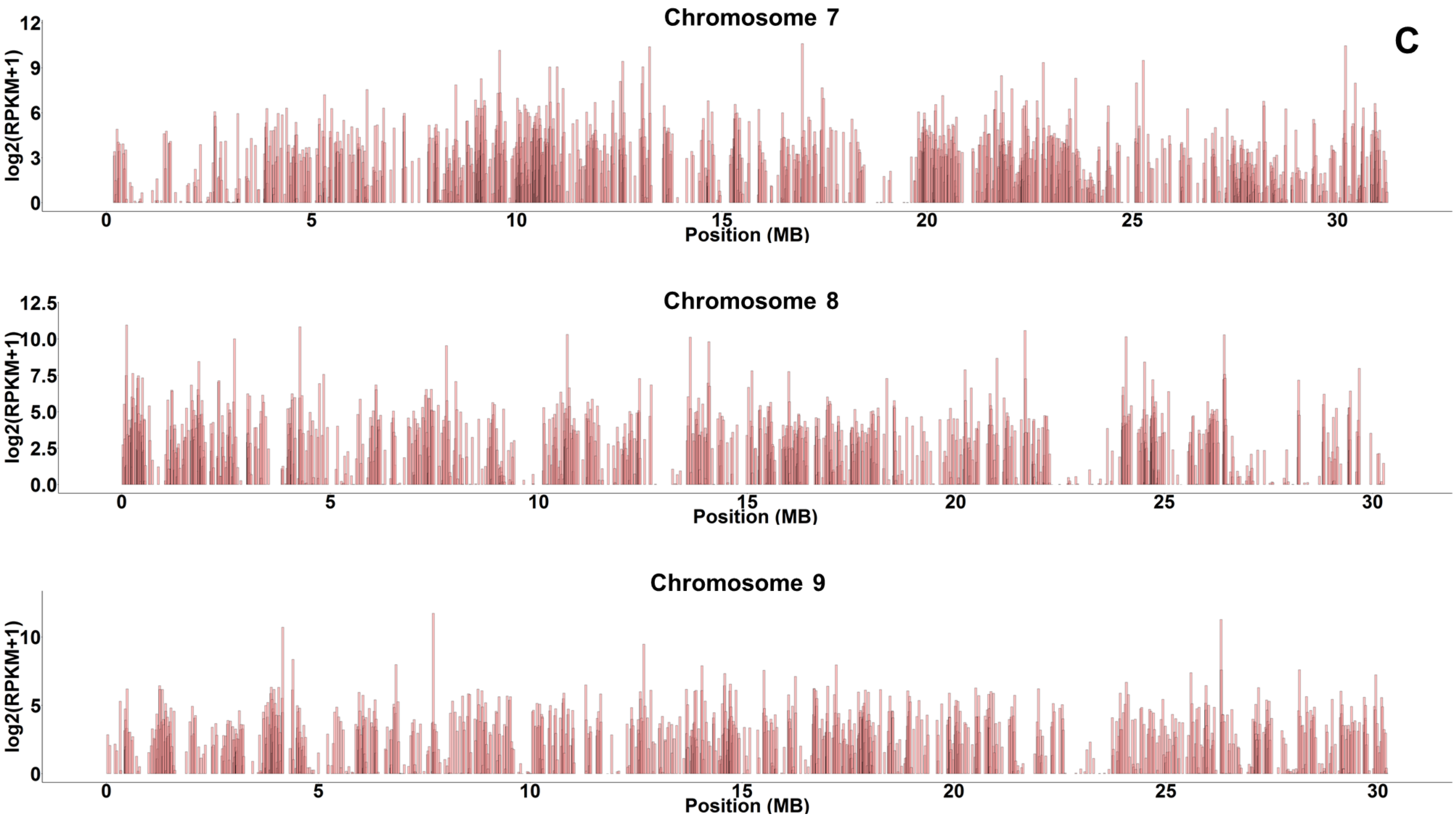

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

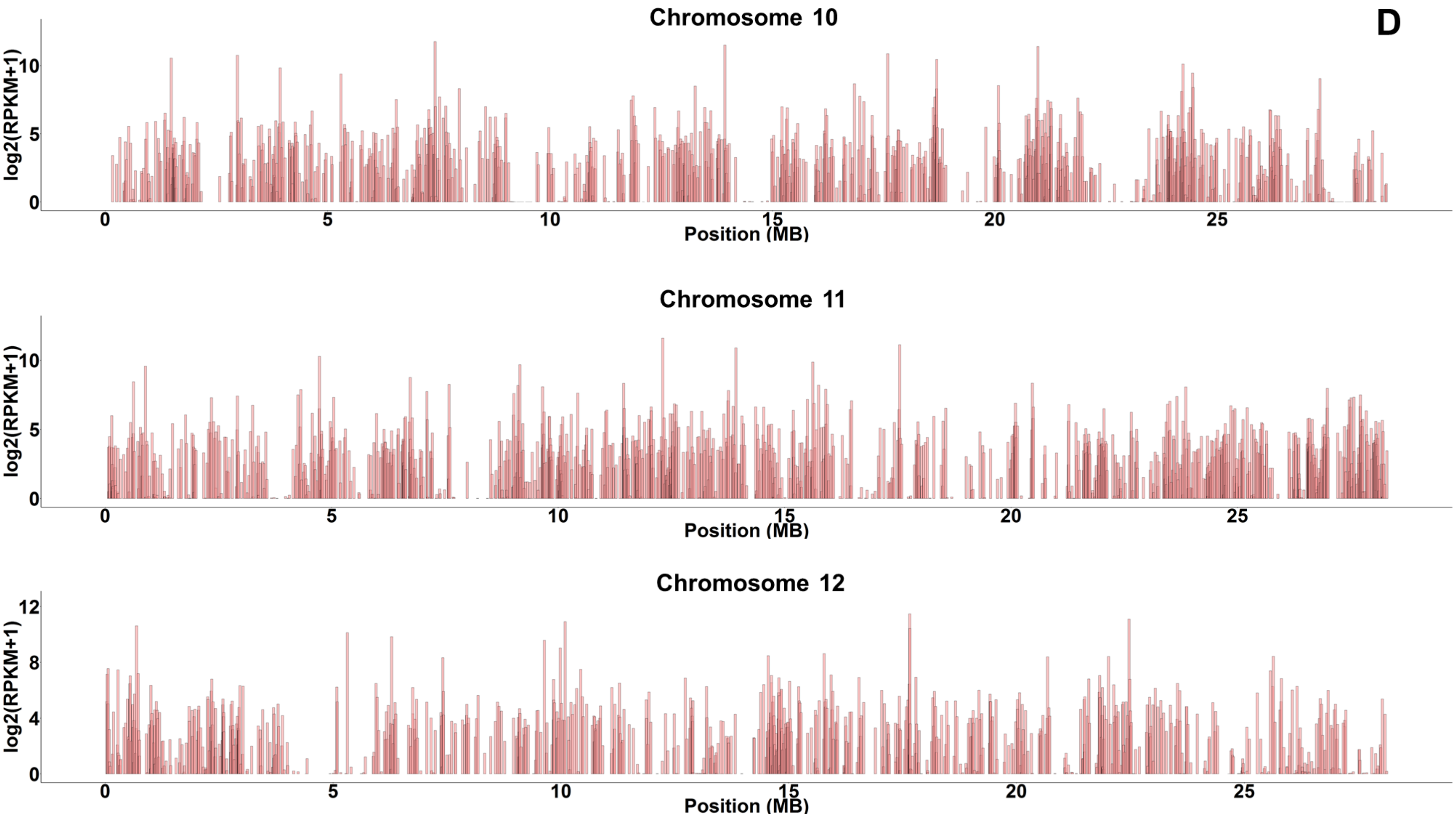

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

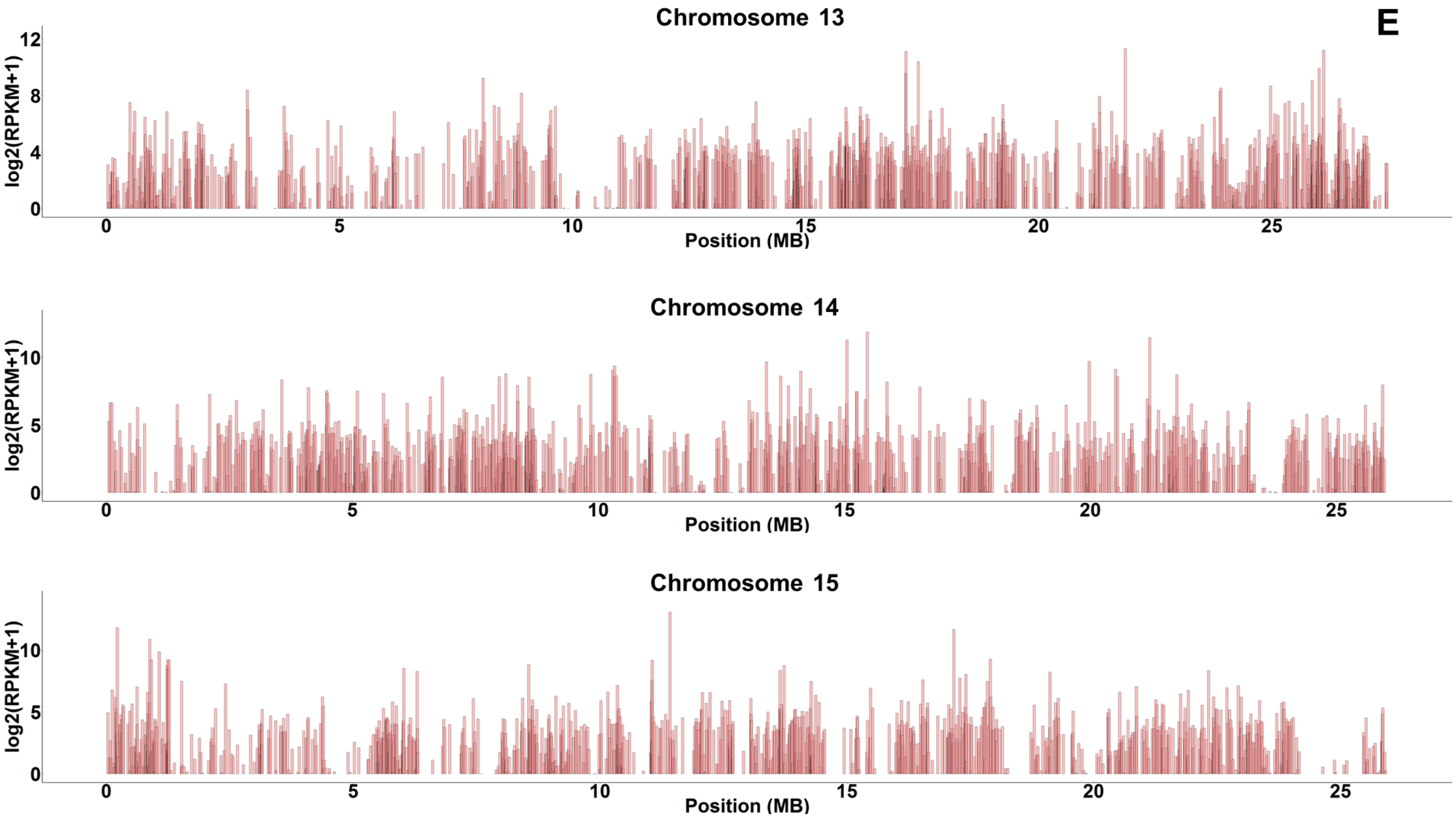

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

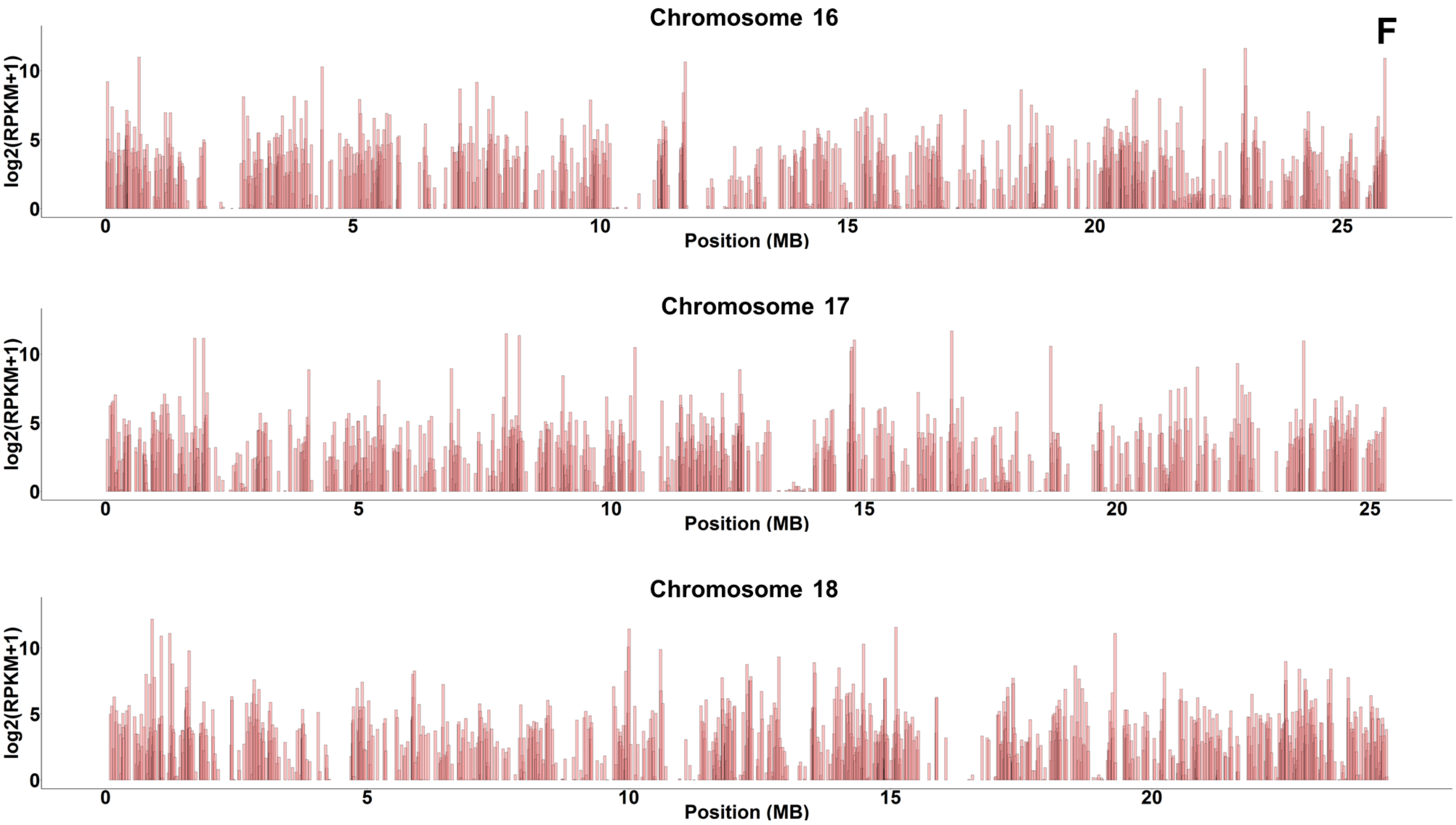

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

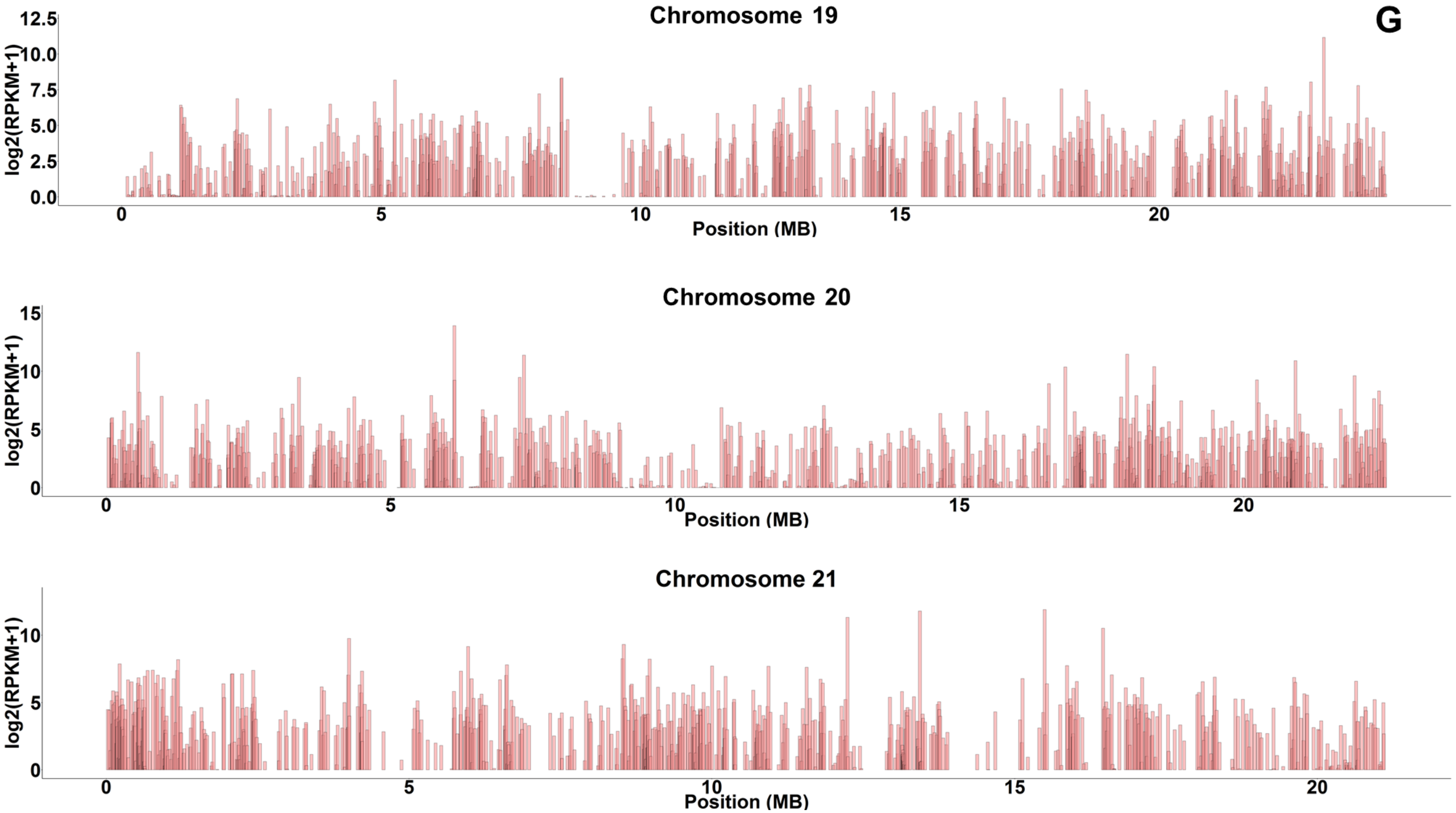

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

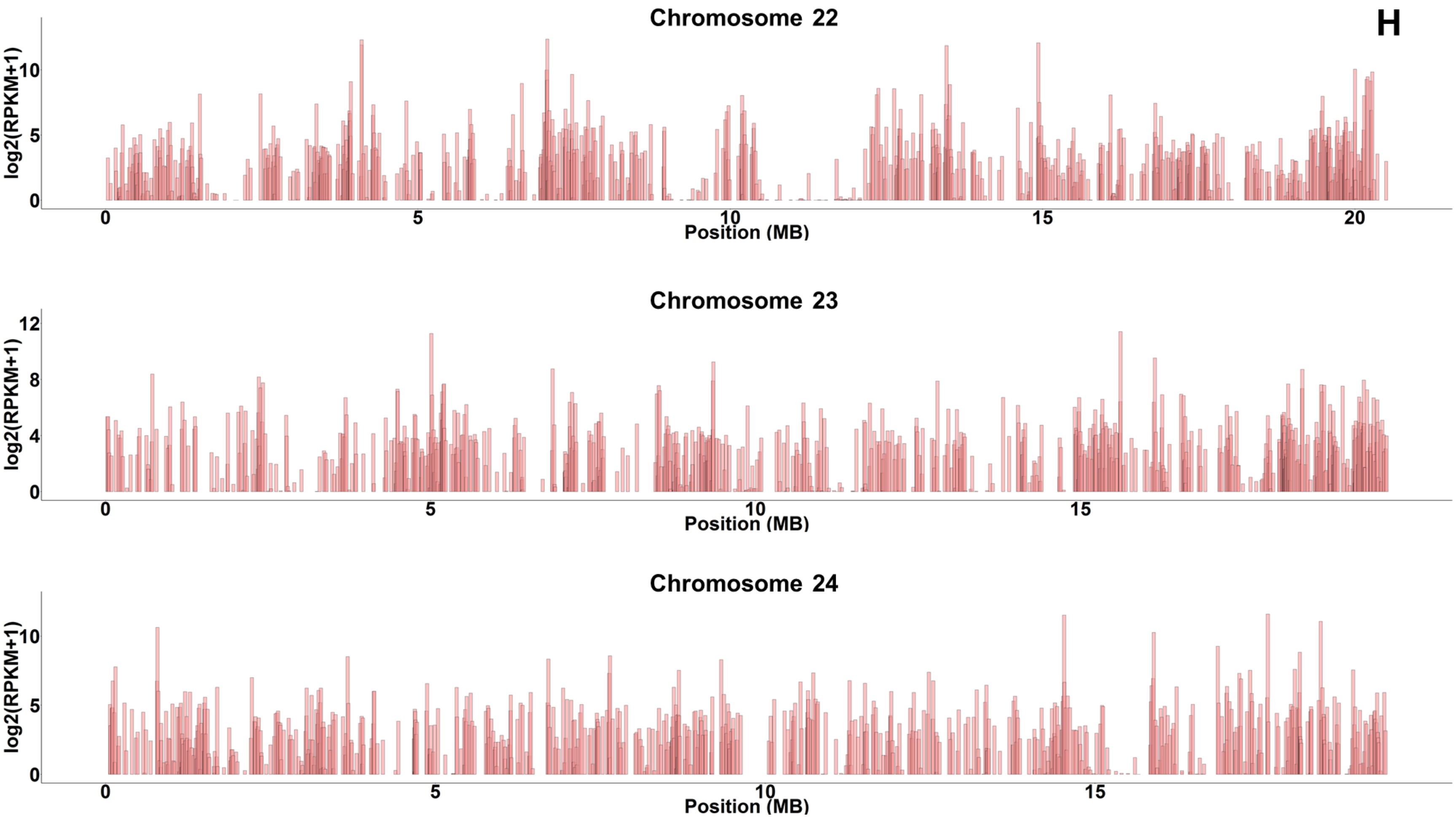

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

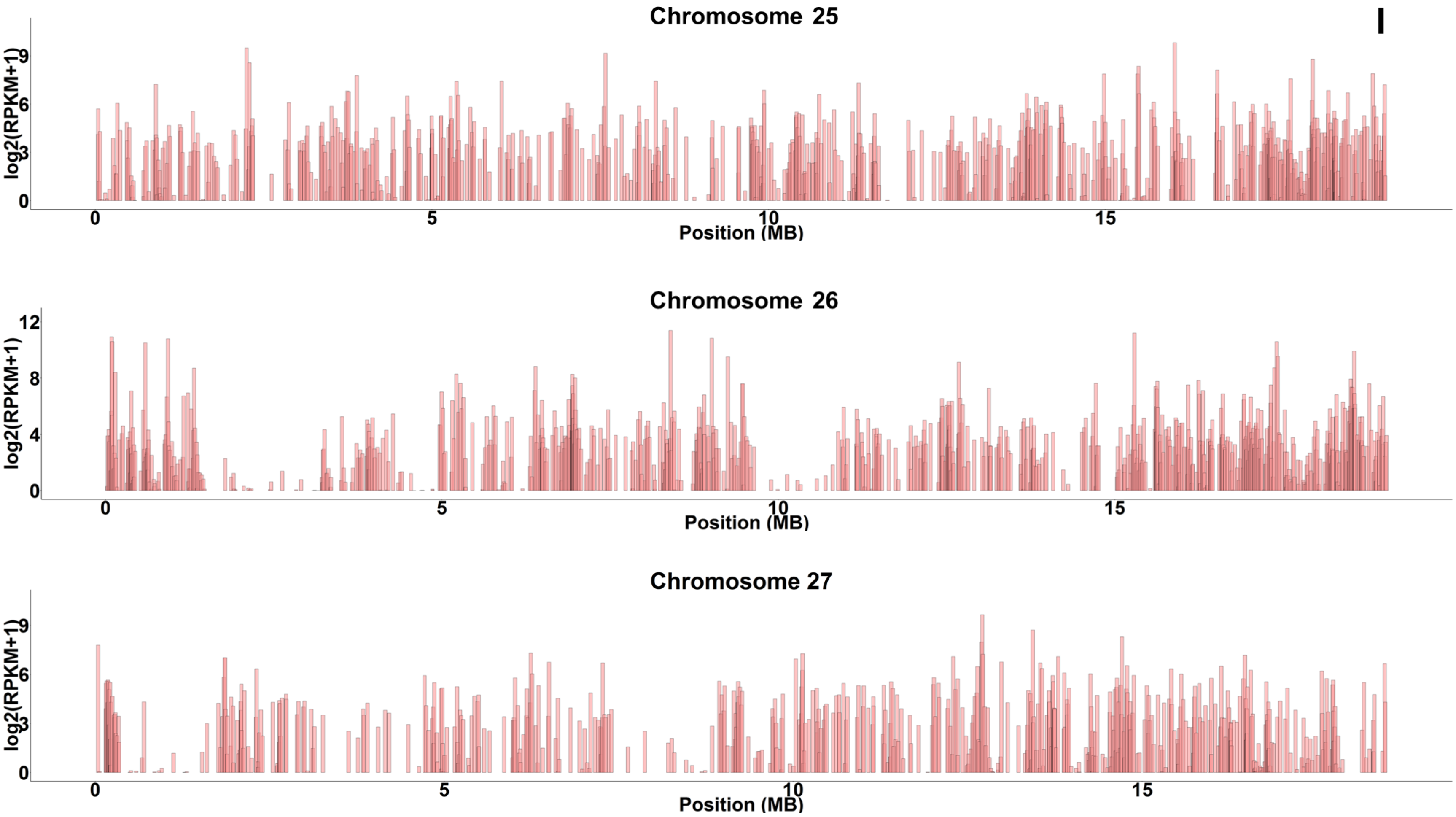

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

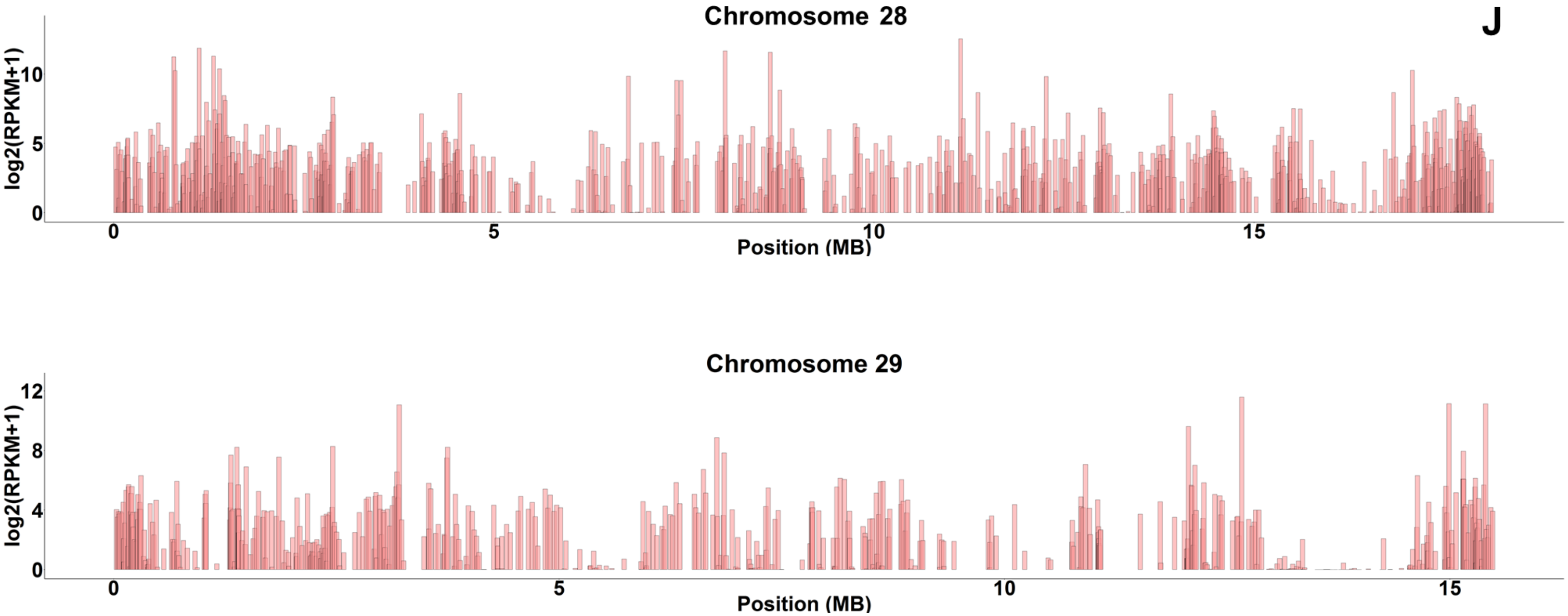

Figure 3. Protein-coding gene expression profiles of channel catfish. X-axis represented the position of the protein-coding genes along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

In order to assess the protein-coding genes expression pattern in normal condition, all normalized expression values of all protein-coding genes were classified and visualized in Figure 4. The normalized RPKMs were log transformed to $\log_2$ (normalized RPKM + 1) for better visualization. As shown in Figure 4, many genes were expressed with a $\log_2$ (normalized RPKM + 1) of less than 1, also normalized RPKM expression value less than 1, which represented relatively lowly expressed group of genes. Liver had the most lowly expressed genes (11,375 genes), followed by intestine (10,705 genes) and ovary (10,253), while testis had the least lowly expressed genes (8,061 genes), followed by skin (9,323 genes) and gill (9,797 genes). Genes that were expressed with a normalized RPKM value of more than 127 was considered highly expressed in this study, which was $\log_2$ (normalized RPKM + 1) more than 7. Liver also had the most highly expressed genes (1,461 genes), followed by ovary (1,287 genes) and intestine (1,201 genes). On the other hand, skin had the least highly expressed genes (641 genes), followed by gill (704 genes) and testis (743 genes). The normalized RPKM expression value between 1 and 127 were then considered as intermediately expressed genes, which testis had the highest amount (17,857 genes) and liver possessed the least (13,826 genes).

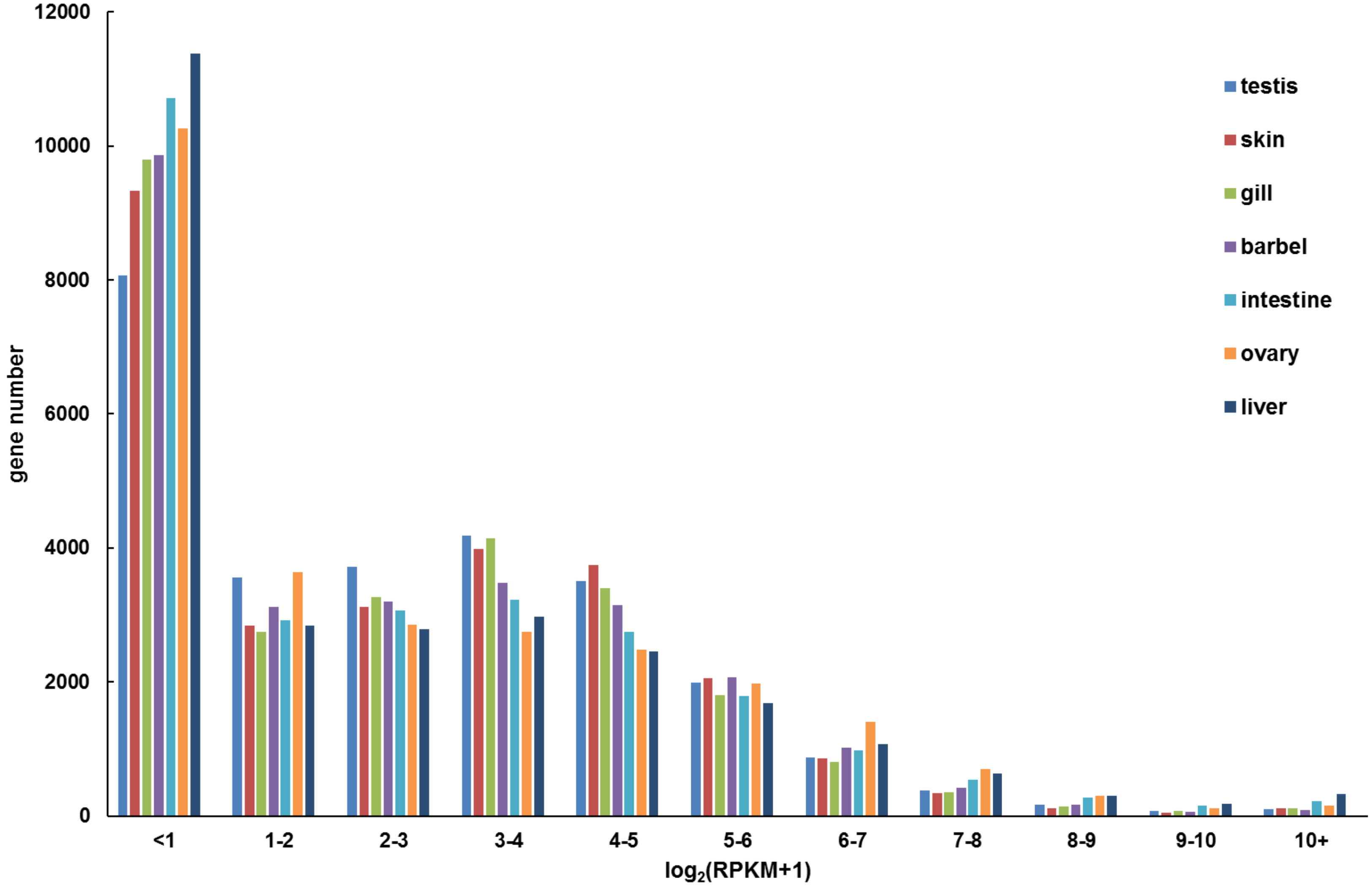

Figure 4. Protein-coding genes expression level distribution in different tissues under normal condition. X-axis represented the log transformed normalized RPKM, $\log_2$ (normalized RPKM + 1) for expression levels, Y-axis represented the number of genes that had certain range of expression values.

Identification of tissue-specific expressed genes in channel catfish

Tissue-specific expressed genes were considered as genes whose functions and expressions were favored in a specific tissue type (Xiao et al., 2010). These genes were important for an organism to maintain specificity and complexity since the tissue-specific expressed genes may affect the process of development, function and maintenance of diverse cell types within an organism (Salem et al., 2015). In the present study, we took advantage of the various previously conducted RNA-Seq datasets, collecting total of eight tissues for comparison to identify tissue-specific expressed genes, including barbel, gill, intestine, liver, skin and testis. By mapping reads from each collected tissue to channel catfish genome, the normalized expression level of each gene in each tissue was calculated, tissue-specific expressed genes were only identified if the fold change of one specific tissues against the rest of the tissues in expression level was at least 32 fold with the FDR (False discovery rate) adjusted p value of less than 0.05. A total of 1,455 genes were identified as tissue-specific expressed genes based on the above criteria, the number of tissue-specific genes predicted in different tissues were listed in Table 6 and detailed tissue-specific genes with their fold changes were summarized in Supplemental Table 2. Liver showed the most tissue-specific expressed genes, which was 377 genes, followed by testis (326 genes) and intestine (244 genes). Conversely, gill showed the lowest number of tissue-specific expressed genes, which was only 82 genes, followed by skin (103 genes) and barbel (108 genes). In order to assess the specificity of these tissue-specific expressed genes, distribution of the differentially expressed fold change of all tissue-specific expressed genes were visualized in Figure 5. As shown in Figure 5, liver had the most tissue-specific expressed genes, and most of the genes (242 genes) were expressed over 128 fold higher than all other seven tissues, while 49 genes were expressed 64 to 128 fold higher than all other six tissues and 86 genes were expressed 32 to 64 fold higher than all others. Similarly, in intestine, most of the tissue-specific genes (132) were expressed over 128 fold higher than all other seven tissues, less genes were expressed 32 to 128 fold higher than all other tissues (63 genes expressed 32 to 64 fold and 49 genes expressed 64 to 128 fold). However, different pattern was seen in testis tissue, most of the tissues-specific genes (167 genes) were expressed 32 to 64 fold higher than all other seven tissues, less genes were expressed over 64 fold higher than all other tissues (84 genes expressed 64 to 128 fold and 75 genes expressed over 128 fold).

Table 6. Number of tissue-specific (at least 32 fold higher than its expression in any of the other tissues) genes predicted in different tissues

Tissues	**Gene**
Barbel	108
Gill	82
Intestine	244
Liver	377
Ovary	215
Skin	103
Testis	326

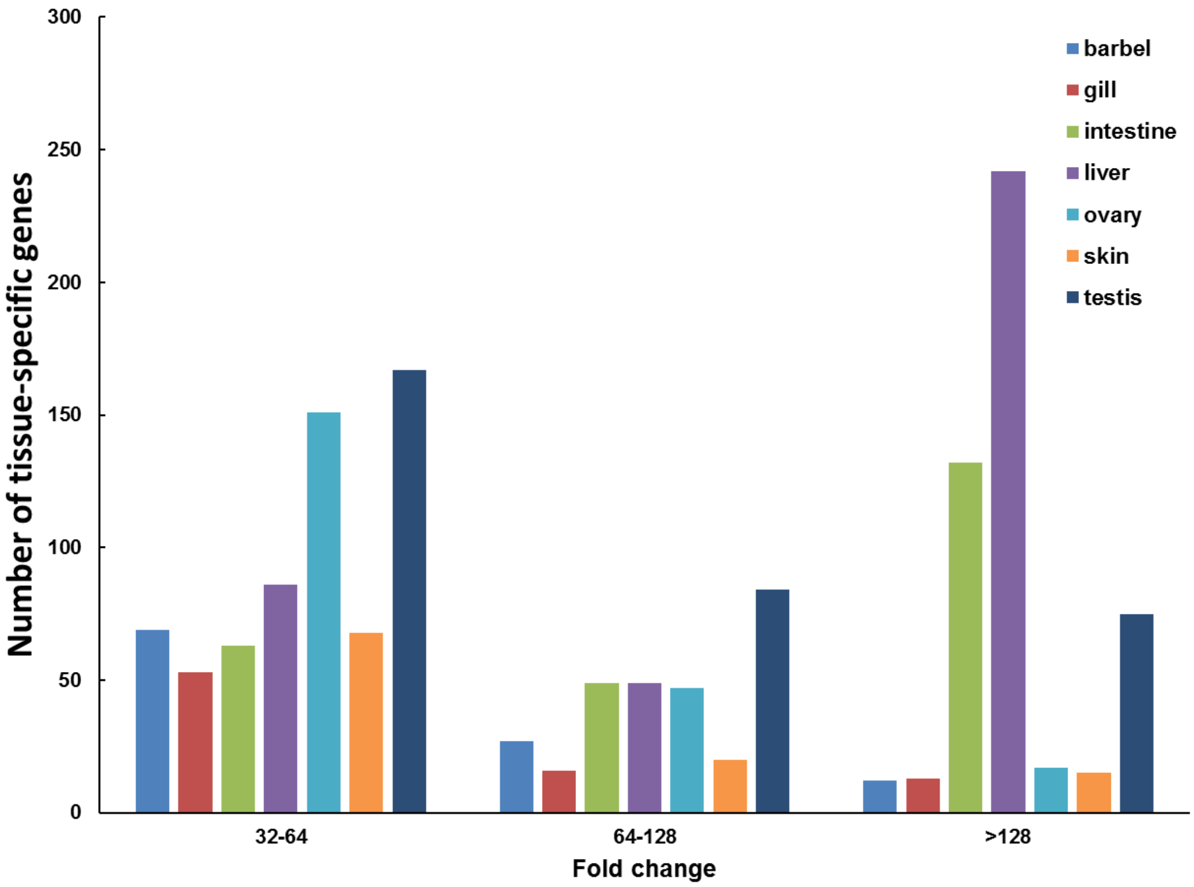

Figure 5. Number of tissue-specific protein-coding genes in different scales of fold change. X-axis represented the three different scales of fold change with the FDR corrected p-value less than 0.05, which were 32 to 64 fold, 64 to 128 fold, and over 128 fold. Y-axis represented the number of tissue-specific protein-coding genes that fell in different fold change scales.

Identification of differentially induced expressed protein-coding genes in channel catfish

Differentially induced genes after different disease and stress treatments were determined by comparing their expression levels in normalized RPKM between treatment groups and control groups. The differentially induced genes were only identified if two-fold change expression were observed in at least one treatment and FDR (false discovery rate) corrected p-value < 0.05, the number of differentially induced expressed genes from different treatments were summarized in

Table 7 and visualized in Figure 6. A total of 8,560 genes were differentially expressed across all used challenged RNA-Seq datasets (Supplemental Table 3), as shown in Table 7, the most differentially expressed genes were observed in ESC challenged RNA-Seq dataset, while the columnaris challenged RNA-Seq dataset (0h, 4h, 24h, and 48h) had the least.

Table 7. Differentially expressed protein-coding genes that were induced in different treatments

Treatment	**Gene**
Columnaris (gill: 0h, 4h, 24h, and 48h)	411
Columnaris (gill: 0h, 1h, 2h, and 8h)	2,918
Esc (liver)	3,362
Heat (gill)	2,867
Heat (liver)	2,652
Feed deprivation (gill and skin)	534

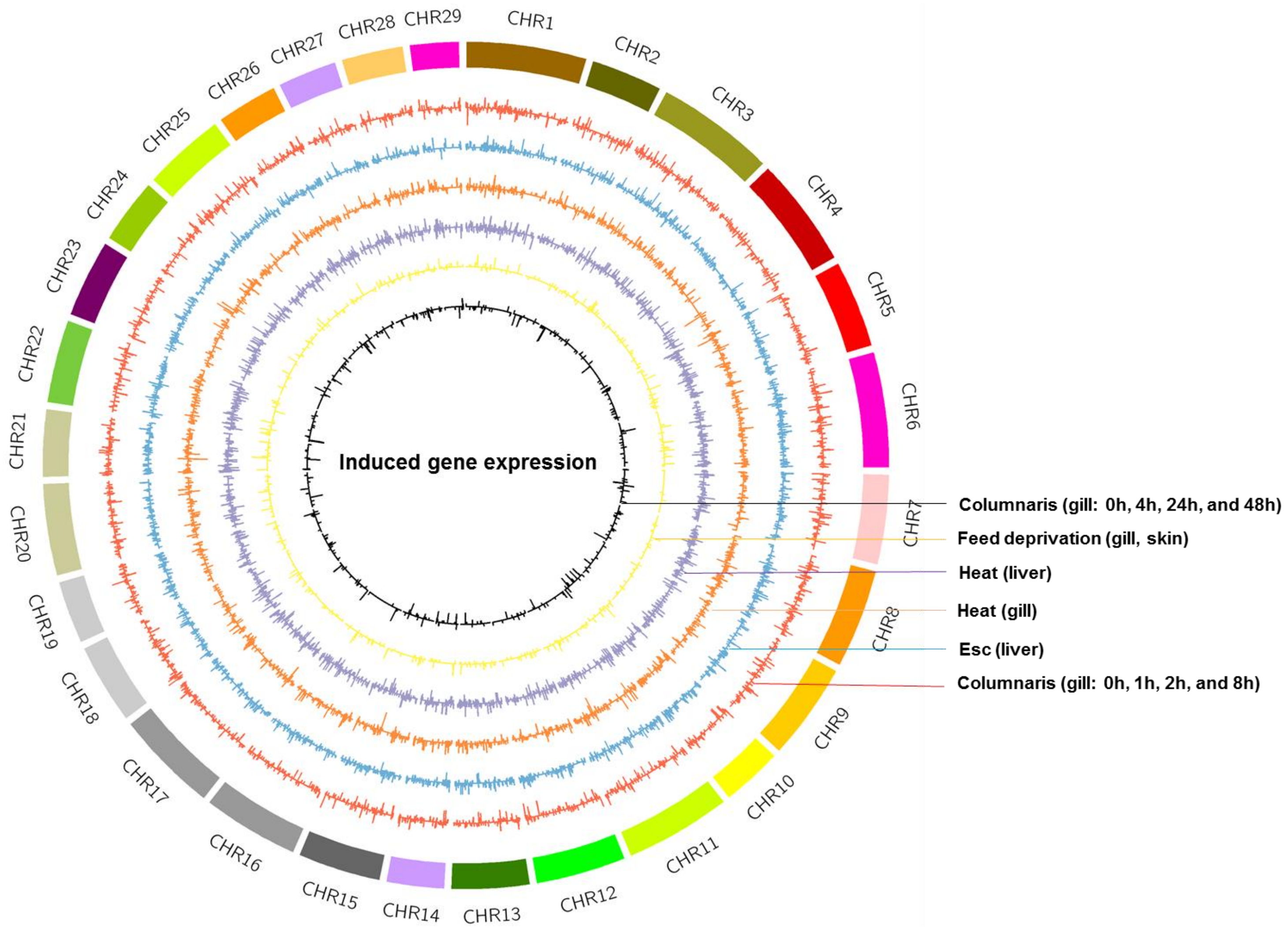

Figure 6. Differentially induced gene expression profiles across different treatments along channel catfish genome. The outer circle represented channel catfish 29 chromosomes, inner six circles represented the highest $\log_2$(fold change) against the control group following different treatments, including columnaris infection (0h, 4h, 24h, and 48h), short-term feed deprivation challenge, heat challenge in two different tissues, ESC infection and two sets different susceptibilities channel catfish challenged with columnaris disease (0h, 1h, 2h, and 8h).

Identification of long non-coding RNAs (LncRNAs) in channel catfish

Genome-guided TopHat-Cufflinks assembled 197,161 transcriptome contigs were used for identification of the lncRNAs. CPAT (Coding Potential Assessment Tool) software was used to eliminate those contigs that have coding capacities, resulting in 77,758 cotings to have coding probability less than 0.38. BLAST analysis was then conducted to eliminate any sequences with hits to known proteins or protein domains. The remaining contigs were further examined to retain those contigs that had an ORF less than 100 amino acids. Lastly, the contigs were excluded if any overlap were observed with either UTRs or channel catfish annotated genes and its 1Kb neighboring region. After all the filtering, a total of 36,266 lncRNAs were identified as lncRNAs.

LncRNAs expression profiles of channel catfish

Channel catfish genome was pervasively transcribed as described above, however, protein-coding genes only account for a small part of genome, in order to process pervasive transcription, non-coding RNAs should also be transcribed. LncRNAs were one of the most important parts of non-coding RNAs, their expression profiles in channel catfish were conducted to initially reveal transcription in non-coding regions. The normalized RPKM expression values were calculated based on raw read counts for each lncRNAs. The normalized RPKMs were depicted along each chromosome of channel catfish genome in Figure 7 with log transformation to $\log_2$ (normalized RPKM + 1) for better visualization. As shown in Figure 7, the highest peak was also shown in chromosome 2, from 3,734,474 bp to 3,734,742 bp, with a $\log_2$ (normalized RPKM + 1) of 16.33 (normalized RPKM equaled to 77,297), which was much larger than the highest expressed protein-coding gene. Similar to protein-coding genes, lncRNAs were also not all highly expressed but with much fewer number, 28 lncRNAs were barely expressed with a $\log_2$ (normalized RPKM + 1) of 0, which was normalized RPKM equaled to 0 either.

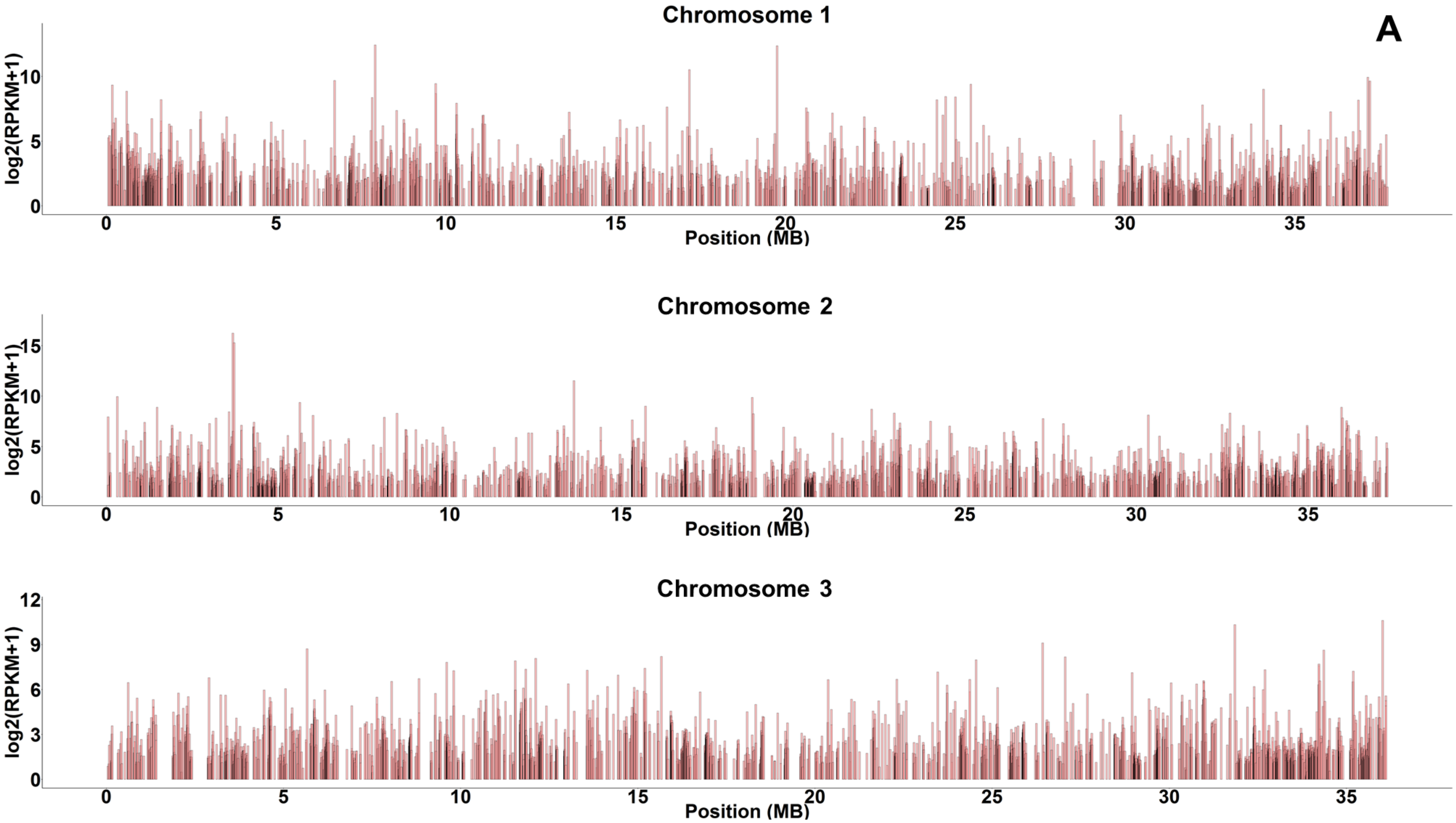

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

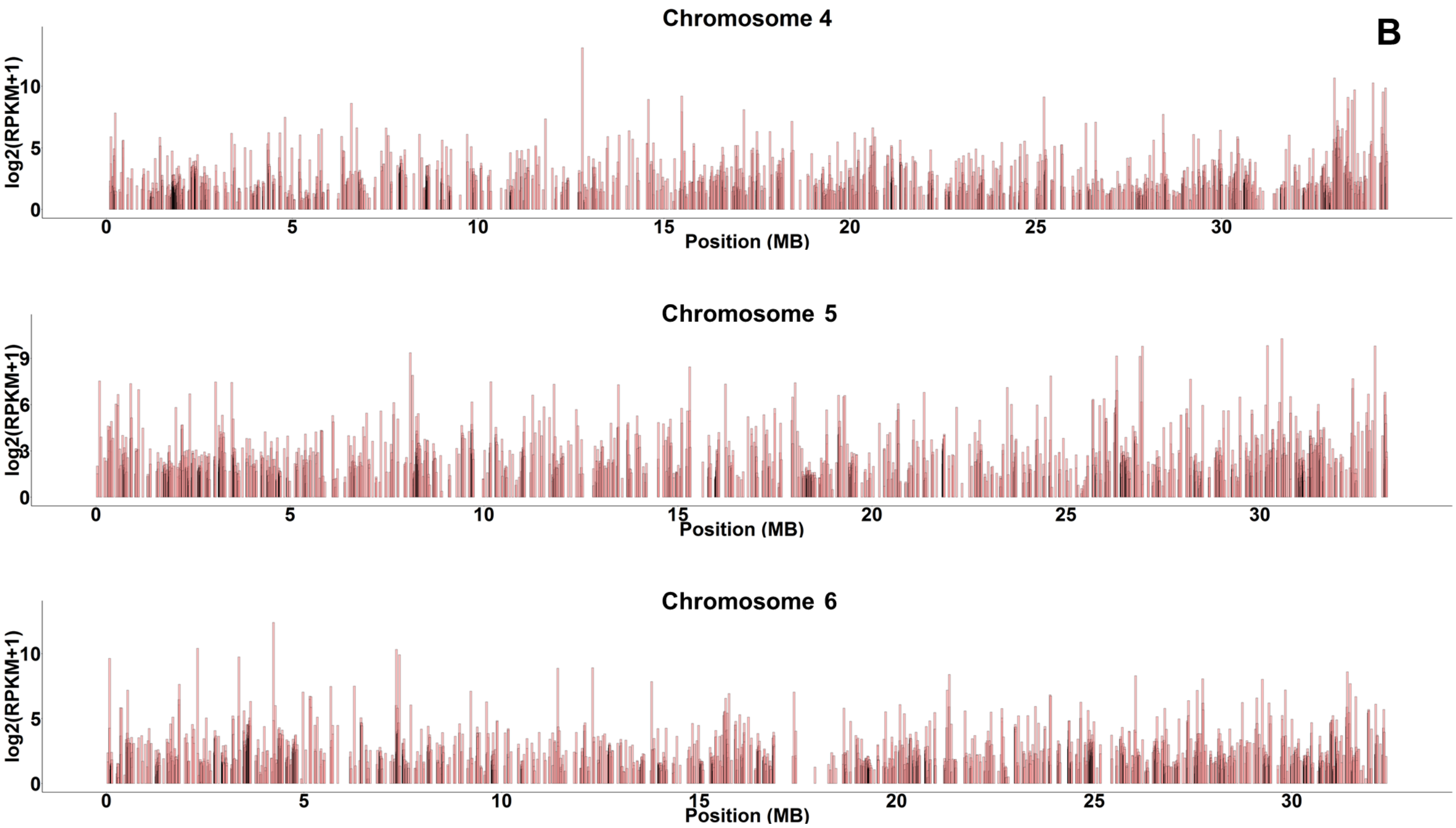

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

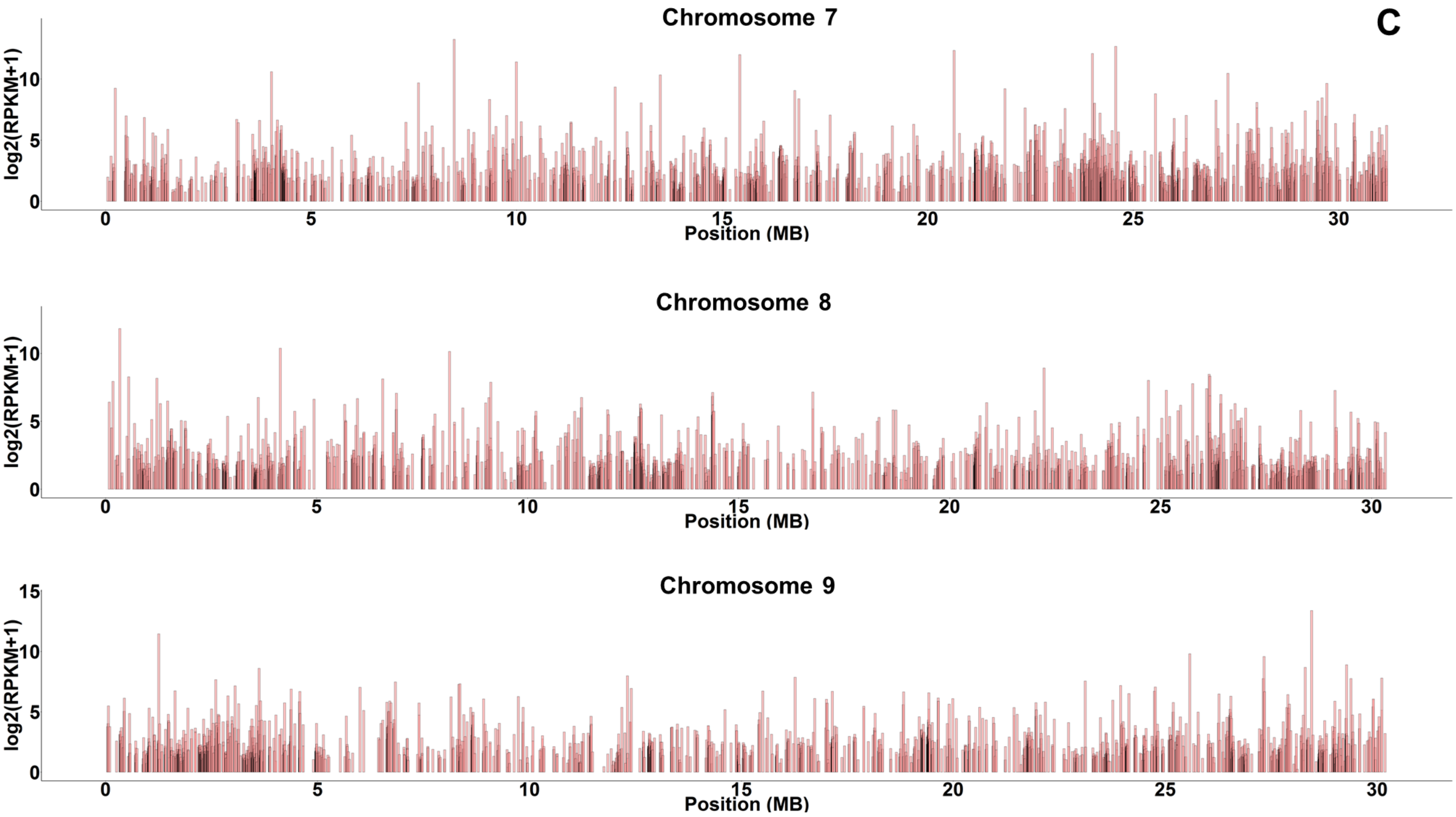

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

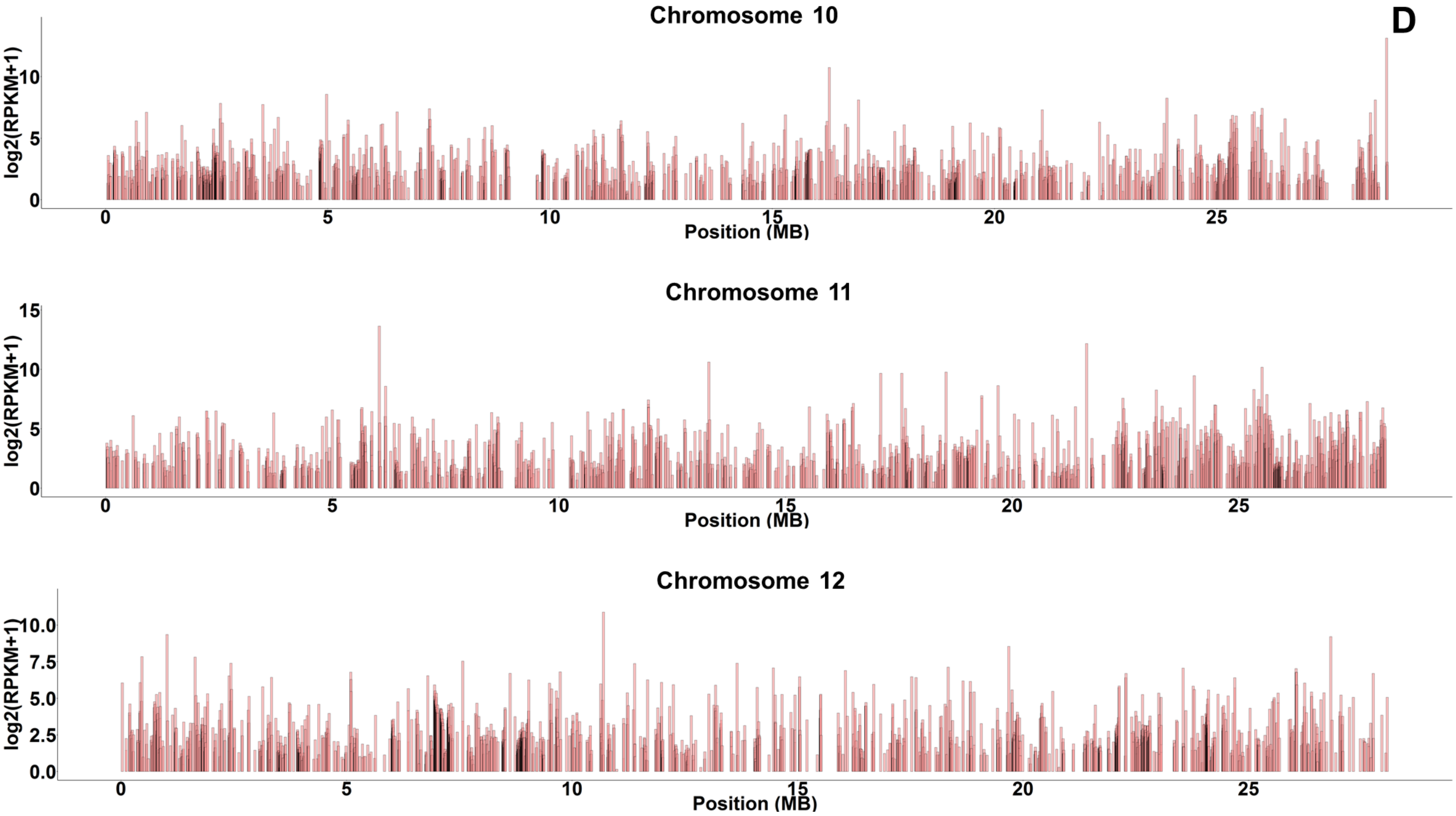

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

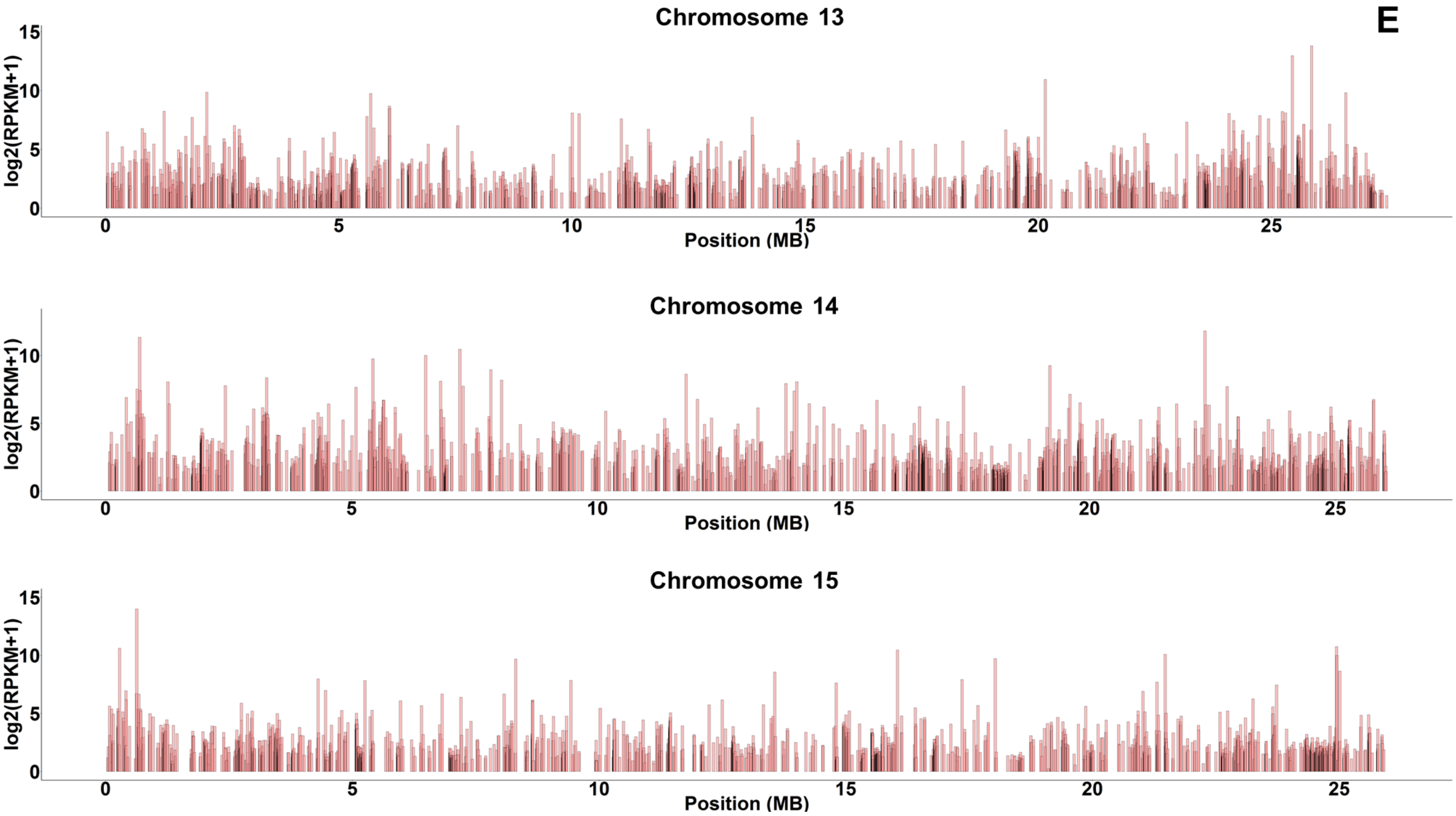

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

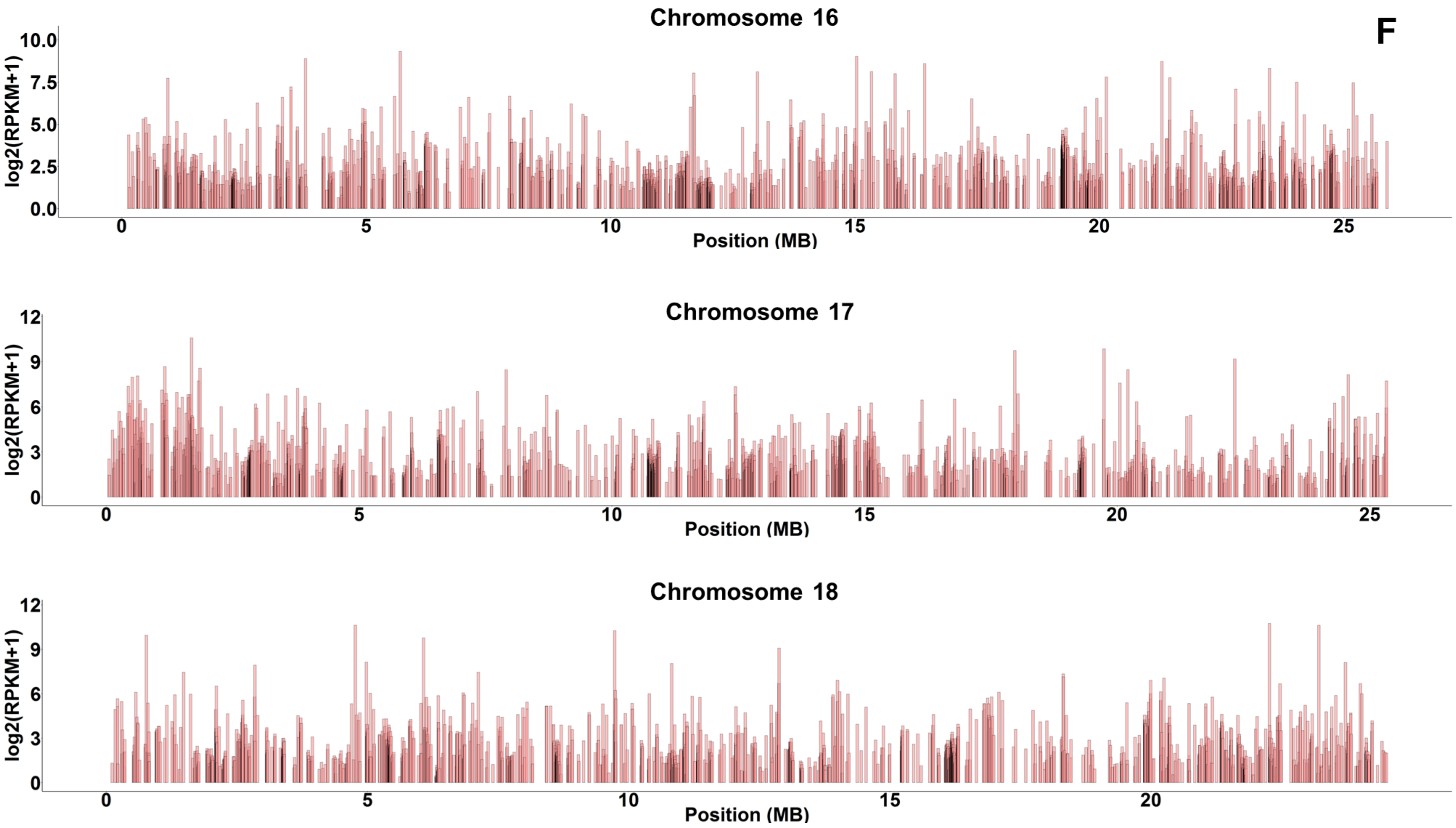

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

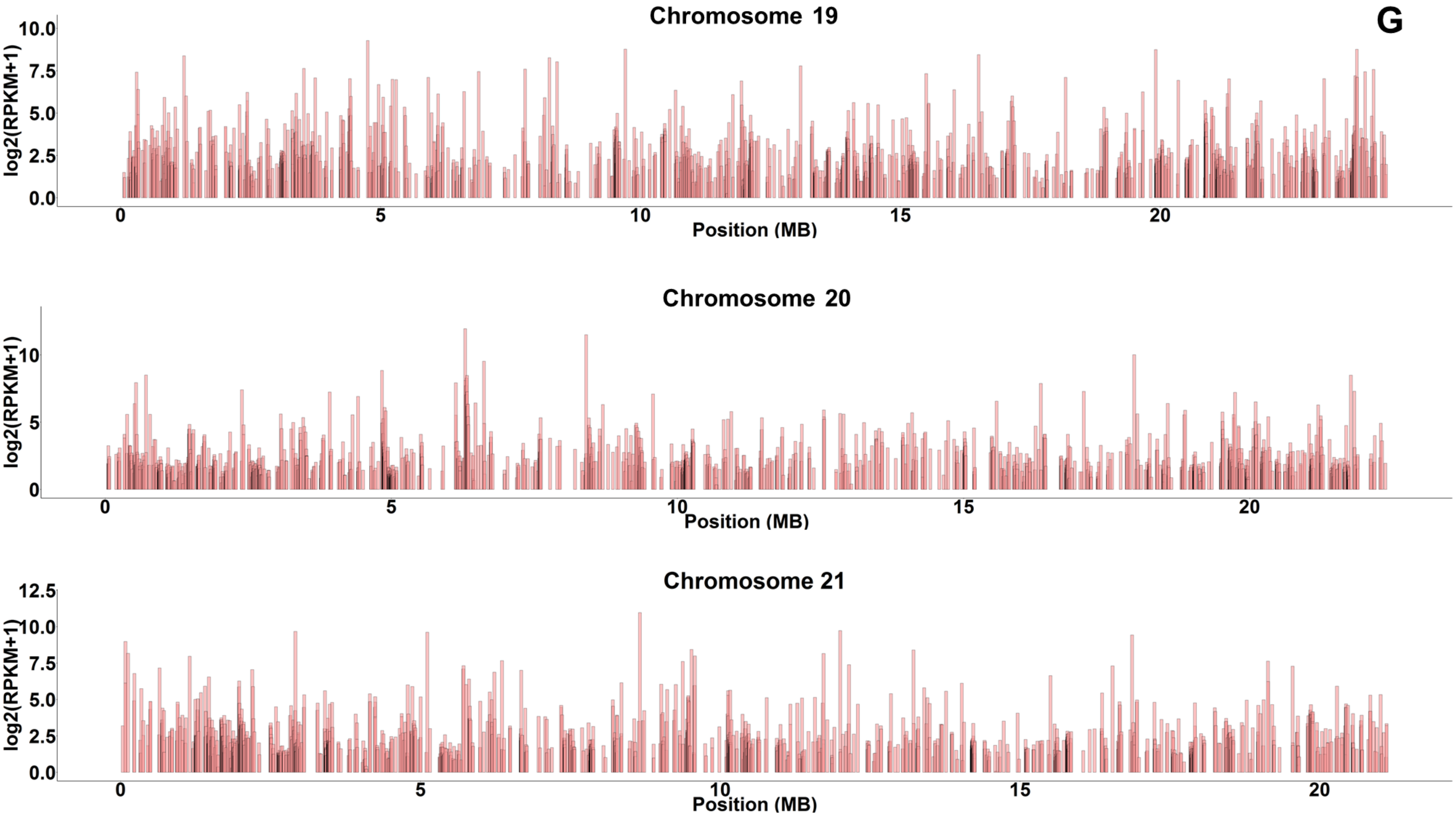

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

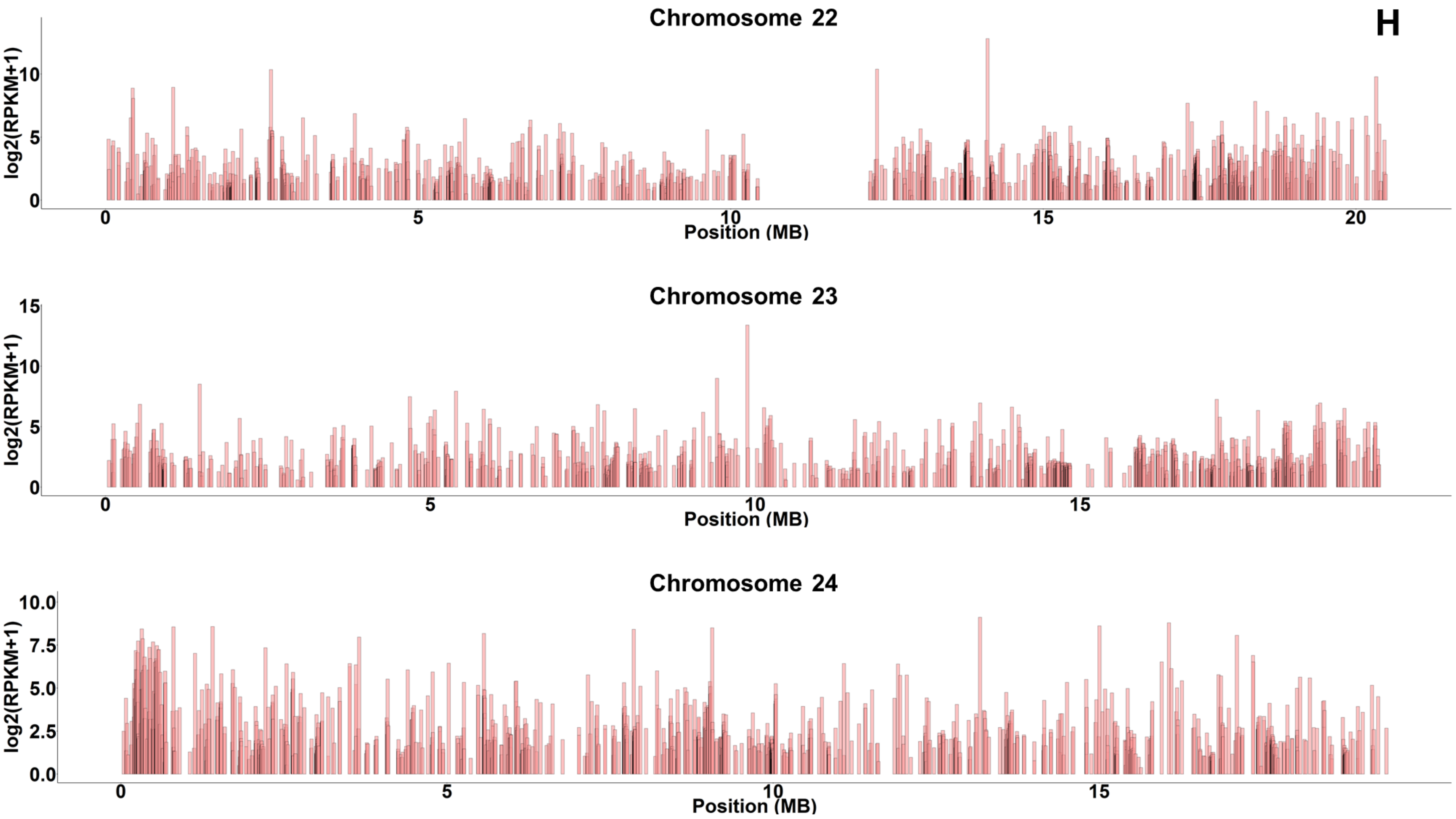

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

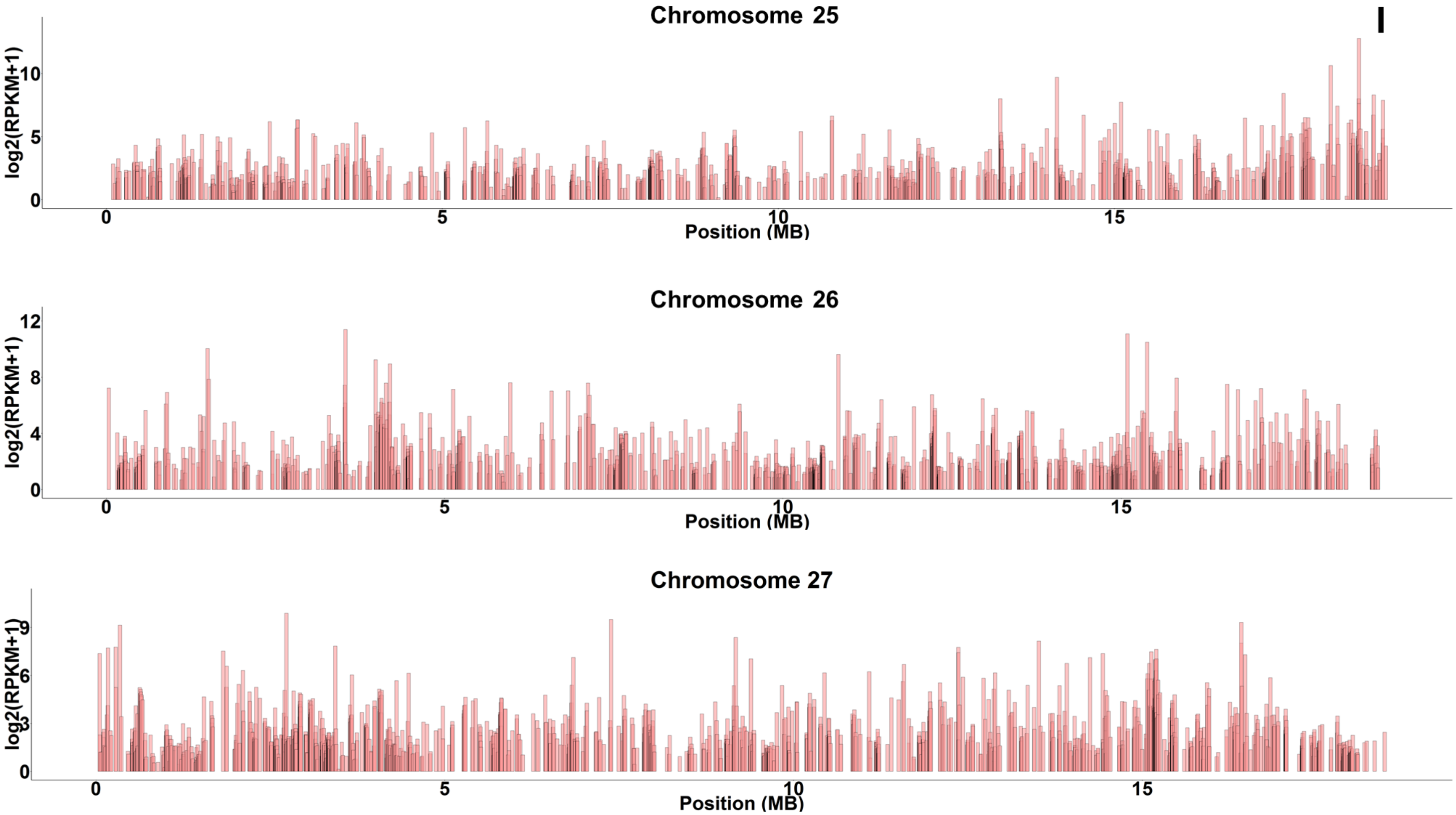

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

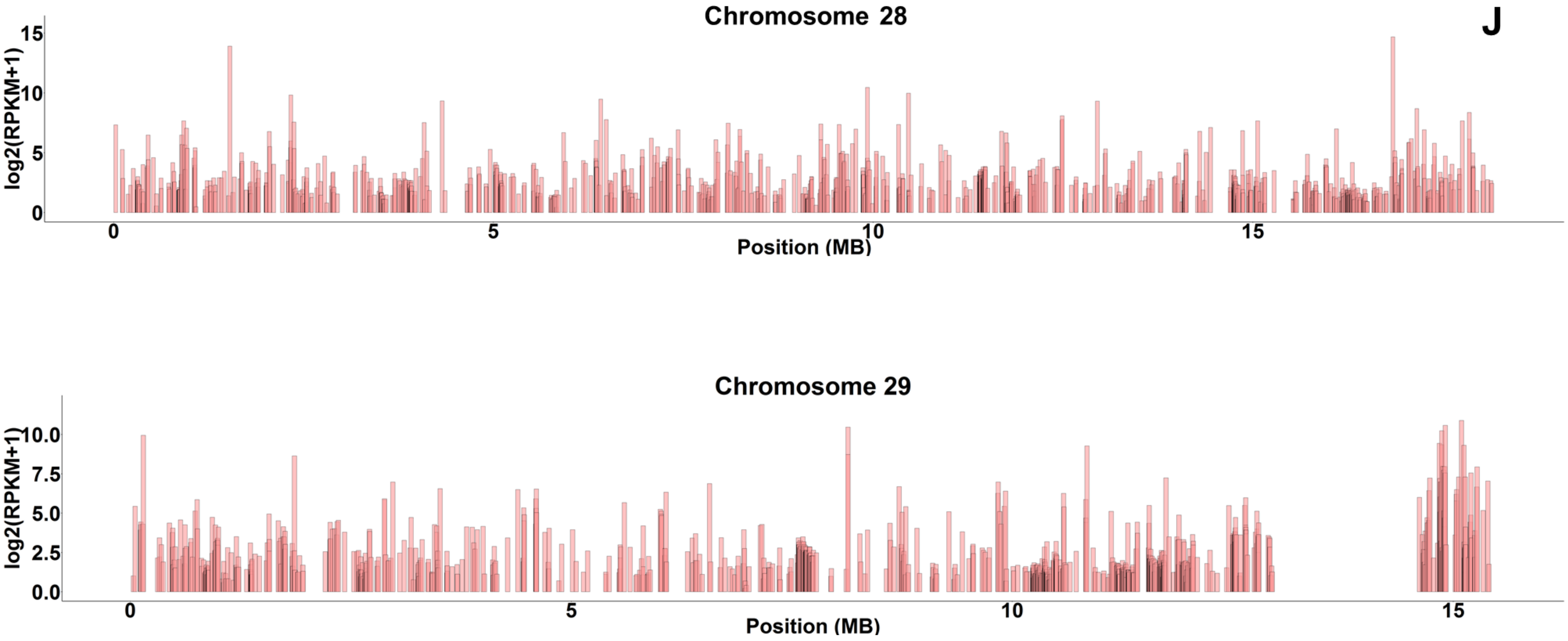

Figure 7. Long non-coding RNAs (lncRNA) expression profiles of channel catfish. X-axis represented the position of the lncRNAs along the chromosomes in mega base pair (MB), Y-axis represented the log transformed expression value $\log_2$ (normalized RPKM + 1). (A) Chromosomes 1-3; (B) Chromosomes 4-6; (C) Chromosomes 7-9; (D) Chromosomes 10-12; (E) Chromosomes 13-15; (F) Chromosomes 16-18; (G) Chromosomes 19-21; (H) Chromosomes 22-24; (I) Chromosomes 25-27; (J) Chromosomes 28 and 29.

Identification of tissue-specific expressed lncRNAs in channel catfish

Tissue-specific expressed lncRNAs were identified as the same criteria as tissue-specific genes, and using the same set of RNA-Seq datasets utilized for identification of tissue-specific expressed genes. By mapping reads from each collected tissue to channel catfish genome, the normalized expression level of each lncRNAs in each tissue was calculated, tissue-specific expressed lncRNAs were only identified if the fold change of one specific tissues against the rest of the tissues in expression level was at least 32 fold with the FDR adjusted p value of less than 0.05. A total of 2,599 lncRNAs were identified as tissue-specific expressed lncRNA based on the above criteria, the number of tissue-specific genes predicted in different tissues were listed in Table 8 and detailed tissue-specific lncRNAs with their fold changes were summarized in Supplemental Table 4. Ovary showed the most tissue-specific expressed lncRNAs (848 lncRNAs), followed by testis (710 lncRNAs) and barbel (605 lncRNAs). Conversely but also similar to tissue-specific gene expression, gill showed the lowest number of tissue-specific expressed lncRNAs, which was only 41 lncRNAs, followed by intestine (89 lncRNAs) and skin (122 lncRNAs). In order to assess the specificity of these tissue-specific expressed lncRNAs, distribution of the differentially expressed fold change of all tissue-specific expressed lncRNAs were visualized in Figure 8. As shown in Figure 8, unlike the pattern in tissue-specific genes, ovary had the most tissue-specific expressed lncRNAs across different range of fold changes, 353 lncRNAs were expressed 32 to 64 fold higher than all other seven tissues, 273 genes were expressed 64 to 128 fold higher than all other seven tissues and 222 genes were expressed over 128 fold higher than all others.

Table 8. Number of tissue-specific lncRNAs predicted in different tissues

Tissues	LncRNA
Barbel	605
Gill	41
Intestine	89
Liver	184
Ovary	848
Skin	122
Testis	710

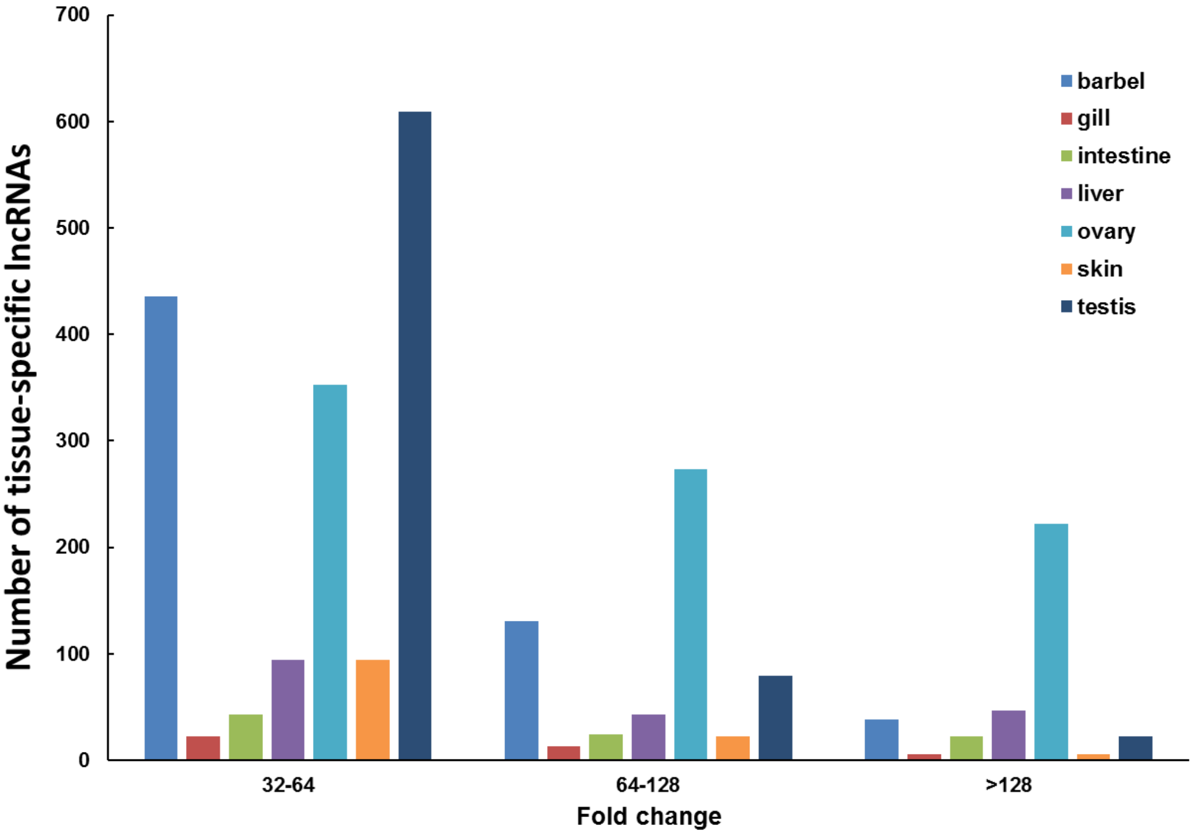

Figure 8. Number of tissue-specific lncRNAs in different scales of fold change. X-axis represented the three different scales of fold change with the FDR corrected p-value less than 0.05, which were 32 to 64 fold, 64 to 128 fold, and over 128 fold. Y-axis represented the number of tissue-specific lncRNAs that fell in different fold change scales.

Identification of induced expression of lncRNAs in channel catfish

Differentially induced lncRNAs after different disease and stress treatments were determined by comparing their expression levels in normalized RPKM between treatment groups and control groups. The differentially induced lncRNAs were only identified if two-fold change expression were observed in at least one treatment and FDR corrected p-value < 0.05, the number of differentially induced expressed lncRNAs from different treatments were summarized in Table 9 (details in Supplemental Table 3). A total of 748 lncRNAs were differentially expressed across all used challenged RNA-Seq datasets, as shown in Table 9, shared the similar pattern in differentially expressed genes, the most differentially expressed lncRNAs were observed in columnaris challenged RNA-Seq dataset (0h, 1h, 2h, and 8h）, while the other columnaris challenged RNA-Seq dataset (0h, 4h, 24h, and 48h) had the least. The differentially expressed lncRNAs and genes were then compared within each stimulus or immune response with their $\log_2$(fold chanlge) value (Figure 9-14). As shown in the comparison figures, the differentially induced lncRNAs were much less than the differentially induced genes in number, the extent of induced expression were not too different, however, after log transformation, the extent of differentially induced lncRNAs were less than those in differentially induced genes.

Table 9. Differentially expressed lncRNAs that were induced in different treatments

Treatment	**LncRNAs**
Columnaris (gill: 0h, 4h, 24h, and 48h)	12
Columnaris (gill: 0h, 1h, 2h, and 8h)	319
Esc (liver)	275
Heat (gill)	144
Heat (liver)	76
Feed deprivation (gill, skin)	25

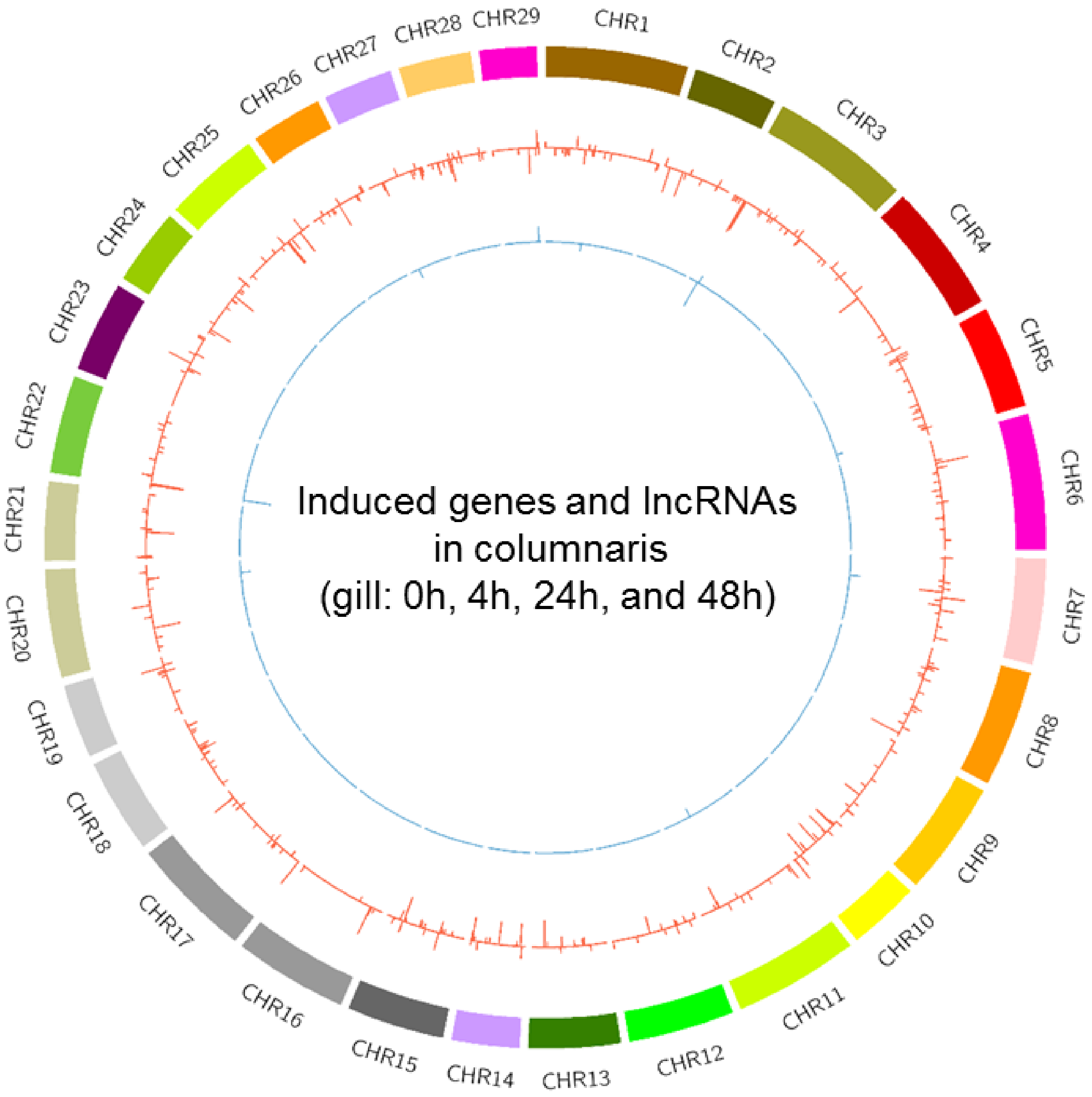

Figure 9. **Differentially induced lncRNAs and protein-coding genes expression profiles comparison after columnaris infection (0h, 4h, 24h, and 48h) along channel catfish genome.** The outer circle represented channel catfish 29 chromosomes, inner red circle represented the highest $\log_2$(fold change) of the protein-coding gene against the control group following the columnaris challenge (0h, 4h, 24h, and 48h), while the inner blue circle represented the highest $\log_2$(fold change) of the lncRNAs against the control group following the same challenge.

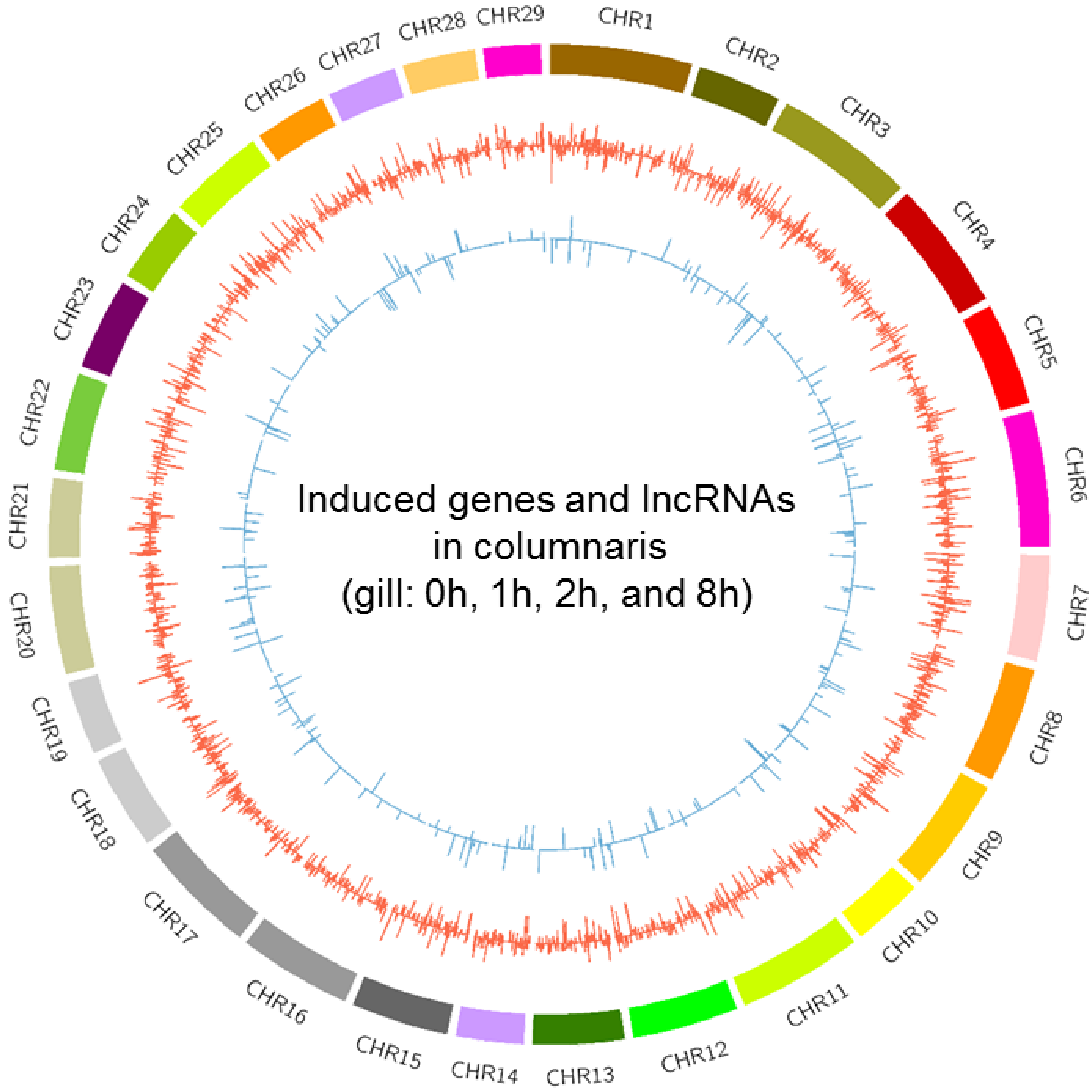

Figure 10. **Differentially induced lncRNAs and protein-coding genes expression profiles comparison after columnaris infection using two sets of channel catfish with different susceptibilities (0h, 1h, 2h, and 8h) along channel catfish genome.** The outer circle represented channel catfish 29 chromosomes, inner red circle represented the highest $\log_2$(fold change) of the protein-coding gene against the control group following the columnaris infection using two sets of channel catfish with different susceptibilities (0h, 1h, 2h, and 8h), while the inner blue circle represented the highest $\log_2$(fold change) of the lncRNAs against the control group following the same challenge.

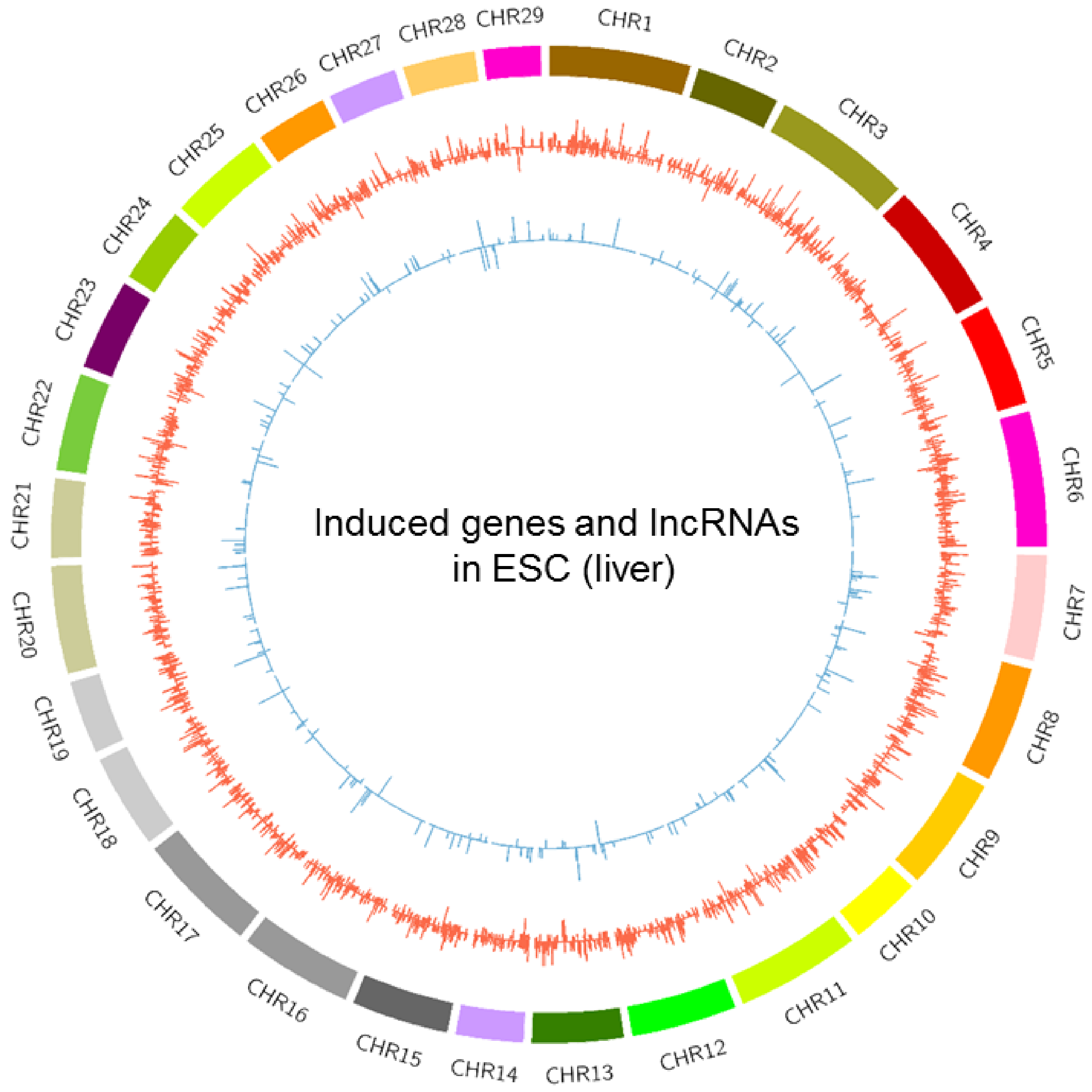

Figure 11. **Differentially induced lncRNAs and protein-coding genes expression profiles comparison after ESC infection along channel catfish genome.** The outer circle represented channel catfish 29 chromosomes, inner red circle represented the highest $\log_2$(fold change) of the protein-coding gene against the control group following the ESC infection, while the inner blue circle represented the highest $\log_2$(fold change) of the lncRNAs against the control group following the same challenge.

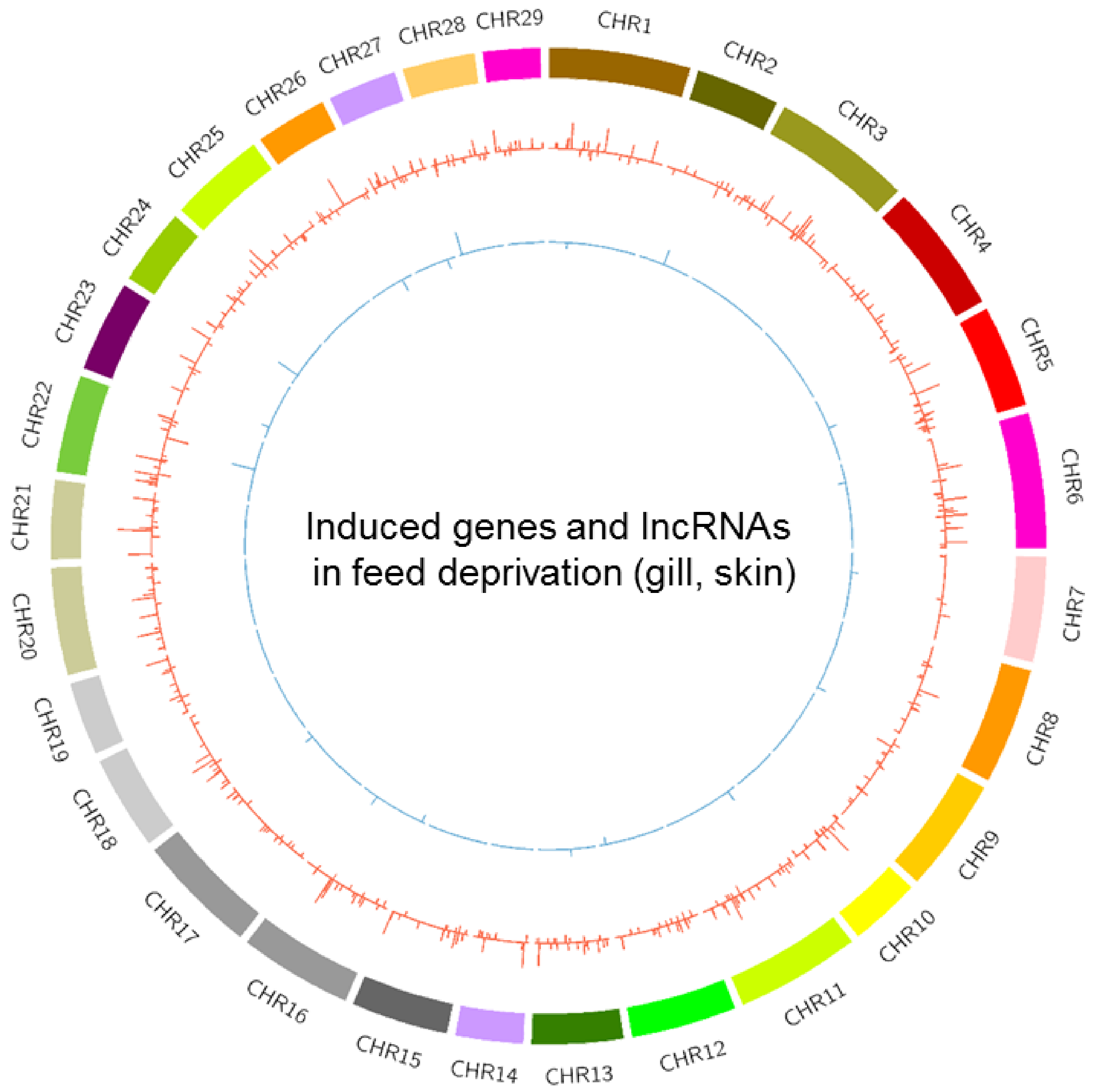

Figure 12. **Differentially induced lncRNAs and protein-coding genes expression profiles comparison after short-term feed deprivation challenge along channel catfish genome.** The outer circle represented channel catfish 29 chromosomes, inner red circle represented the highest $\log_2$(fold change) of the protein-coding gene against the feed group following the short-term feed deprivation challenge, while the inner blue circle represented the highest $\log_2$(fold change) of the lncRNAs against the control group following the same challenge.

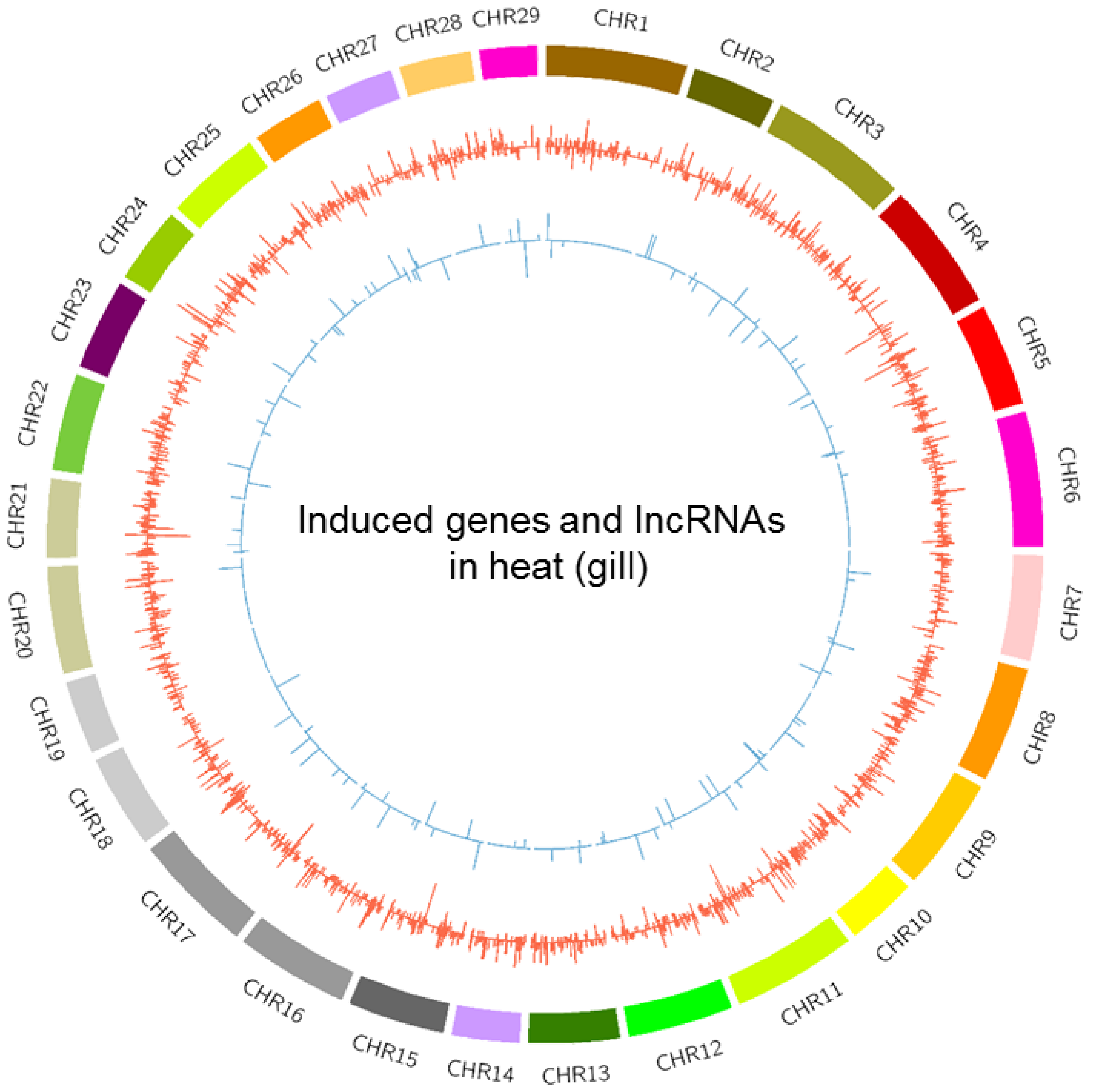

Figure 13. **Differentially induced lncRNAs and protein-coding genes expression profiles comparison after heat stress challenge in gill along channel catfish genome.** The outer circle represented channel catfish 29 chromosomes, inner red circle represented the highest $\log_2$(fold change) of the protein-coding gene against the control group following the heat stress challenge in gill, while the inner blue circle represented the highest $\log_2$(fold change) of the lncRNAs against the control group following the same challenge same tissue.

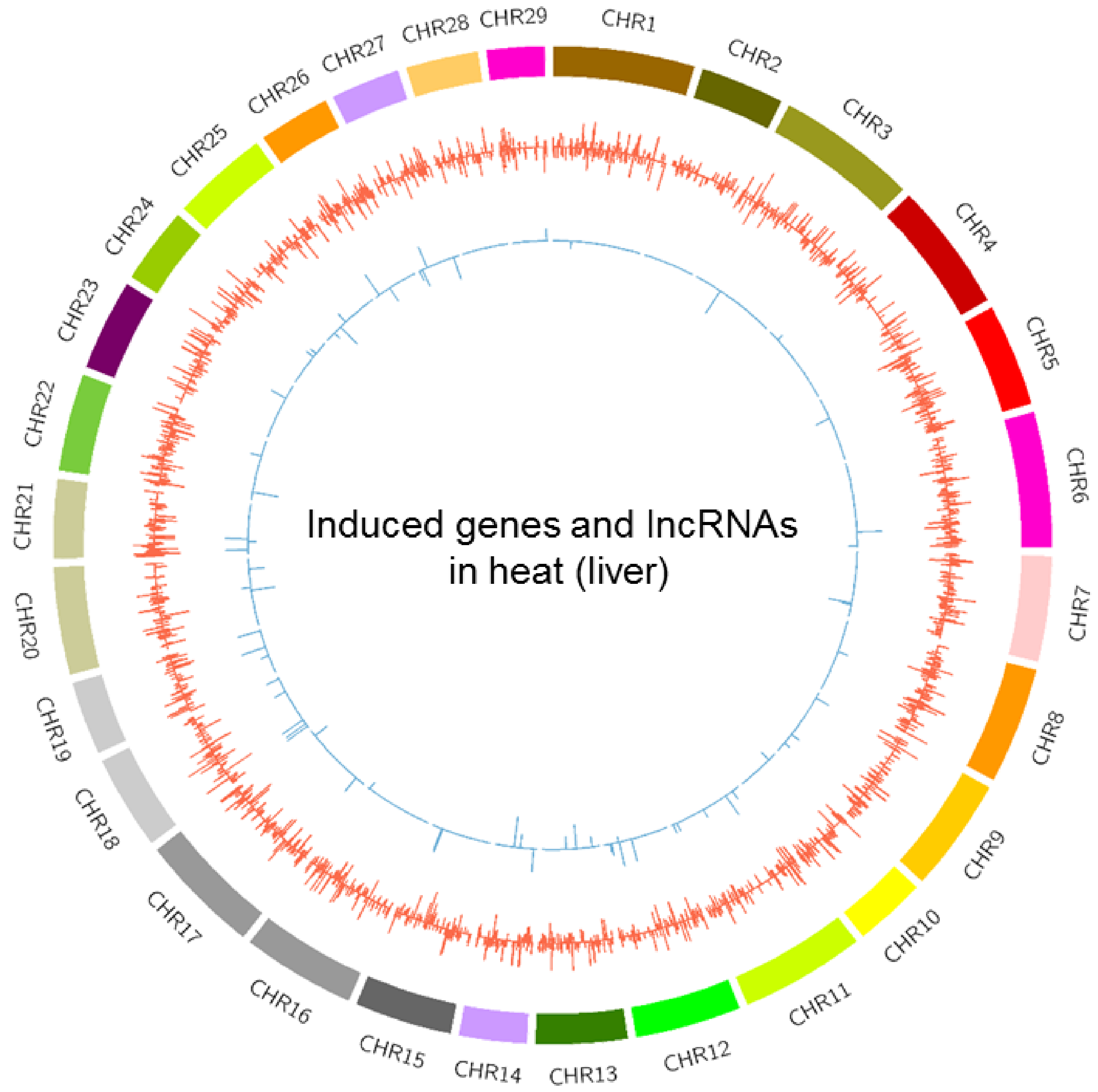

Figure 14. **Differentially induced lncRNAs and protein-coding genes expression profiles comparison after heat stress challenge in liver along channel catfish genome.** The outer circle represented channel catfish 29 chromosomes, inner red circle represented the highest $\log_2$(fold change) of the protein-coding gene against the control group following the heat stress challenge in liver, while the inner blue circle represented the highest $\log_2$(fold change) of the lncRNAs against the control group following the same challenge same tissue.

Identification of correlated co-expressed lncRNAs and genes

As shown above, many differentially induced genes and lncRNAs were observed to the same expression pattern, in order to assess if there were correlation between them, correlation analysis was conducted to construct correlation matrix using normalized RPKM of all the significantly differential expressed genes and lncRNAs in each time point, tissue and treatment group. A total of 1,754 genes and 253 lncRNAs were significantly correlated co-expressed with the correlation coefficient treater than 0.9 or smaller than -0.9 with p-value smaller than 0.05 (Supplemental Table 6). All of the co-expressed lncRNAs and genes were positive correlated. Among these co-expressed lncRNAs and genes, as shown in Table 10, the most co-expressed sets were in ESC infection, which was 479 genes co-expressed with 112 lncRNAs. During columnaris challenge (0h, 4h, 24h, and 48h) and feed deprivation challenge, much less co-expressed genes and lncRNAs were found.

Table 10. Correlated co-expressed lncRNAs and genes that were induced after different treatments

Treatment	Co-expressed gene	Co-expressed lncRNAs
Columnaris (gill: 0h, 4h, 24h, and 48h)	49	3
Columnaris (gill: 0h, 1h, 2h, and 8h)	260	75
ESC (liver)	479	112
Heat (gill)	222	42
Heat (liver)	258	35
Feed deprivation (gill, skin)	8	5
Total	1,754	253

Identification of induced co-localized lncRNAs and genes

Despite of the statistical correlation between the genes and lncRNAs, spatial correlations of the co-localization between genes and lncRNAs were also analyzed. Co-localized expression were only identified if the paired lncRNA and gene sets to be differentially expressed in at least one same treatment, time point and tissue and also the paired lncRNA genes sets were located next to each other based on their position along channel catfish genome. A total of 260 pairs of differentially expressed lncRNAs and genes were observed to be co-localized (Supplemental Table 7). As shown in the Table 11, most co-localized lncRNAs and genes sets were also observed in ESC challenge and least co-localized sets were observed in columnaris challenge (0h, 4h, 24h, and 48h), similar as the result from correlated co-expressed lncRNAs and genes. Among these 260 pairs of co-localization expressed genes and lncRNAs, 113 lncRNAs located on the 5’ side of their paired genes, 101 lncRNAs located on the 3’ side, while three pairs of were the combination of two genes and one lncRNA, and other 43 pairs were the combination of two lncRNAs and one gene. Among all of the co-localized pairs, 8 pairs were observed co-localized differentially expressed in at least two bacterial infections, including nucleotide-binding oligomerization domain protein 3, interferon-induced protein 44-like, krueppel-like factor 7, interferon regulatory factor 1, zinc finger HIT domain-containing protein 1, interferon-induced protein 44, urokinase plasminogen activator surface receptor, and butyrophilin subfamily 1 member A1-like.

Table 11. Induced co-localized lncRNAs and genes that were induced after different treatments

Treatment	**Pair**
Columnaris (gill: 0h, 4h, 24h and 48h)	2
Columnaris (gill: 0h, 1h, 2h, and 8h)	92
Esc (liver)	122
Heat (gill)	37
Heat (liver)	27
Feed deprivation (gill, skin)	3
Total	260

Identification of co-induced co-localized lncRNAs and genes

The final pool of differentially induced co-localized lncRNAs and genes sets were constructed using the pool of correlated lncRNAs and genes and co-localized lncRNAs and genes sets. A total of 45 sets of lncRNAs and genes sets were identified as co-induced and co-localized (Supplemental Table 8). The two sets of channel catfish gill of different susceptibilities challenged with columnaris disease (0h, 1h, 2h, and 8h) had the most co-induced and co-localized lncRNAs and genes pairs, while the other columnaris challenge (0h, 4h, 24h and 48h) had the least pairs (Table 12).

Table 12. Number of co-induced co-localized lncRNAs and genes pairs in different treatments

Treatment	**Pair**
Columnaris (gill: 0h, 4h, 24h, and 48h)	2
Columnaris (gill: 0h, 1h, 2h, and 8h)	16
Esc (liver)	17
Heat (gill)	6
Heat (liver)	7
Feed deprivation (gill, skin)	2
Total	45

Discussion

The expression of the transcriptome is dynamic and variable with multi-dimensional factors over space (different tissues or organs), time (different stages of growth and development), physiological conditions (e.g., hormonal regulation, reproduction etc.), various environmental conditions (e.g., high temperature, low oxygen level, etc.), as well as interactions of all these factors. In order to fully assess the expression profiles for transcription, first and foremost, a complete and well annotated transcriptome is needed. In this study, taken advantage of the large scales of RNA-Seq experiments that has been utilized in channel catfish, a relatively complete set of transcriptome was constructed with 769,270 *de novo* contigs and 197,161 genome-guided contigs. The assemblies each recovered 25,888 genes and 25,987 genes, which all together represented 27,448 total genes transcribed from genome. This revealed the most well annotated transcriptome that's been carried out in channel catfish compared to any other individual assembled catfish transcriptome, with an average number of annotated genes of 23,082 from nine published channel catfish transcriptome analyses (Li et al., 2012a; Liu et al., 2013a; Liu et al., 2012; Liu et al., 2011; Liu et al., 2013b; Peatman et al., 2013; Sun et al., 2012; Sun et al., 2013; Wang et al., 2013b). The assembled transcriptome was also considered well annotated across many other teleost. Transcriptome analysis during four early development stages were conducted in zebrafish, resulting in 13,610 genes overlapped with zebrafish reference gene models defined in the UCSC RefSeq Genes track (Vesterlund et al., 2011). Another parallel transcriptome analysis in zebrafish larvae aligned to 25,255 distinct loci on the Zv9 genome (Palmblad et al., 2013). While transcriptome analysis during zebrafish optic vesicle morphogenesis revealed a total of 31,731 genes were identified by combining transcripts assembled using Cufflinks and public databases RefSeq, Ensembl and GenBank (Yin et al., 2014). Besides the well-studied zebrafish, transcriptome analysis conducted in common carp revealed a total of 19,165 unique proteins from their assembled contigs (Ji et al., 2012). The characterization of the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome revealed 24,624 transcript scaffolds with 4,030 gene ontology definitions (Fox et al., 2014). The reference transcriptome is especially essential since with the development of the sequencing technology, sequenced read length is becoming longer and longer, however, longer reads are more likely to span multiple exons, therefore the mapping of long junction reads is becoming more and more challenging without the assistance of reference transcriptome (Zhao, 2014).

With the development of advanced technologies such as high solution tiling arrays, RNA-Seq, and large-scale chromatin immunoprecipitation experiments (ChIP-chip), it became possible and convenient to assess the complexity of the genome expression. More and more evidence had proved that genome is pervasively transcribed through this global profiling of transcription. Many organisms had reported its heavy transcription including human (Birney et al., 2007), yeast (David et al., 2006; Dutrow et al., 2008), plants (Li et al., 2006), *Drosophila* (Stolc et al., 2004) and some other mammals (Berretta and Morillon, 2009). However, pervasive transcription has not been reported in teleost. In the present study, approximately six billion of Illumina sequenced reads were used to map to channel catfish genome, resulted in covering 79.7% of whole channel catfish genome length, indicating that channel catfish was also pervasively transcribed. However, the extent of pervasive transcription in channel catfish was smaller than that in human (93%) or in yeast (85%), two possible reasons were proposed that the varying extent of transcription might play an important role in evolution, or even though the large number of RNA-Seq datasets were utilized, the RNA libraries were constructed using same method from same company which could affect the ability to capture the full sets of transcripts. In addition, the RNA-Seq experiments that conducted in channel catfish were all non-strand specific, therefore estimation of pervasive transcription might be overestimated, and transcription on one strand of the genome could decrease to a minimum of 39.9%. However, antisense transcription (transcription from the opposite strand to a protein-coding or sense strand) was proposed in mammalians (Katayama et al., 2005). Katayama et al. reported that a large proportion of the genome can produce transcripts from both strands, the antisense transcripts played an important role in gene regulation involving degradation of the corresponding sense transcripts as well as gene silencing at the chromatin level (Katayama et al., 2005). In addition, yeast *Saccharomyces cerevisiae* was also reported to observe transcripts that overlap known genes in antisense orientation by quantifying RNA expression on both strands using a high-density oligonucleotide tiling array (David et al., 2006). As lncRNAs are estimated to qualitatively represent about 98% of expressed transcripts in human cells (Morris and Vogt, 2010), if the antisense transcripts are also widespread in channel catfish, the pervasive transcription of channel catfish could be ranged between 39.9% to 79.7%. However, this hypotheses need further validation using strand-specific RNA-Seq data.

Tissue-specific expression plays a fundamental role in maintaining specificity and determining complexity of an organism. Studies on tissue-specific expression pattern could help reveal the molecular mechanisms underlying tissue development, gene function, and transcriptional regulation of biological processes (Song et al., 2013). Here in this work, a total of seven tissues were collect to perform tissue-specific expression in channel catfish, resulted in 1,455 tissue-specific genes and 2,599 tissue-specific lncRNAs. During the identification, the threshold for tissue-specific was set to 32 fold, in order to exclude those genes or lncRNAs that were significantly expressed in two or more tissues, and to gather only genes or lncRNAs with high specificity. Similar work has been done in rainbow trout (Salem et al., 2015) with a cutoff of 8 fold, however, a single double haploid rainbow trout was used to identify tissue-specific gene expression, which could lower the extent of specificity. Using the above stringent criteria, liver had the most tissue-specific genes and ovary had the most tissue-specific lncRNAs, while liver had the highest tissue-specific genes and lncRNAs, followed by intestine. Some similar expression patterns of tissue specificity were also observed in other organism. In channel catfish liver, the highest tissue-specific gene was antihemorrhagic factor cHLP-B, a member of fetuin family, which play an important role in inhibition of hemorrhagic activity but also proteolytic activities, and is expressed by the liver (Dietzel et al., 2013). Despite of the uncharacterized genes, another highly tissue-specific gene (fifth highest) was apolipoprotein A-I-1, which is the major protein component of high density lipoprotein (HDL) in plasma, promotes cholesterol efflux from tissues to the liver for excretion, and primarily expressed in liver (Delcuve et al., 1992). Similarly in channel catfish intestine, the highest non-uncharacterized tissue-specific gene was acidic mammalian chitinase, may participate in the defense against nematodes, fungi and other pathogens and play a role in T-helper cell type 2 (Th2) immune response. High level of activity was detected in the stomach and intestine (Boot et al., 2001). The next highest tissue-specific gene was ladderlectin (RTLL), a multimeric serum lectin that binds Sepharose and LPS of *Aeromonas salmonicida* and its isoform 1 is highly expressed in intestine (Russell et al., 2008).

Since it is generally accepted that genome is pervasively transcribed, lncRNAs, as one of the important component of non-coding RNA, are becoming a hotspot. In order to access the overall picture of lncRNAs, the systemically large scale lncRNA identification and characterization projects start to performed in several model species including human (Wapinski and Chang, 2011), mouse (Guttman et al., 2010), chicken (Li et al., 2012b), zebrafish (Pauli et al., 2012) and *Caenorhabditis elegans* (Nam and Bartel, 2012). However, identification of lncRNA is still blank in non-model species, especially in fish. In this study, 36,266 lncRNAs were identified and their expression profiles were determined using multiple RNA-seq datasets along with protein-coding genes, 8,560 genes and 748 lncRNAs were differentially expressed after different stress treatments. Several studies have discovered the correlated co-expression pattern (Li et al., 2012b; Pauli et al., 2012), and co-localization expression pattern between lncRNAs and protein-coding genes (Ponjavic et al., 2009). Similarly, both correlated co-expression and co-localization expression patterns were observed in the present study, notably, 45 set of protein-coding genes and lncRNAs displayed co-induced co-localized expression after different disease infections and stress inductions, suggesting the presence of "run-on" joint transcripts, or certain cooperative mechanism between lncRNAs and genes were involved during their transcription and related to their functions of stimulus and immune responses. LncRNAs have also been reported to correlated to transcription factors therefore regulate the expression of the protein-coding genes (Guttman et al., 2011; Herriges et al., 2014). If this applied to channel catfish, the observed co-induced co-localized expression pattern between lncRNAs and genes sets could be caused by the same or similar transcription factors. Therefore, when the lncRNAs and genes have the same or similar sets of transcription factor binding sites, the expression pattern of these genes and lncRNAs could be induced at similar expression level at same tissue under different treatments.

The large sets of transcripts assembled in this study will provide valuable resources for future functional research, gene family structures, and digital gene expression analysis. The identified set of tissue-specific genes and lncRNAs enabled greater understanding of organismal development, complexity at the system level. The identification of the lncRNAs followed with the initial characterization of expression profiles along with the protein-coding genes provided a starting point for the study of lncRNA biology in catfish, while the established protocal for rapid molecular genetic analyses of lncRNAs could contribute to the future studies of the function and mechanisms of lncRNAs.

Reference

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

Berretta, J., Morillon, A., 2009. Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO reports 10, 973-982.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Boot, R.G., Blommaart, E.F.C., Swart, E., Ghauharali-van der Vlugt, K., Bijl, N., Moe, C., Place, A., Aerts, J.M.F.G., 2001. Identification of a novel acidic mammalian chitinase distinct from chitotriosidase. Journal of Biological Chemistry 276, 6770-6778.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., Steinmetz, L.M., 2006. A high-resolution map of transcription in the yeast genome. Proceedings of the National Academy of Sciences of the United States of America 103, 5320-5325.

Delcuve, G.P., Sun, J.M., Davie, J.R., 1992. Expression of Rainbow-Trout Apolipoprotein-a-I Genes in Liver and Hepatocellular-Carcinoma. J Lipid Res 33, 251-262.

Dietzel, E., Wessling, J., Floehr, J., Schafer, C., Ensslen, S., Denecke, B., Rosing, B., Neulen, J., Veitinger, T., Spehr, M., Tropartz, T., Tolba, R., Renne, T., Egert, A., Schorle, H., Gottenbusch, Y., Hildebrand, A., Yiallouros, I., Stocker, W., Weiskirchen, R., Jahnen-

Dechent, W., 2013. Fetuin-B, a Liver-Derived Plasma Protein Is Essential for Fertilization. Dev Cell 25, 106-112.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.

Dutrow, N., Nix, D.A., Holt, D., Milash, B., Dalley, B., Westbroek, E., Parnell, T.J., Cairns, B.R., 2008. Dynamic transcriptome of *Schizosaccharomyces pombe* shown by RNA-DNA hybrid mapping. Nature genetics 40, 977-986.

Engstrom, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Ratsch, G., Goldman, N., Hubbard, T.J., Harrow, J., Guigo, R., Bertone, P., Consortium, R., 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. Nature methods 10, 1185-1191.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic acids research 39, W29-W37.

Fox, S.E., Christie, M.R., Marine, M., Priest, H.D., Mockler, T.C., Blouin, M.S., 2014. Sequencing and characterization of the anadromous steelhead (*Oncorhynchus mykiss*) transcriptome. Marine genomics 15, 13-15.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.D., Chen, Z.H., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644-U130.

Grillo, G., Turi, A., Licciulli, F., Mignone, F., Liuni, S., Banfi, S., Gennarino, V.A., Horner, D.S., Pavesi, G., Picardi, E., Pesole, G., 2010. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic acids research 38, D75-D80.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J.L., Root, D.E., Lander, E.S., 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477, 295-300.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S., Regev, A., 2010. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28, 503-510.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N., Regev, A., 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols 8, 1494-1512.

Herriges, M.J., Swarr, D.T., Morley, M.P., Rathi, K.S., Peng, T., Stewart, K.M., Morrisey, E.E., 2014. Long noncoding RNAs are spatially correlated with transcription factors and regulate lung development. Genes & development 28, 1363-1379.

Huang, X., Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome Res 9, 868-877.

Jain, P., Krishnan, N.M., Panda, B., 2013. Augmenting transcriptome assembly by combining *de novo* and genome-guided tools. Peerj 1, e133.

Ji, P., Liu, G., Xu, J., Wang, X., Li, J., Zhao, Z., Zhang, X., Zhang, Y., Xu, P., Sun, X., 2012. Characterization of common carp transcriptome: sequencing, de novo assembly, annotation and comparative genomics. PloS one 7, e35152.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K.C., Hallinan, J., Mattick, J., Hume, D.A., Lipovich, L., Batalov, S., Engstrom, P.G., Mizuno, Y., Faghihi, M.A., Sandelin, A., Chalk, A.M., Mottagui-Tabar, S., Liang, Z., Lenhard, B., Wahlestedt, C., 2005. Antisense transcription in the mammalian transcriptome. Science 309, 1564-1566.

Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14, R36.

Li, C., Zhang, Y., Wang, R., Lu, J., Nandi, S., Mohanty, S., Terhune, J., Liu, Z., Peatman, E., 2012a. RNA-seq analysis of mucosal immune responses reveals signatures of intestinal barrier disruption and pathogen entry following *Edwardsiella ictaluri* infection in channel catfish, *Ictalurus punctatus*. Fish & shellfish immunology 32, 816-827.

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

Li, L., Wang, X.F., Stolc, V., Li, X.Y., Zhang, D.F., Su, N., Tongprasit, W., Li, S.G., Cheng, Z.K., Wang, J., Deng, X.W., 2006. Genome-wide transcription analyses in rice using tiling microarrays. Nature genetics 38, 124-129.

Li, T., Wang, S., Wu, R., Zhou, X., Zhu, D., Zhang, Y., 2012b. Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. Genomics 99, 292-298.

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659.

Liu, L.S., Li, C., Su, B.F., Beck, B.H., Peatman, E., 2013a. Short-Term Feed Deprivation Alters Immune Status of Surface Mucosa in Channel Catfish (*Ictalurus punctatus*). PloS one 8, e74581.

Liu, S., Zhang, Y., Zhou, Z., Waldbieser, G., Sun, F., Lu, J., Zhang, J., Jiang, Y., Zhang, H., Wang, X., Rajendran, K.V., Khoo, L., Kucuktas, H., Peatman, E., Liu, Z., 2012. Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. BMC genomics 13, 595.

Liu, S., Zhou, Z., Lu, J., Sun, F., Wang, S., Liu, H., Jiang, Y., Kucuktas, H., Kaltenboeck, L., Peatman, E., Liu, Z., 2011. Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC genomics 12, 53.

Liu, S.K., Wang, X.L., Sun, F.Y., Zhang, J.R., Feng, J.B., Liu, H., Rajendran, K.V., Sun, L.Y., Zhang, Y., Jiang, Y.L., Peatman, E., Kaltenboeck, L., Kucuktas, H., Liu, Z.J., 2013b. 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, 462-476.

McCarthy, D.J., Chen, Y., Smyth, G.K., 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic acids research 40, 4288-4297.

Morris, K.V., Vogt, P.K., 2010. Long antisense non-coding RNAs and their role in transcription and oncogenesis. Cell Cycle 9, 2544-2547.

Nam, J.W., Bartel, D.P., 2012. Long noncoding RNAs in *C. elegans*. Genome Res 22, 2529-2540.

Palmblad, M., Henkel, C.V., Dirks, R.P., Meijer, A.H., Deelder, A.M., Spaink, H.P., 2013. Parallel deep transcriptome and proteome analysis of zebrafish larvae. BMC research notes 6, 428.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Research 22, 577-591.

Peatman, E., Li, C., Peterson, B.C., Straus, D.L., Farmer, B.D., Beck, B.H., 2013. Basal polarization of the mucosal compartment in *Flavobacterium columnare* susceptible and resistant channel catfish (*Ictalurus punctatus*). Molecular immunology 56, 317-327.

Ponjavic, J., Oliver, P.L., Lunter, G., Ponting, C.P., 2009. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS genetics 5, e1000617.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The European molecular biology open software suite. Trends in Genetics 16, 276-277.

Roberts, A., Pimentel, H., Trapnell, C., Pachter, L., 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27, 2325-2329.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y.S., Newsome, R., Chan, S.K., She, R., Varhol, R., Kamoh, B., Prabhu, A.L., Tam, A., Zhao, Y.J., Moore, R.A., Hirst, M., Marra, M.A., Jones, S.J.M., Hoodless, P.A., Birol, I., 2010. *De novo* assembly and analysis of RNA-seq data. Nature methods 7, 909-U962.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

Russell, S., Young, K.M., Smith, M., Hayes, M.A., Lumsden, J.S., 2008. Cloning, binding properties, and tissue localization of rainbow trout (*Oncorhynchus mykiss*) ladderlectin. Fish & shellfish immunology 24, 669-683.

Salem, M., Paneru, B., Al-Tobasei, R., Abdouni, F., Thorgaard, G.H., Rexroad, C.E., Yao, J., 2015. Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout. PloS one 10, e0121778.

Schmieder, R., Edwards, R., 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS one 6, e17288.

Song, Y., Ahn, J., Suh, Y., Davis, M.E., Lee, K., 2013. Identification of Novel Tissue-Specific Genes by Analysis of Microarray Databases: A Human and Mouse Model. PloS one 8, e64483.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M.F., Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P.E., Bussemaker, H.J., White, K.P., 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science 306, 655-660.

Sun, F., Peatman, E., Li, C., Liu, S., Jiang, Y., Zhou, Z., Liu, Z., 2012. Transcriptomic signatures of attachment, NF-kappaB suppression and IFN stimulation in the catfish gill following columnaris bacterial infection. Developmental and comparative immunology 38, 169-180.

Sun, F.Y., Liu, S.K., Gao, X.Y., Jiang, Y.L., Perera, D., Wang, X.L., Li, C., Sun, L.Y., Zhang, J.R., Kaltenboeck, L., Dunham, R., Liu, Z.J., 2013. Male-Biased Genes in Catfish as Revealed by RNA-Seq Analysis of the Testis Transcriptome. PloS one 8.

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511-515.

Vesterlund, L., Jiao, H., Unneberg, P., Hovatta, O., Kere, J., 2011. The zebrafish transcriptome during early development. BMC developmental biology 11, 30.

Wang, L., Park, H.J., Dasari, S., Wang, S.Q., Kocher, J.P., Li, W., 2013a. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic acids research 41, e74-e74.

Wang, R.J., Sun, L.Y., Bao, L.S., Zhang, J.R., Jiang, Y.L., Yao, J., Song, L., Feng, J.B., Liu, S.K., Liu, Z.J., 2013b. 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, 929.

Wapinski, O., Chang, H.Y., 2011. Long noncoding RNAs and human disease. Trends in cell biology 21, 354-361.

Xiao, S.J., Zhang, C., Zou, Q., Ji, Z.L., 2010. TiSGeD: a database for tissue-specific genes. Bioinformatics 26, 1273-1275.

Yin, J., Morrissey, M.E., Shine, L., Kennedy, C., Higgins, D.G., Kennedy, B.N., 2014. Genes and signaling networks regulated during zebrafish optic vesicle morphogenesis. BMC genomics 15, 825.

Zhao, S.R., 2014. Assessment of the Impact of Using a Reference Transcriptome in Mapping Short RNA-Seq Reads. PloS one 9, e101374.