FishExp: A comprehensive database and analysis platform for gene expression and alternative splicing of fish species

Suxu Tan a, Wenwen Wang c, Wencai Jie d, Jinding Liu a,b,*

a Department of Animal Science, Michigan State University, East Lansing, MI 48824, USA
b Bioinformatics Center, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China
c School of Fisheries, Aquaculture and Aquatic Sciences, Auburn University, Auburn, AL 36849, USA
d Institute for Plant Molecular Biology, State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, Jiangsu 210023, China

A R T I C L E   I N F O

Article history:
Received 23 March 2022
Received in revised form 7 July 2022
Accepted 7 July 2022
Available online 11 July 2022

Keywords:
Alternative splicing
Gene expression
Fish
Database
Analysis platform

A B S T R A C T

The publicly archived RNA-seq data has grown exponentially, while its valuable information has not yet been fully discovered and utilized, such as alternative splicing and its integration with gene expression. This is especially true for fish species which play important roles in ecology, research and the food industry. Furthermore, there is a lack of online platform to analyze users’ new data individually and jointly with existing data for the comprehensive analysis of alternative splicing and gene expression. Here, we present FishExp, a web-based data platform covering gene expression and alternative splicing in 26,081 RNA-seq experiments from 44 fishes. It allows users to query the data in a variety of ways, including gene identifier/symbol, functional term, and BLAST alignment. Moreover, users can customize experiments and tools to perform differential/species expression and alternative splicing analysis, co-expression and cross-species analysis. In addition, functional enrichment is provided to confer biological significance. Notably, users are allowed to submit their own data and perform various analyses using the new data alone or alongside existing data in FishExp. Results of retrieval and analysis can be visualized on the gene-, transcript- and splicing event-level webpage in a highly interactive and intuitive manner. All data in FishExp can be downloaded for more in-depth analysis. The manually curated sample information, uniform data processing and various tools make it efficient for users to gain new insights from these large data sets, facilitating scientific hypothesis generation. FishExp is freely accessible at https://bioinfo.njau.edu.cn/fishExp.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Fishes are the largest group of vertebrates, with over 34,000 species [1], more than all other vertebrate species combined. They are the earliest vertebrates on Earth and have evolved for more than 500 million years [2,3]. Living in a variety of habitats globally, fishes play vital structural and functional roles in the aquatic ecosystem. With the wide range of time and geographical scale, they exhibit extremely high levels of biodiversity in terms of morphology, behavior, ecology, and among others. Fish species have been utilized as excellent models in studies of development, physiology, behavior, toxicology, evolution and genetics. For instance, zebrafish and medaka are valuable models for studying human genetics and disease [4–6]. Additionally, many fish species serve as important food sources in fisheries and aquaculture [7].

With the advancement of high-throughput sequencing technologies, fish research has entered the era of omics and profoundly revolutionized our understanding of biology, diversity and disease [8]. So far, transcriptomic studies mainly focused on differential gene expression, with negligence of many other crucial layers of regulations such as splicing. Alternative splicing (AS) is the mechanism by which a single pre-mRNA molecule generates different mature mRNAs (transcripts or isoforms), enhancing proteomic diversity and gene expression modulation. Studies have demonstrated its prevalence and importance in eukaryotes, particularly in vertebrate species. For example, nearly all multi-exonic genes in human exhibit AS events, significantly increasing the complexity and function of gene expression [9,10]. AS functions in a cell-, tissue-, or condition-specific manner, and plays key roles in development, disease, stress response and evolution [11–13].
Efficient retrieval and display of such important information can improve our understanding of gene regulatory networks and facilitate genome annotation and future functional research. A number of useful databases have been developed, including TCGA SpliceSeq [14], ASAPedia [15], VastDB [16], ASlive [17] and MeDAS [18]. These databases, however, typically accommodate a small number of model organisms or provide specialized scope of knowledge and analysis, and none of them is specifically designed for fish species. Some intriguing individual studies have been carried out in fish [19–23], for example, the use of an alternative 5' splice site of the gene MSX2A in freshwater three-spined stickleback (Gasterosteus aculeatus) resulted in shorter dorsal spines [20]; Tan et al. investigated the genome-wide changes of AS profiles and revealed the enrichment of RNA binding and splicing after both biotic and abiotic stresses in catfish [21–23]. Nevertheless, AS research in fish is still at an early stage, especially for genome-wide AS analysis and non-model fish species. We thereby created FishExp, a user-friendly, highly interactive database and analysis platform for the comprehensive analysis of AS and gene expression. We developed a uniform bioinformatic pipeline to integrate genomic/transcriptomic data with detailed metadata to systematically analyze gene expression/AS profiles of 44 fishes with sufficiently publicly available genomic and transcriptomic data. Genome annotations of most fish species were greatly improved, i.e., many novel transcripts and splicing events were discovered and displayed. Furthermore, users can flexibly customize RNA-seq data sets of interest and examine differentially expressed genes (DEG) and differentially alternatively spliced (DAS) events/genes, provided with functional enrichment analysis. One highlight is that the platform allows users to analyze their new data alone or co-analyze with the existing data of FishExp. This study provides an added-value resource for public repertoire and offers convenient tools of retrieval, analysis and visualization for the fish genome research community.

2. Materials and methods

2.1. Data collection and database content

The reference genome assemblies and original annotation of fish species were collected from Ensembl, alternatively from NCBI. RNA-seq data were collected from Sequence Read Archive (SRA) database by querying the SRA metadata (as of July 2020) [24]. Illumina sequencing data was exclusively collected due to its ubiquity and high base quality. The metadata is manually curated and organized, including strain, genotype, tissue, development and treatment. A total of 26,081 high-quality RNA-seq experiments of 664 studies from 44 fish species were harbored by FishExp (Table 1). A phylogenetic tree of the studied species was generated based on NCBI taxonomy using PhyloT (https://phylot.biobyte.de) and visualized using iTol [25] (Fig. 1). Fig. 2 briefly illustrates the data collection, manual curation and data processing, and highlighted features of FishExp.

2.2. Gene annotation improvement

The reference genome sequences and annotation of 41 and 3 species are from the Ensembl and NCBI, respectively (Table 1). Accurate and complete gene annotation is extremely important for the unambiguous quantification of expression or splicing from RNA-seq experiments [26]. Whereas online genome annotations are largely incomplete, even for widely studied organism like human. To address this, we improved the gene/transcript model using the following steps. First, high-quality RNA-seq data sets were mapped to the reference genome by HISAT2 [27], and then assembled into transcripts using StringTie2 [28]. Second, we kept novel multi-exonic transcripts with enough length (≥200 bps) and average coverage (≥2 x per transcript and ≥1 x per exon). At last, we retained only the novel transcripts with high confidence, shown in at least 1/3 of all experiments and at least 3 experiments. All improved annotations are available for download on the Summary page of the FishExp database.

2.3. Gene expression estimation and differential analysis

We used HISAT2 to perform read alignment against the reference genome and used StringTie2 to assemble them into transcripts and obtain expression levels for genes/transcripts. For co-expression analysis, the R package WGCNA was utilized [29]. For differential expression analysis, two most widely used tools were employed: DESeq2 [30] and edgeR [31]. It is noted that edgeR allows three test models including exact test, likelihood test and quasi-likelihood F test; it also supports comparison without sample replicates, which is useful for exploring many early sequencing experiments.

2.4. Alternative splicing detection and differential analysis

By generating multiple isoforms from a single gene, AS influences diverse cellular processes, including stability, localization, binding and enzymatic properties [13]. AS and DAS analyses could provide new insights into biological processes and disease conditions, however, they were underestimated and far from being fully mined in existing RNA-seq data sets. To overcome this, rMATS [32] and customized scripts were used to perform analyses of AS and DAS. Five canonical AS types were considered, including exclusion or inclusion of individual exon (SE), alternative 5' splice site and 3' splice site (A5SS and A3SS), retention of intron (RI), mutually exclusive splicing of adjacent exon (MXE). Among them, SE is the most common type, accounting for approximately 38.8% of all AS events detected in this study, followed by RI, A3SS and A5SS, while MXE only occurs in about 2.5% of AS events (Fig. 3A). Approximately 13.4% of the total AS events were novel detected by rMATS based on read mapping (Fig. 3B).

2.5. Functional enrichment

To help better understand the functions of the biological system, we employed ClusterProfiler R package [33] to conduct functional enrichments of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway. For genomes from Ensembl, GO terms come from its functional annotation, while Blast2GO [34] was used to obtain GO annotations for RefSeq genomes. GO annotation of some species, especially model species, can also be directly extracted from AnnotationHub [35]. The pathway assignment was achieved by submitting protein sequence to KAAS (KEGG automatic annotation server) [36], which works mainly based on sequence similarities. The pathway annotation of widely studied species can be extracted directly from KEGG database. In addition to the typical enrichment analysis using hypergeometry test, we also provide options for enriching GO and pathways using Gene Set Enrichment Analysis (GSEA) [37].

2.6. Orthologous gene and AS event

Gene conservation indicated positive natural selection, reflecting their evolutionary and functionally important roles. In an attempt to explore the gene conservation between fish species, orthologous genes between species were identified by OrthoFinder [38] using the longest protein sequence of each gene. Gene splicing can be found in almost all eukaryotic species. The formation and disappearance of AS events occur during evolution, and conserved AS provides strong evidence of biological function [39]. To explore
the conservation of AS, we carried out the following steps. First, the protein sequences within one orthologous gene group were aligned using MAFFT [40]. Second, the protein alignments were converted to codon alignments using PAL2NAL [41]. Finally, we assigned new coordinates of exons in transcripts based on the conservation of AS. We carried out the following steps. First, the protein sequences within one orthologous gene group were aligned using MAFFT [40]. Second, the protein alignments were converted to codon alignments using PAL2NAL [41]. Finally, we assigned new coordinates of exons in transcripts based on the conservation of AS.

2.7. Implementation of FishExp

The data in FishExp are stored and managed in the relational databases MySQL. The web interfaces are based on HTML, CSS and JavaScript. The backend processing scripts use PHP, Perl and R language. A genome browser is implemented using JBrowse [42] to facilitate convenient visualization of genes, transcripts and AS.

3. Results

3.1. Summary of genomic and transcriptomic data

The species table on the homepage lists all 44 species. The summary icon in the species table leads the user to view a brief biological introduction of the species, genomic statistics and the information of RNA-seq experiments. Of them, the genomic statistics include the number of original gene/transcript, improved transcript/exon/splice and AS event. Statistics of improvement results for all species are listed in Supplementary Table S1. On average, the number of exons and splice junctions increased by approximately 5.8% and 9.8%, respectively; the number of transcripts per multi-exon gene increased from 1.9 to 2.3; and the proportion of genes with alternative transcripts increased from 36.4% to 51.7% among all multi-exon genes. All the genomic and transcriptomic data, such as genome sequence/annotation, gene/transcript expression and AS data, are downloadable for users to conduct off-line analyses.

3.2. Advanced searching

The search term for FishExp can be in any of the following formats: gene ID/symbol or functional categories including protein family, gene ontology and KEGG pathway (Fig. 4A). Moreover, the BLAST tools, including blastn, blastp and blastx, enable the user to search for targeted genes by supplying protein or nucleotide sequences from the current or other species (Fig. 4B). The search results provide annotations with links to external mainstream databases including Swissport, pfam, GO and KEGG (Fig. 4C). Furthermore, clicking on the expression tab will direct the user to
3.2. Gene page

The gene page, which displays 1) the general information of the gene, including the orthogroup with a popup window displaying the orthologous genes of other fish species, helping to explore the cross-species gene conservation; 2) a genome browser for intuitive visualization of the gene model and associated transcripts/AS event, from which the sequence of the region of interest can be obtained; 3) a hierarchical bar chart displaying gene expression profiles in experimental groups of a selected study at the first level and sequencing experiments at the second level by clicking a certain group. The gene page can lead the user to the transcript page and splicing page which have the similar layout (Fig. 4D-G). Both transcript page and splicing page present expression value using TPM (transcript per million) or FPKM (Fragments Per Kilobase of transcript per Million mapped reads) based on user selection, whereas the splicing page shows the percent spliced-in (PSI) indicating the exon inclusion level for a certain AS event.

3.3. Differential expression and alternative splicing

Most transcriptomic studies strive to identify and investigate differences between groups, which can provide insights into the underlying molecular mechanisms and generate new hypothesis. Here in the FishExp server, both differential gene expression (DGE) and differential alternative splicing (DAS) analyses are offered for users. The comparative analyses can be customized in many aspects. First, based on various information (strain, genotype, tissue, development and treatment), two groups of RNA experiments can be flexibly selected by clicking the corresponding icon in the “Control” and “Treatment” column (Fig. 5A). In addition, the search box in the upper right corner can be used to locate certain study, RNA experiment or data source. Second, a number of analysis tools and parameters of DGE/DAS and GO/KEGG enrich-
ment analyses are user-defined (Fig. 5B and C). For instance, the DAS analysis provides different statistical models including MATS LRT, rMATS unpaired and rMATS paired [32].

The link of the result page including various tables and figures will be sent to the user-provided email after job submission. The top of the result page displays the basic analysis information of the user-selected samples, analysis tools and parameters. The remaining page is separated into two sections of analysis results: differential expression and differential splicing. Each section contains two clustering plots, PCA and heatmap (Fig. 5D and E), which...
provide an overview of the variation of RNA-seq data and help check the group selection. It is followed by figures of enriched GO (Fig. 5F) and pathway and the interactive table of DEG or DAS (Fig. 5G) if available. Moreover, a download button is offered for users to obtain all the tables/high-quality figures for publication and other detailed information for further analysis.

3.4. Specificity analysis

Specifically expressed genes (SEG) or specifically spliced genes (SSG) are of great interest since they may exert specific functions in a certain tissue, genotype or disease condition. The FishExp server provides the specificity analysis as an extended function of the...
differential analysis. The only difference is that at least four groups need to be selected in the specificity analysis, and the pairwise comparison between groups will be performed. The specifically expressed gene is referred to the gene which expressed significantly higher or lower in one group than that in any other groups. Similarly, the specifically alternatively spliced gene refers to the gene whose PSI value in one group is significantly higher or lower than that in other groups. The specificity result page is similar to that of the differential analysis between two groups, except that it additionally contains the chart displaying the number of SEG/SSG with significantly higher or lower expression/splicing levels than other groups.

3.5. Co-expression and cross-species analysis

Cluster of highly correlated genes may be responsible for specific traits of interest. Co-expression analysis allows us to explore the potentially important hub genes of a module (gene cluster), the relationships between different modules, and the modules correlated with certain samples (traits or features). In FishExp, various
WGCNA parameters can be set for co-expression analysis, such as tests for correlation coefficient (Person, Spearman, and Kendall), soft thresholding power, and network types (signed and unsigned). In addition to the standard WGCNA analysis pipeline, enrichment of GO terms and KEGG pathways for each gene module is available to reveal its biological function.

The cross-species analysis focuses on the gene expression of orthologous gene and gene group of two specified species. As above, DESeq2 and edgeR can be selected for differential gene expression; Person, Spearman, and Kendall tests are available for the measure of correlation. As an example of an analysis, single-copy ortholog of two species which are both differentially expressed between two study groups are listed, which may imply evolutionarily conservation and crucial function in certain biological process. For ease of understanding and further exploration, rich figures and tables are displayed in the result page.

3.6. Your-own-data analysis

Users can submit their own data and perform all of the above analyses using their data alone or in conjunction with existing data in FishExp. All they need to do is follow the pipeline and scripts we provide in the “co-analysis” page.

3.7. A case study for differential splicing patterns

We demonstrate here how the FishExp can reveal novel regulatory information from published research, promote the understanding of underlying mechanisms and help generate new hypothesis for future studies. A simple design in zebrafish was selected [43], which compared the gene expression in proximal intestine of adult zebrafish with and without short bowel syndrome (SBS). The SBS zebrafish model underwent treatment of laparotomy, proximal stoma and distal ligation, while the control fish experienced laparotomy alone. This study focused on overall changes in gene expression. We further performed the analyses of AS and DAS using FishExp web server with a few clicks and obtained the results in minutes (Fig. 5).

Both PCA of gene expression and PCA of AS (Fig. 5D) show that samples are distinguishably clustered to two groups in accordance to the treatment and control group, reflecting that SBS zebrafish exhibit evident changes in both gene expression and AS profiles. Using the selected tools and parameters (Fig. 5B and C), a total of 76 DAS events of 66 DAS genes were identified, seven of which were also differentially expressed genes (Supplementary Table S2).

Remarkably, the DAS genes were almost entirely enriched in the biological process of RNA splicing/processing and molecular function of RNA binding (Fig. 5F). In addition, there is only one enriched KEGG pathway, spliceosome (d003040), which plays central roles in the splicing process of eukaryotic genes. This observation is consistent with the findings that genes encoding RNA binding proteins including splicing factors and spliceosomal components themselves often undergo alternative splicing [44-48].

Splice site selection on pre-mRNA under specific conditions is determined by the binding of splicing factors, which recruit numerous spliceosomal components and thereby the spliceosome [49,50]. Changes in levels or activity of splicing factors may have profound effects on the expression of downstream target genes [46]. The main classes of splicing factors are Ser/Arg-rich (SR) proteins and heterogeneous nuclear ribonucleoprotein particle (hnRNPs) proteins. Among the DAS events in this study, we identified four DAS events of three SR genes including srsf3a (SE and A3SS), srsf4 (IR) and srsf7a (SE) as well as one differential ASSS event of hnrnpr. The expression or activity of these factors were regulated by alternative splicing, which may affect downstream gene network. For instance, the ASSS event on hnrnpr could result in an insertion of 18 nucleotides on the transcript of ENSDART00000172319, which overlap the translated acidic sequence segment domain (PF18360). The DAS events in the present study imply the involvement of alternative splicing in SBS and may generate new hypotheses for disease progression and treatment.

4. Discussion

The advent of RNA-seq has revolutionized our understanding in the complexity and function of gene expression regulation, and emphasized the considerable roles of AS in various biological processes. The study of AS is still lacking in fish species, whose RNA-seq data is sitting in public repertoire but not being fully explored. The hidden information may reveal valuable information on gene expression regulation and suggest new functional association for further investigations. We thereby created FishExp to help researchers address the complexity in analyzing and visualizing the gene expression and alternative splicing profiles.

Serving as an added-value resource, FishExp not only provides a comprehensive survey of the profiles of gene expression/splicing, but also is ideally suited to study the DEG/DAS genes/events, co-expression and cross-species gene regulation, with functional enrichment. Most importantly, new data from users can be submitted for above-mentioned analyses or co-analysis with data stored in FishExp. As such, the platform would be of great interest to a broad range of users. In addition to provide a wealth of information and analysis tools, we sought to make the database easy to use. The database is convenient to navigate with logical, hierarchical and interactive webpage, displayed with a highly interactive interface. Many external links are embedded for further exploration, and all data can be downloaded in batch for additional offline analysis.

In the future, we strive to frequently update FishExp to cover newly generated transcriptome data for current and additional fish species, including long sequencing reads for accurate transcript assembly and splicing detection. We envision to extend FishExp to provide more analyses, such as alternative polyadenylation and RNA-editing. We anticipate FishExp to become a very useful resource to explore the profiles and functions of gene expression/splicing, to expand our understanding of gene expression regulation, and to promote hypothesis generation for further research.

CRediT authorship contribution statement

Suxu Tan: Conceptualization, Data curation, Investigation, Visualization, Writing – original draft, Writing – review & editing. Wenwen Wang: Data curation, Writing – review & editing. Wencai Jie: Software. Jinding Liu: Investigation, Software, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the work of the genome and RNA-seq data producers.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.07.015.
