Transcriptome analysis of *Culter alburnus* gonad tissues for discovery of sex-related genes

Jianbo Zheng*, Yongyi Jia*, Shili Liu, Wenping Jiang, Meili Chi, Shun Cheng, Zhimin Gu

Zhejiang Institute of Freshwater Fisheries, *Huzhou 313001, China*

Correspondence should be addressed to Gu Zhi-Min: guzhimin2006@163.com

* Authors contributed equally to this work

Abstract: *Culter alburnus* is an important commercially fish species for freshwater breeding in China, and the females grow faster than the males. However, the molecular genetic mechanism of sex determination in *C.alburnus* is still poorly characterized. Here, we performed *de novo* assembly of a transcriptome from adult fish tissues of different gender using short read sequencing technology (Illumina). Our results showed that a total of 364,650 unigenes using Trinity software were obtained, giving rise to an average of 561.92 bp per read. Among them, 70,215 sequences matched known genes, including 5,892 male-biased unigenes and 942 female-biased unigenes. Many sex-related genes and pathways were identified based on annotation information. These results would provide new insights into the genetic mechanism of *C.alburnus* sex determination and also establish an important foundation for further research on aquaculture breeding.

Key words: *Culter alburnus*, transcriptome sequencing, differential gene expression, sex determination, sex-related pathways

Culter alburnus is one kind of important fish species for freshwater breeding in

China, and the female individual grows faster than the male one. Thus, developing sex control technology to produce all female population can improve breeding yield and increase economic income. Although we succeeded to obtain all female species through gynogenesis and sex reversal, the genetic mechanism of sex determination in *C.alburnus* is still poorly characterized. These remaining problems severely limit sex control breeding technology application in *C.alburnus*.

Recently, with advancements in next-generation sequencing technologies, transcriptome profiling provides an invaluable tool for gene discovery and functional analysis [1]. So far, several fish species that involved in sex determination and differentiation by RNA-seq are reported, including *Oreochromis niloticus* [2], *Xiphophorus maculatus* [3], *Ictalurus punctatus* [4], *Paralichthys olivaceus* [5], *Danio rerio* [6], *Oncorhynchus mykiss* [7], and so on. These transcriptome sequencing data with identification of the sex-differentially expressed genes is particularly helpful to understand and elucidate the gene regulatory mechanism of sex determination and differentiation.

Previous studies on *C.alburnus* have mainly focused on embryonic development, growth and reproduction, little involved in the levels of molecular genetic mechanism. In this study, we performed *de novo* assembly of a transcriptome from adult fish tissues of different gender using short read sequencing technology (Illumina). These assembled, annotated transcriptome sequences and gene expression profiles would offer a rapid approach to identify key candidate regulatory genes underlying the processes of gonad development. Moreover, demonstrating of the regulatory mechanisms associated with sex determination will be essential for accelerating and advancing aquaculture breeding programs on *C.alburnus* in the future.

Materials and Methods

Fish samples preparation and RNA extraction

The fish (three males and three females) used in this study were provided from

the Balidian breeding base of Zhejiang Institute of Freshwater Fisheries (Huzhou, Zhejiang Province). The tissues of testes and ovaries were immediately frozen in liquid nitrogen, and stored at -80℃. Total RNA was extracted by SV Total RNA Isolation System (Promega) and was treated with DnaseI (Ambion, USA) to remove contaminative DNA.

Library construction and Illumina sequencing

cDNA library was generated by the SMART cDNA library construction kit (Clontech, USA) according to the manufacturer's protocol. Firstly, mRNA was sheared with fragmentation buffer to produce short bands (200 bp), from which the first-strand DNA was synthesized using random primers and reverse transcriptase. Subsequently, the synthesized cDNA was subjected to end-repair by adding End Repair Mix, 'A' base addition and ligated to adapters according to Illumina's library construction protocol. Finally, the completed libraries were sequenced on the Illumina HiSeq 2000 platform [8, 9].

Sequence data analysis and assembly

The raw reads were initially pre-processed by removing adaptor sequences, low quality sequences, empty reads, and higher N rate sequences to obtain the high quality clean data using SeqPrep (https://github.com/jstjohn/SeqPrep) and Sickle (https://github.com/najoshi/sickle). The trimmed and size-selected reads were then *de novo* assembled by Trinity program (http://trinityrnaseq.sourceforge.net/) [10].

Gene annotation and function classification

All assembled unigenes were compared against the protein databases, such as non-redundant (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG), Search Tool for the Retrieval of Interacting Genes (String), Swissprot and Pfam (Swissprot), to obtain function annotations using BlastX with a typical cut-off E-value<$1e^{-5}$. GO annotation was performed using Blast2GO (http://www.blast2go.com/b2ghome), Clusters of Orthologous Groups (COG) (http://www.ncbi.nlm.nih.gov/COG/) and

KEGG (http://www.genome.jp/kegg/) were used to predict possible functional classifications and metabolic pathways [9].

Identifcation of sex-related differentially expressed genes

RPKM (Reads Per Kilobase of exon model per Million mapped reads) was directly used to compare the difference of gene expression level between male and female individuals [11]. The formula was as follows:

$$\mathrm{RPKM} = \frac{\text{total exon read}}{\text{mapped reads (Millions)} * \text{exon length (KB)}}$$

FDR<0.05 and $\log_2|FC| \geqq 1$ were set as the calculation criterion for significantly differential expression analysis. Based on searching sex-related keywords and other published strategies were to predict sex differentially expressed genes.

KEGG enrichment analysis of differentially expressed genes

KEGG (Kyoto Encyclopedia of Genes and Genomes) was a public database for revealing high level genomic function [12]. In this study, the KEGG functional enrichment analysis of differential expressed genes was tested by KOBAS software (http://kobas.cbi.pku.edu.cn/home.do) [13,14].

Results

Sequencing and *de novo* transcriptome assembly of *C.alburnus*

In order to identify sex-related genes and clarify the mechanism of sex determination in *C.alburnus*, two cDNA libraries were constructed from testes and ovaries for transcriptome analysis. After eliminating adapter sequences and filtering out the low-quality reads, the Illunima HiSeq sequencing produced 127,931,976 raw reads from the testes and 27,215,890 raw reads from the ovaries, respectively (Table 1). The *de novo* assembly of high quality transcriptomic reads generated 364,650 unigenes using Trinity software and the lengths were distributed as 561.92 bp, 22,423 bp, 224 bp, 640 bp, 266 bp from average, longest, shortest, N50 and N90 length,

respectively (Fig 1 a). In addition, most of these unigenes (90.47%) were distributed in the 200-1000 bp region (Fig 1 b).

Function annotation, classifictiaon and bioinformatical analysis

The unigenes were initially processed by searching the non-redundant protein databases using BlastX and a total of 70,215 sequences (19% of unigenes) matched known genes. Among those distinct sequence, the majority of sequences (30,379) had strong homology with *Danio rerio*, followed by *Oncorhynchus mykiss* (3,190), *Astyanax mexicanus* (2,481), *Mus musculus* (1,445), *Tetraodon nigroviridis* (784), *Stegastes partitus* (767), and other or unknown species, which made up 44.4% of total genes (Figure 2).

Gene Ontology (GO) and Cluster of Orthologous Groups of proteins (COG) were used to predict and classify possible functions of the unigenes. For GO analysis, all annotated unigenes (30,654) were classified into three functional categories, including biological process, cellular component and molecular function (Fig 3). In biological process, genes were divided into 25 classifications and cellular process (62.34%) was the largest subcategory. In cellular component, genes involved in cell (44.5%) and cell part (44.5%) were the most abundant. In molecular function, genes were grouped into 20 classifications and the most represented molecular function were bingding (57.8%). For COG analysis, 20,349 unigenes were annotated and divided into 25 specific categories (Fig 4). The most common category was ‘general function prediction only’ with 3,974 unique sequences, followed by ‘Signal transduction mechanisms’ (2,619), ‘Posttranslational modification, protein turnover, chaperones’ (1,876), ‘Replication, recombination and repair’ (1,239). ‘Cell motility’ (15), ‘Extracellular structures’ (0), ‘Nuclear structure’ (13) were the smallest COG categories.

To systematically understand the biological functions of genes, the unigene metabolic pathway analysis was conducted using the Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation system. According to the KEGG results, We

mapped 29,467 unigenes to 362 KEGG pathways (Fig 5). 'Pathways in cancer' pathway had the largest number of unigenes (1,233), followed by 'PI3K-Akt signaling pathway' (1,022), 'Calcium signaling pathway' (921), 'Neuroactive ligand-receptor interaction' (921), 'MAPK signaling pathway' (912) and 'cAMP signaling pathway' (883).

Sex-related genes idetification from comparison of gonad tissues transcriptomes

Gene differential expression analysis was performed by edgeR software, which generated gene read count for differential expression calculation [15]. Based on calculation criterion of FDR<0.05 and log2|FC|≧1, 38,167 unigenes showied signifcant expression difference between male and female individuals, including 5,892 (15.44%) male-biased unigenes and 942 (2.47%) female-biased unigenes (Fig 6). According to our sequence analysis and other published search strategies [5], sex-related well-documented genes were identifed (Table 2). Obviously, many sex-linked genes, especially those located upstream of gonadal development regulatory network, had much more significant difference trend in expression levels, such as *dmrt1*, *foxl2, cyp19a*, and so on.

KEGG enrichment analysis of sex-differential expression genes

KEGG enrichment analysis was applied in revealing high level genomic function to discover the metabolic processes and signal transduction pathways. Statistic and analysis of sex-differential expression genes involving in different category of metabolic pathways, found that these genes mainly enriched in Pathways in cancer, Calcium signaling pathway, cAMP signaling pathway, PI3K-Akt signaling pathway, Endocytosis, MAPK signaling pathway, Neuroactive ligand-receptor interaction, cGMP-PKG signaling pathway. Among them, several pathways associated with gonadal development and sex maintenance were identified, such as ovarian steroidogenesis, estrogen signaling pathway, progesterone-mediated oocyte maturation, prolactin signaling pathway, oocyte meiosis, TGF-beta signaling pathway, steroid hormone biosynthesis.

Discussion

Previous study with identification sex-linked genetic markers using AFLP analysis (amplified fragment length polymorphism) in our lab showed that *C.alburnus* sex differentiation may be controlled by strict genetic regulation. However, the role of genetic factors in sex determination and differentiation is unclear and little evidence can demonstrate its genetic mechanism. Therefore, the discovery of sex-related genes and biological regulatory pathways would provide important clues for systematic elucidation sex determination mechanism of *C.alburnus*.

In this study, we performed the *de novo* transcriptome sequencing of *C.alburnus* testes and ovaries tissues by using Illumina Hiseg platform [8,16]. A total of 364,650 unigenes using Trinity software were obtained, giving rise to an average of 561.92 bp per read. About 19% of unigenes were found to match known genes using the public databases and the majority of sequences showed the greatest similarity to *Danio rerio*.

There has several reasons for explaining the rest of unigenes without annotations: i) most of these unigenes are short fragments; ii) these unigenes are probably new genes or unique to the species of *C.alburnus*. iii) these genes may be non-coding RNA sequences [17]. Hence, these transcriptome data would provide a useful resource for future genetic or genomic studies on this species.

It is known that the mainly aim of transcriptome sequencing is to identify a large number of candidate genes potentially involved in specific biological processes, such as growth, reproduction, and gonad development. In present study, our results revealed large quantities of sex-biased genes that showed sexual dimorphism between ovary and testis by analysis the gene expression profiles of *C.alburnus* gonads. Moreover, multiple sex-related genes were identifed to be involved in several biological pathways associated with gonadal development and sex maintenance, including ovarian steroidogenesis, estrogen signaling pathway, Gn RH signaling pathway, oocyte meiosis, TGF-beta signaling pathway, steroid hormone biosynthesis,

and Wnt signaling pathway.

Taken together, this was the first attempt to perform RNA-seq technology to identify differentially expressed genes between ovaries and testes on *C.alburnus*. Additionly, a significant number of sex-related biological pathways associated with the unique sequences were found. These results would provide new insights into the genetic mechanism of *C.alburnus* sex determination and also establish an important foundation for further research on aquaculture breeding.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

This work was financially supported by grants from, Natural science foundation for young scientists of Zhejiang province (LQ18C190001), Zhejiang Science and Technology Major Program (2016C02055-1), and Natural Science foundation of Huzhou (2017YZ02).

References

1. Lei T, Yue Z, Ying W, Wang Q, Yuan H, Zhao L, Guo W, You X. Transcriptome profiling and digital gene expression by deep sequencing in early somatic embryogenesis of endangered medicinal eleutherococcus senticosus, maxim. Gene. 2016; 578(1):17-24.

2. Tao W, Yuan J, Zhou L, Sun L, Sun Y, Yang S, Li M, Zeng S, Huang B, Wang D. Characterization of gonadal transcriptomes from Nile Tilapia (*Oreochromis niloticus*) reveals differentially expressed genes. Plos One.

2013; 8(5):e63604.

3. Zhang Z, Wang Y, Wang S, Liu J, Warren W, Mitreva M, Walter RB. Transcriptome analysis of female and male *Xiphophorus maculatus* Jp 163 a. Plos One. 2011; 6(4):e18379.

4. Sun F, Liu S, Gao X, Jiang Y, Perera D, Wang X, Li C, Sun L, Zhang J, Kaltenboeck L, Dunham R, Liu Z. Male-biased genes in catfish as revealed by rna-seq analysis of the testis transcriptome. Plos One. 2013; 8(7):e68452.

5. Fan Z, You F, Wang L, Weng S, Wu Z, Hu J, Zou Y, Tan X, Zhang P. Gonadal transcriptome analysis of male and female olive flounder (*Paralichthys olivaceus*). Biomed Research International. 2014; 2014(2014):291067.

6. Small CM, Carney GE, Mo Q, Vannucci M, Jones AG. A microarray analysis of sex- and gonad-biased gene expression in the zebrafish: evidence for masculinization of the transcriptome. BMC Genomics. 2009; 10(1):579.

7. Baron D, Montfort J, Houlgatte R, Fostier A, Guiguen Y. Androgen induced masculinization in rainbow trout results in a marked dysregulation of early gonadal gene expression profiles. BMC Genomics. 2007; 8(1):357.

8. Erlich Y, Mitra PP, Delabastide M, Mccombie WR, Hannon GJ. Alta-cyclic: a self-optimizing base caller for next-generation sequencing. Nature Methods. 2008; 5(8):679-682.

9. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. Blast plus: architecture and applications. BMC Bioinformatics. 2009; 10.

10. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I. Trinity: reconstructing a full-length transcriptome without a genome from RNA-seq data. Nature Biotechnology. 2011; 29(7):644-652.

11. Li B, Dewey CN. Rsem: accurate transcript quantification from rna-seq data

with or without a reference genome. BMC Bioinformatics. 2011; 12(1):323.

12. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M. Kegg for linking genomes to life and the environment. Nucleic Acids Research. 2008; 36: 480-484.

13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995; 57(1):289-300.

14. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L. Kobas 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Research. 2011; 39:316-322.

15. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–140.

16. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. Nucleic Acids Research. 2009; 38(6):1767-1771.

17. Yano A, Guyomard R, Nicol B, Jouanno E, Quillet E, Klopp C. An immune-related gene evolved into the master sex-determining gene in rainbow trout, *Oncorhynchus mykiss*. Current Biology. 2012; 22(15):1423-1428.

Fig 1. Length distribution of unigenes assembled by Trinity. (a) Sequence length distriution from average, longest, shortest, N50 and N90 length, respectively. (b) Size distribution for transcript sequence.

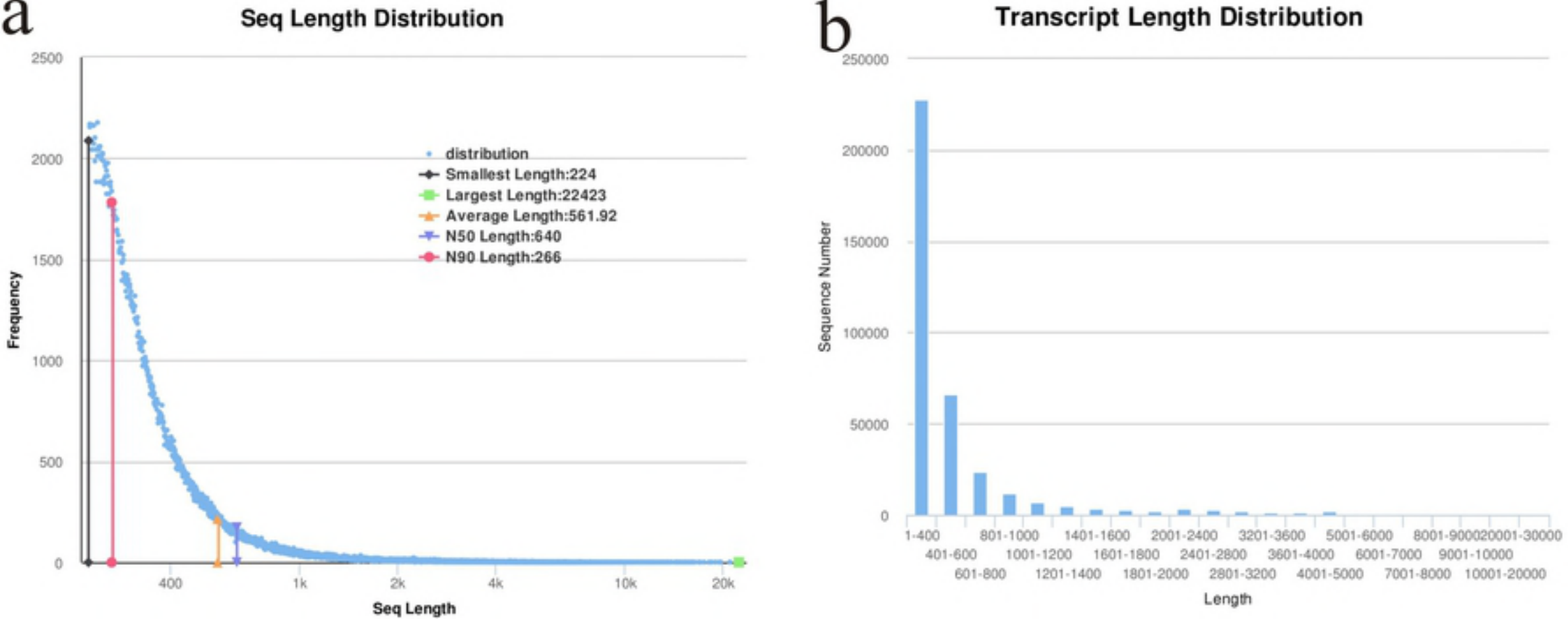
Fig 2. Species distribution that match to the sequences of *C.alburnus*. Each sector represents a species. The number of homologous sequence corresponding each species using BlastX are indicated near the sector.

Fig 3. Gene ontology (GO) category for the transcriptome of *C.alburnus*. Green represents biological process, blue represents cellular component, pink represents molecular function.

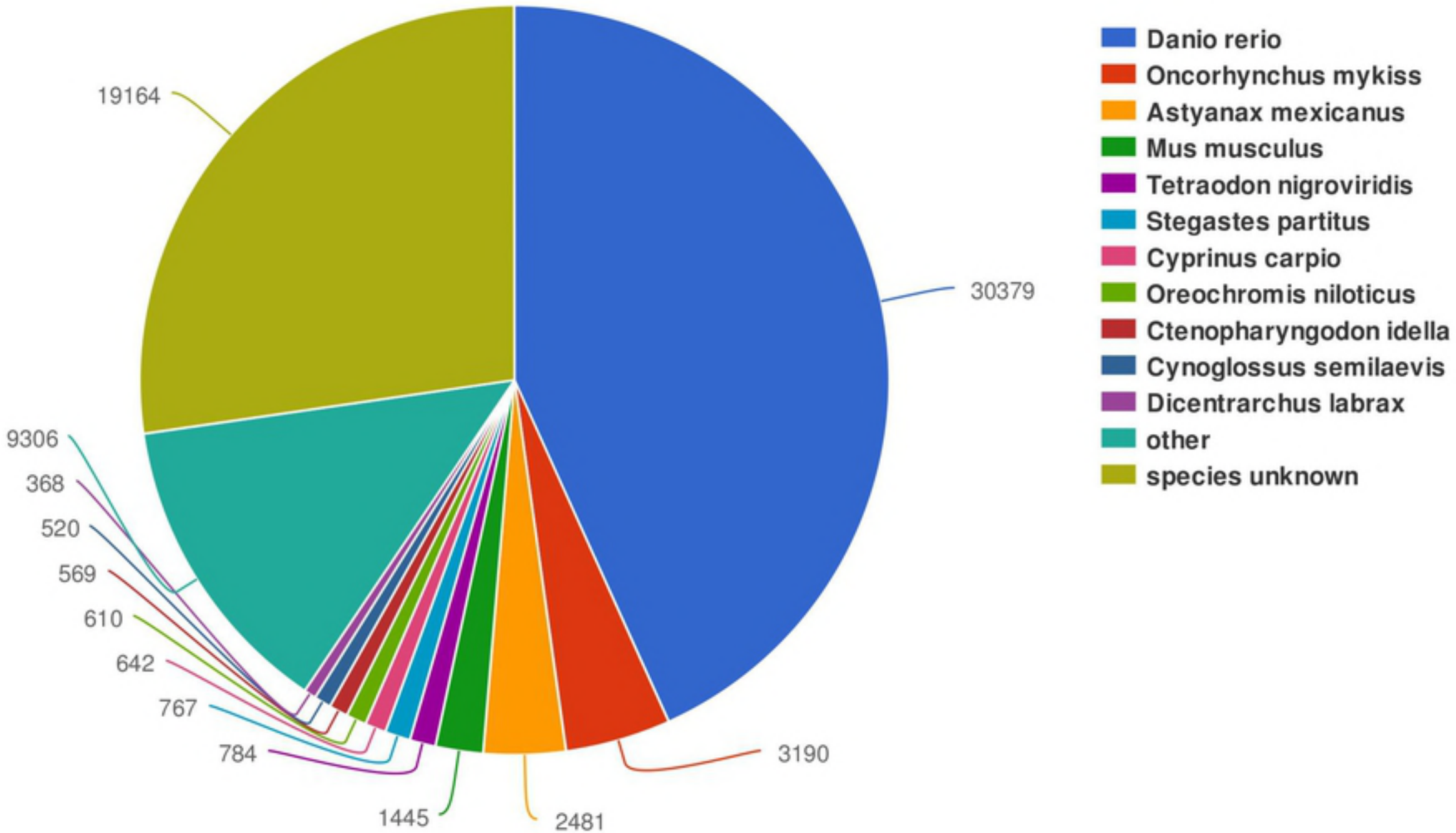
Fig 4. Clusters of orthologous groups (COG) Classification and annotation for *C.alburnus* unigenes. Each color column represents the functional classification of a COG, which indicated by the capital letter A-Z.

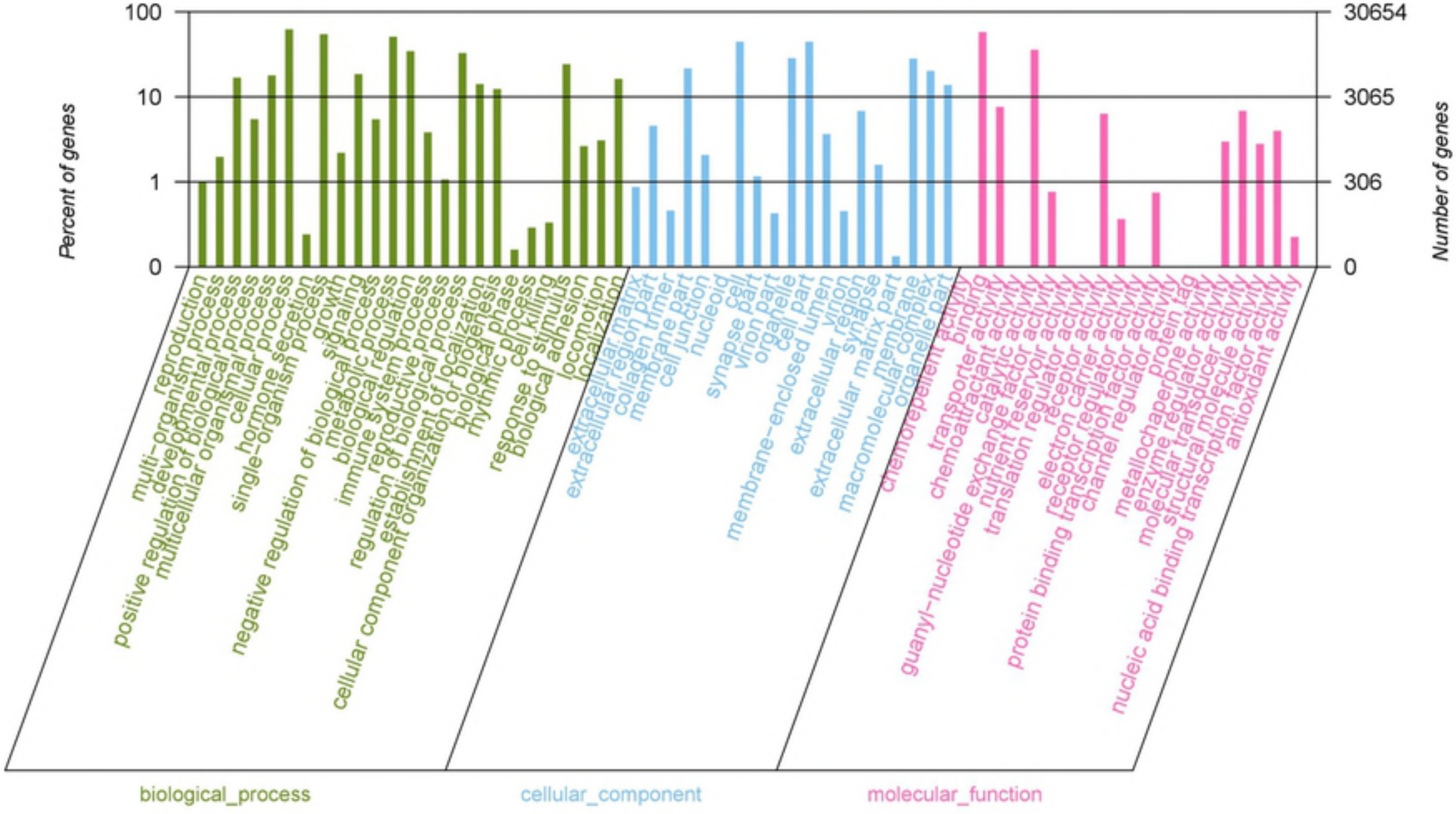
Fig 5. The top twenty pathways with most abundant unigenes.

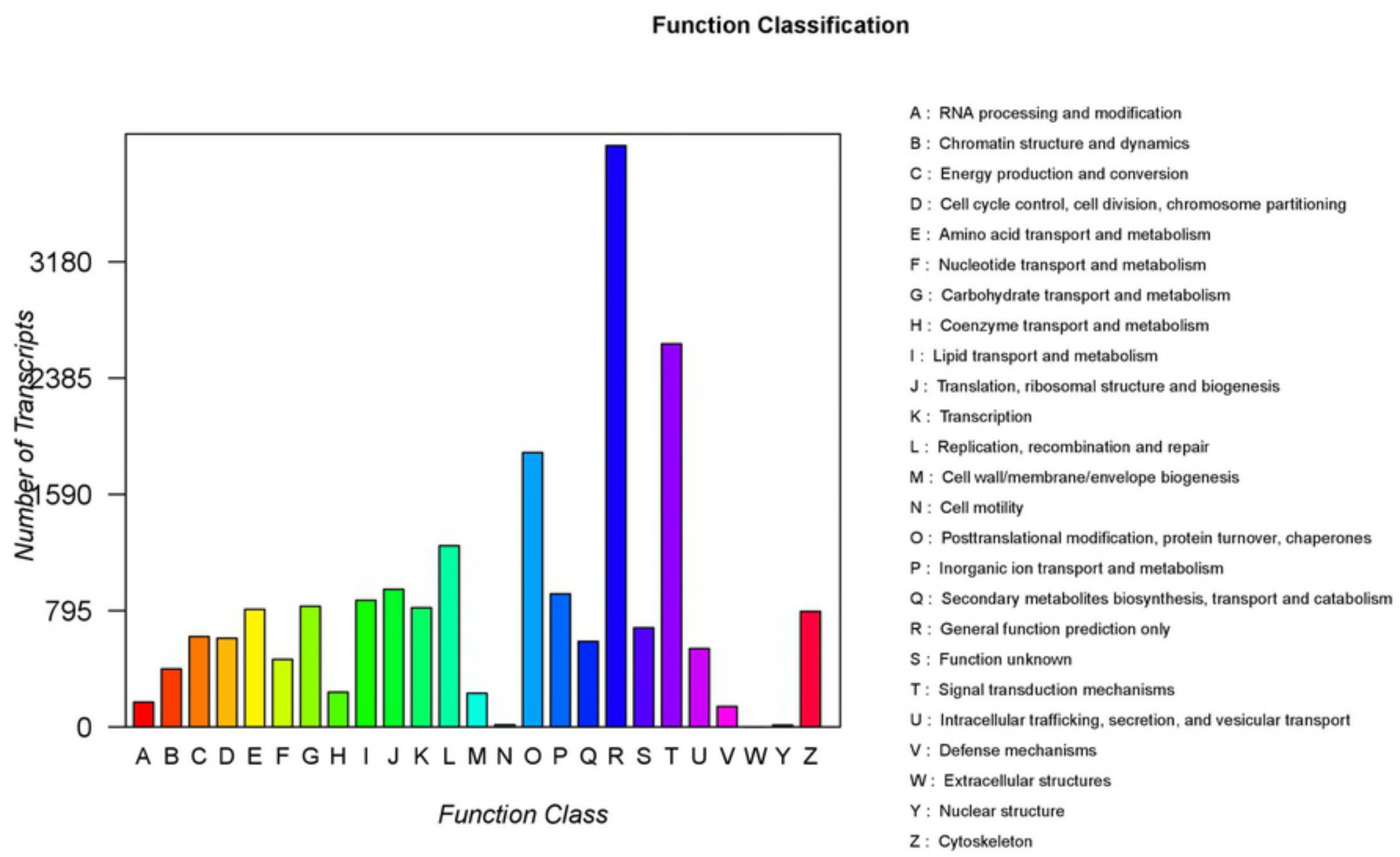
Fig 6. Comparisons of gene expression between females and males of *C.alburnus.* XX represents male individuals, CX represents female individuals.

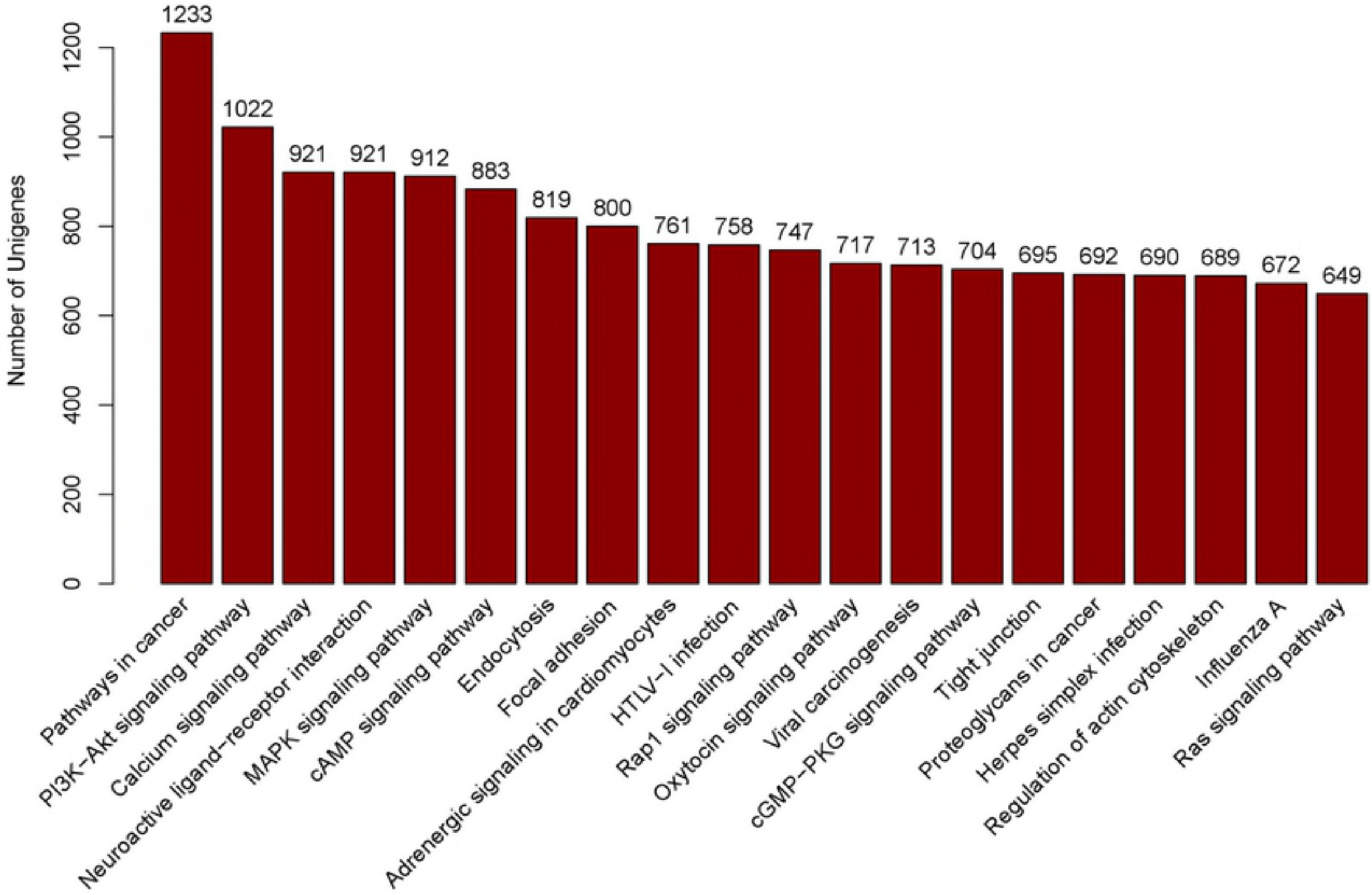

a
Seq Length Distribution
distribution
Smallest Length:224
Largest Length:22423
Average Length:561.92
N50 Length:640
N90 Length:266
Frequency
Seq Length
2500
2000
1500
1000
500
0
400
1k
2k
4k
10k
20k
b
Transcript Length Distribution
Sequence Number
Length
250000
200000
150000
100000
50000
0
1-400
401-600
601-800
801-1000
1001-1200
1201-1400
1401-1600
1601-1800
1801-2000
2001-2400
2401-2800
2801-3200
3201-3600
3601-4000
4001-5000
5001-6000
6001-7000
7001-8000
8001-9000
9001-10000
10001-20000
20001-30000

Species Distribution




Percent of genes
100
10
1
0
30654
3065
306
0
Number of genes
reproduction
multi-organism process
developmental process
positive regulation of biological process
multicellular organismal process
cellular process
single-organism process
hormone secretion
negative regulation of biological process
growth
signaling
metabolic process
biological regulation
immune system process
regulation of biological process
reproductive process
cellular component organization or biogenesis
establishment of localization
biological phase
rhythmic process
cell killing
response to stimulus
biological adhesion
locomotion
localization
extracellular matrix
extracellular region part
collagen trimer
membrane part
cell junction
nucleoid
synapse part
cell
virion part
organelle
membrane-enclosed lumen
cell part
extracellular region
virion
extracellular matrix
synapse
macromolecular complex
membrane part
organelle part
chemorepellent activity
binding
transporter activity
chemoattractant activity
catalytic activity
guanyl-nucleotide exchange factor activity
nutrient reservoir activity
translation regulator activity
protein binding transcription factor activity
receptor activity
electron carrier activity
receptor regulator activity
channel regulator activity
transcription regulator activity
metallochaperone activity
protein tag
enzyme regulator activity
molecular transducer activity
structural molecule activity
nucleic acid binding transcription factor activity
antioxidant activity
biological_process
cellular_component
molecular_function


Function Classification
Number of Transcripts
0
795
1590
2385
3180
A B C D E F G H I J K L M N O P Q R S T U V W Y Z
Function Class
A : RNA processing and modification
B : Chromatin structure and dynamics
C : Energy production and conversion
D : Cell cycle control, cell division, chromosome partitioning
E : Amino acid transport and metabolism
F : Nucleotide transport and metabolism
G : Carbohydrate transport and metabolism
H : Coenzyme transport and metabolism
I : Lipid transport and metabolism
J : Translation, ribosomal structure and biogenesis
K : Transcription
L : Replication, recombination and repair
M : Cell wall/membrane/envelope biogenesis
N : Cell motility
O : Posttranslational modification, protein turnover, chaperones
P : Inorganic ion transport and metabolism
Q : Secondary metabolites biosynthesis, transport and catabolism
R : General function prediction only
S : Function unknown
T : Signal transduction mechanisms
U : Intracellular trafficking, secretion, and vesicular transport
V : Defense mechanisms
W : Extracellular structures
Y : Nuclear structure
Z : Cytoskeleton


Number of Unigenes
0
200
400
600
800
1000
1200
1233
1022
921
921
912
883
819
800
761
758
747
717
713
704
695
692
690
689
672
649
Pathways in cancer
PI3K−Akt signaling pathway
Calcium signaling pathway
Neuroactive ligand−receptor interaction
MAPK signaling pathway
cAMP signaling pathway
Endocytosis
Focal adhesion
Adrenergic signaling in cardiomyocytes
HTLV−I infection
Rap1 signaling pathway
Oxytocin signaling pathway
Viral carcinogenesis
cGMP−PKG signaling pathway
Tight junction
Proteoglycans in cancer
Herpes simplex infection
Regulation of actin cytoskeleton
Influenza A
Ras signaling pathway

CX_vs_XX.scatter

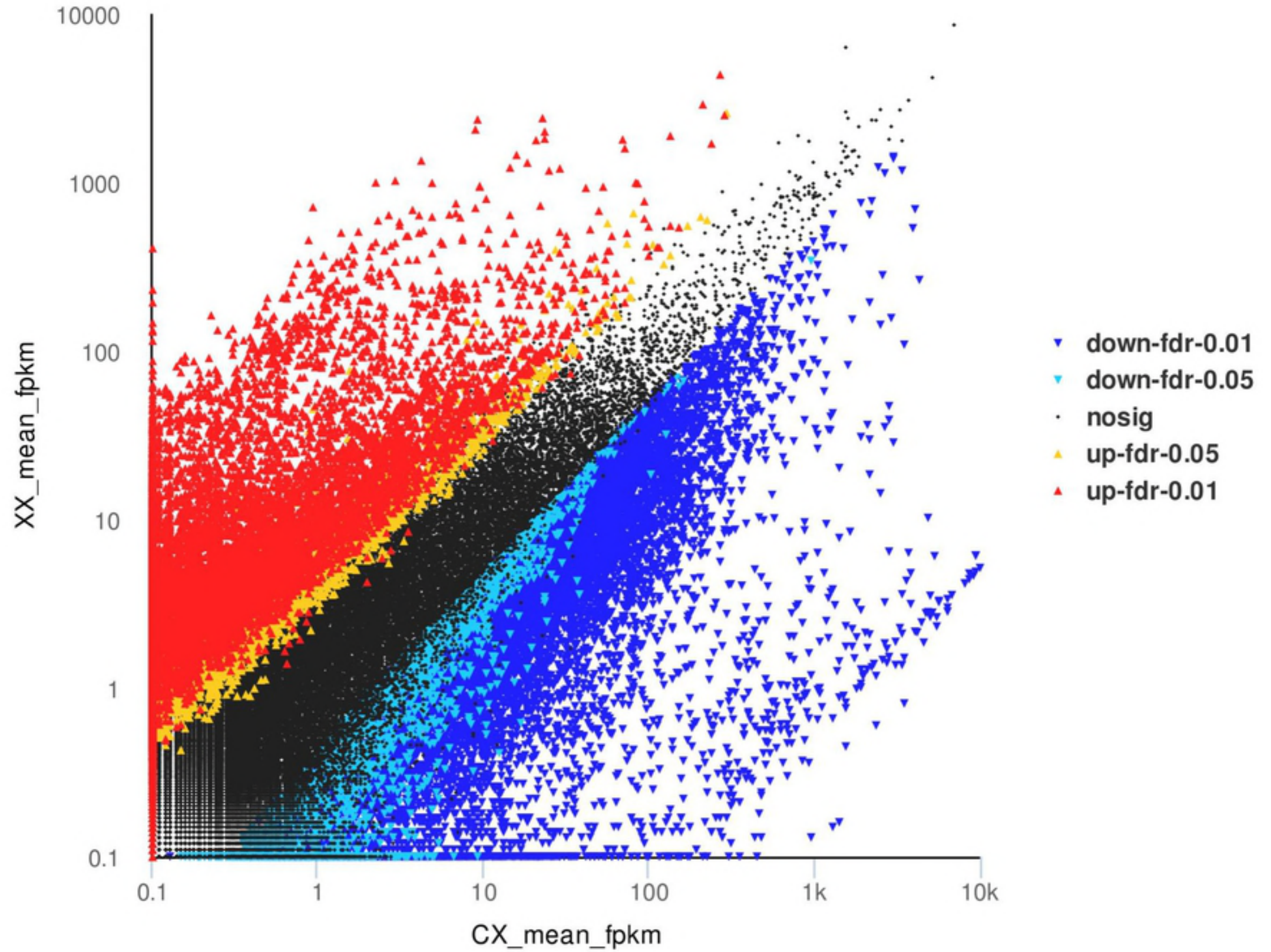


Table 1. Statistics of the sequencing results from testes and ovaries of *Culter alburnus*.

The sample name	Number of reads	Number of bases (bp)	Error%	Q20%	Q30%	GC%
Testes	127,931,976	17,360,771,986	0.0141	97.01	93.055	48.485
Ovaries	27,215,890	3,704,339,375	0.0127	97.785	94.585	46.83

Table 2. Representative sex-related differentially expressed genes in *C.alburnus*.

seq id	Gene	length	log_2fold	function
TR10442|c0_g2	*cyp19a*	2066	-3.5	cytochrome P450 aromatase
TR5369|c1_g1	*fel*	577	-4.73	Fish-egg lectin
TR76297|c3_g9	*zp4*	2193	-9.93	zona pellucida sperm-binding protein 4-like
TR2143|c2_g3	*zar1*	1741	-8.77	Oocyte-specific maternal effect factor
TR41583|c0_g1	*zp3*	2413	-9.11	Zona pellucida sperm-binding protein 3
TR15030|c0_g2	*foxl2*	1378	-6.74	forkhead box protein L2-like
TR41118|c0_g3	*sox4*	3751	-4.65	SRY (sex determining region Y)-box 4a
TR13699|c0_g1	*nasp*	617	-4.31	nuclear autoantigenic sperm protei
TR70355|c0_g1	*nanog*	2060	-5.97	ovary-expressed homeobox protein
TR3462|c0_g2	*sox11b*	1770	-5.37	SRY (sex determining region Y)-box 11b
TR28684|c0_g1	*srm*	2403	-4.15	spermidine synthase isoform X1
TR75077|c1_g1	*smox*	1827	-3.65	Smox protein
TR86814|c0_g2	*SSAT-1*	1250	-3.2	spermidine/spermine N1-acetyltransferase
TR28499|c4_g2	*stpg2*	1236	7.7	sperm-tail PG-rich repeat-containing protein 2
TR36879|c0_g11	*sox30*	663	5.91	SRY (sex determining region Y)-box 30
TR36862|c2_g5	*scmh1*	4908	3.89	sex comb on midleg homolog 1
TR6055|c0_g4	*spata22*	1534	6.06	spermatogenesis associated 22

TR77837|c1_g1	*dmrt1*	2856	6.64	dsx and mab-3 related transcription factor 1
TR8130|c5_g1	*odf3b*	317	7.12	outer dense fiber of sperm tails 3B
TR487|c8_g1	*izumo1*	1606	6.97	izumo sperm-egg fusion protein 1-like
TR8607|c0_g1	*spag16*	2589	7.49	sperm associated antigen 16
TR81938|c1_g1	*spef*	2870	6.79	sperm flagellar protein 1

Note: Fold indicated RPKM (male)/RPKM (female).