Chou *et al. BMC Genomics* 2018, **19**(Suppl 2):103
DOI 10.1186/s12864-018-4463-x

BMC Genomics

RESEARCH Open Access

The aquatic animals' transcriptome resource for comparative functional analysis

Chih-Hung Chou[1,2†], Hsi-Yuan Huang[1,2†], Wei-Chih Huang[1,2†], Sheng-Da Hsu[1], Chung-Der Hsiao[3], Chia-Yu Liu[1], Yu-Hung Chen[1], Yu-Chen Liu[1,2], Wei-Yun Huang[2], Meng-Lin Lee[2], Yi-Chang Chen[4] and Hsien-Da Huang[1,2*]

From The Sixteenth Asia Pacific Bioinformatics Conference 2018
Yokohama, Japan. 15-17 January 2018

Abstract

Background: Aquatic animals have great economic and ecological importance. Among them, non-model organisms have been studied regarding eco-toxicity, stress biology, and environmental adaptation. Due to recent advances in next-generation sequencing techniques, large amounts of RNA-seq data for aquatic animals are publicly available. However, currently there is no comprehensive resource exist for the analysis, unification, and integration of these datasets. This study utilizes computational approaches to build a new resource of transcriptomic maps for aquatic animals. This aquatic animal transcriptome map database dbATM provides de novo assembly of transcriptome, gene annotation and comparative analysis of more than twenty aquatic organisms without draft genome.

Results: To improve the assembly quality, three computational tools (Trinity, Oases and SOAPdenovo-Trans) were employed to enhance individual transcriptome assembly, and CAP3 and CD-HIT-EST software were then used to merge these three assembled transcriptomes. In addition, functional annotation analysis provides valuable clues to gene characteristics, including full-length transcript coding regions, conserved domains, gene ontology and KEGG pathways. Furthermore, all aquatic animal genes are essential for comparative genomics tasks such as constructing homologous gene groups and blast databases and phylogenetic analysis.

Conclusion: In conclusion, we establish a resource for non model organism aquatic animals, which is great economic and ecological importance and provide transcriptomic information including functional annotation and comparative transcriptome analysis. The database is now publically accessible through the URL http://dbATM.mbc.nctu.edu.tw/.

Background

Aquatic animals have significant economic benefits for humans and also play key roles in the development of medical applications [1, 2]. Aquaculture transforms natural aquatic resources such as fish, shrimp, and mussels into socially-valued commodities [3]. Chemicals derived from aquatic organisms are used to develop pharmaceutical compounds with important clinical applications [2]. In most of the aquaculture industries, prophylactic antibiotics have been used to prevent bacterial and virus infections [4].

To investigate the genomic resource of aquatic animals, next generation sequencing (NGS) approach has been adapted to uncover new genes and biological mechanisms. To effectively increase the our knowledge on aquatic animal gene discovery at mRNA level, RNA-seq has been applied to several important aquaculture species such as, *Plecoglossus altivelis* [5], *Cyprinus carpio* [6], *Penaeus monodon* [7], *Hyriopsis cumingii* [8], *Mytilus galloprovincialis* [9], and *Anguilla anguilla* [10]. To gain insights into the immunogenetics and immune response system, RNA-seq has been successfully adapting to

* Correspondence: bryan@mail.nctu.edu.tw
†Equal contributors
[1]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan
[2]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan
Full list of author information is available at the end of the article

study the host defense gene activities against bacteria and viruses [11]. *Clupea pallasii*, *Fundulus grandis*, *Pandalus latirostri*, etc. [12–15], have recently come under significant additional anthropogenic and environmental pressure, and SNPs and a set of genes associated with immune function in these species have been identified as providing high differentiation between two regions, and this insight can be applied in developing ecological monitoring systems. Evolutionary research on species including *Sinocyclocheilus angustiporus* (a comparative investigation between surface and cave species) revealed reduced expression of a series of visual photo-transduction and retinal disease-related genes in cave-dwelling species [16]. *Astyanax mexicanus* was also used to clarify heritable genetic changes governing adaptation to cave environments [17]. Beyond NGS transcriptomic data analysis, the MitoFish database focuses on fish mitochondrial genome annotation to explore a new resource for resolving fish phylogenies and identifying new fish species [18].

NGS technology has emerged as a powerful tool to illustrate the blueprints of novel species. Whole genome sequencing (WGS) and RNA sequencing (RNA-seq) can be used to investigate the individual differences, evolution and gene function [19, 20]. However, whole genome sequencing is more costly in terms of expense and computing resources than RNA-seq, and cannot obtain comprehensive information for transcriptomes directly by gene prediction. A better strategy is using RNA-seq to elucidate the molecular basis of biological functions [21]. Given the economical importance of these aquatic animals, a comprehensive resource of transcriptomes should be establsihed despite the lack of draft genomes of these non model organisms.

RNA-seq technology pave the way for transcriptome data investigation. With the techniques, transcripts can not only be detected in accordance with the reference genome but also can be used for de novo the assembly of RNA-seq reads to discover genes and profile their expression in organisms without a reference genome [22–24]. However, even if these data are deposited in the NCBI Sequence Read Archive (SRA) [25] which provides high-throughput unassembled and un-annotated raw NGS reads for more than 1000 Terabase pairs, public access is not provided to allow for searches of the well-annotated data. To address this problem, we have created a queryable database from the annotated transcriptomic database, called dbATM (http://dbATM.mbc.nctu.edu.tw). A similar approach was reported for ASGARD [26], which collects the annotated transcriptomes from three model arthropod species and NHPRTR [21], focusing on non-human primate NGS transcriptomic sequencing and analysis. dbATM is an easily-accessible database which integrates functional analysis applications.

Aquatic animals have high economic value and also play important roles in the monitoring of environmental pollution in ecosystems. Aquatic animals' related studies have seen exponentially increased in recent years (Additional file 1: Figure S1). Many studies have used NGS approaches to investigate genomic diversity among aquatic animals at genomic and transcriptomic levels. However, these studies largely focus on single species, and no databases exist so far to facilitate homologous gene mining among multiple aquatic species. In this study, we aim to establish an integrated aquatic animal transcriptomic database to facilitate studies in the fields of genomics, evolution, and phylogeny. We initially performed whole organism RNA-seq on four aquatic animals and collected RNA-seq data for eighteen species from the public domain. Reference genome free RNA-seq data of aquatic animals download from the three common NGS reads database, NCBI Sequence Read Archive (SRA), European Nucleotide Archive (ENA), and DDBJ Sequence Read Archive (DRA). Then, a computational pipeline was developed to annotate these NGS transcriptomes. Finally, all data were summarized and provided in the dbATM database. Figure 1 illustrates how the database can be used for mining transcriptomics, functions such Gene Ontology and KEGG pathways and comparative analysis of aquatic animals. The dbATM collects transcriptomes for twenty-two aquatic animals' transcriptomes and provides an invaluable tool for homologous evolution and evo-devo studies.

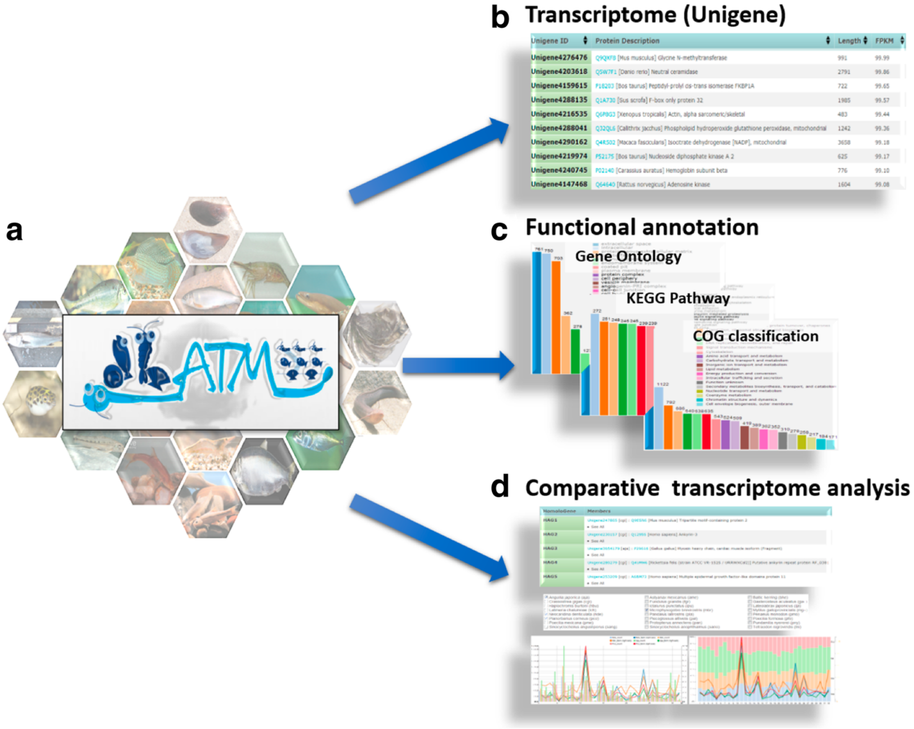
Results

Overview of dbATM content

Additional file 1: Table S1 displays all aquatic animal RNA-seq data sources and original data types. Following systematic analysis (Fig. 2), the individual species information and comparative analysis can be accessed from dbATM. Table 1 shows the species annotation of the transcripts, unigenes, proteins, and homologous genes. To facilitate evolutionary studies and comparative analysis across the different aquatic animal species, a taxonomic tree was constructed according to biological classification (Fig. 3), for shellfish, shrimp, and fish from invertebrate to vertebrate.

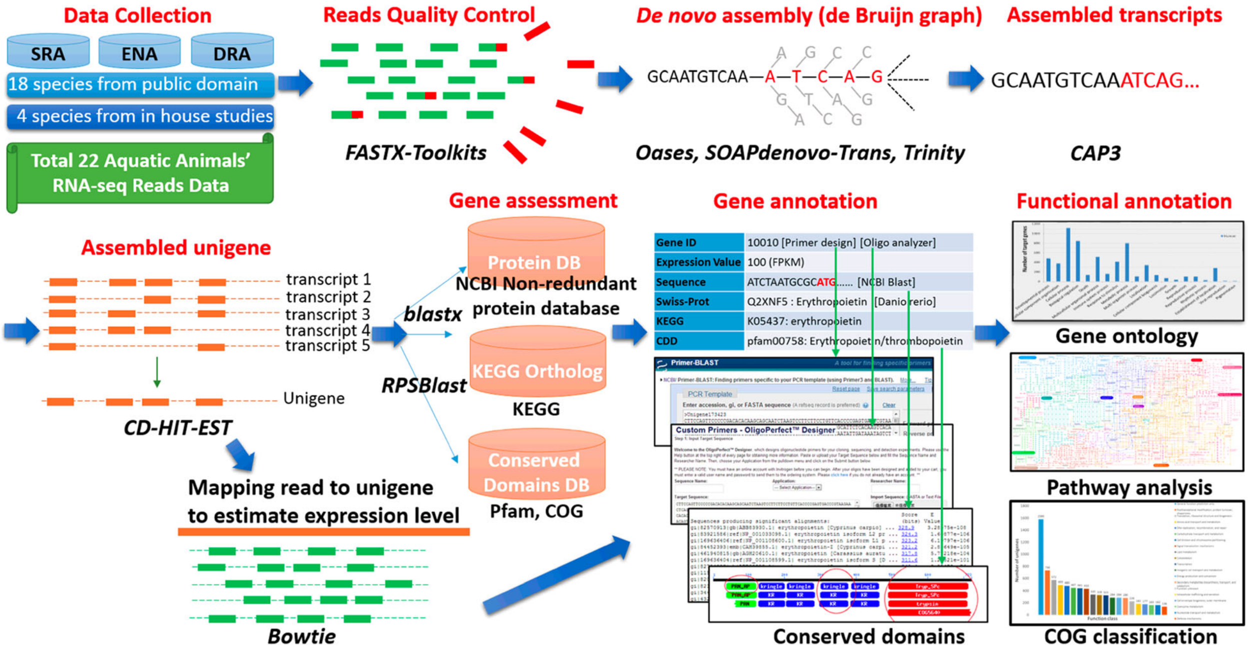
Web interface

The dbATM provides various query interfaces and graphical visualization pages to access to aquatic transcriptome data (Fig. 4). The summary table shows species information including scientific name (or common name), followed by total sequencing read size (in bp), sequencing platform, type of read sequences, number of reads and a link to NCBI SRA and NCBI Taxonomy and Wikipedia. Several browser functions were designed to allow users to obtain transcriptomic information. Users can browse by species, range of gene expression in FPKM, gene length, Gene Symbol, and Gene Ontology [27] based



Unigene ID	Protein Description	Length	FPKM
Unigene4276476	Q9QXF8 [Mus musculus] Glycine N-methyltransferase	991	99.99
Unigene4203618	Q5W7F1 [Danio rerio] Neutral ceramidase	2791	99.86
Unigene4159615	P18203 [Bos taurus] Peptidyl-prolyl cis-trans isomerase FKBP1A	722	99.65
Unigene4288135	Q1A730 [Sus scrofa] F-box only protein 32	1985	99.57
Unigene4216535	Q6P8G3 [Xenopus tropicalis] Actin, alpha sarcomeric/skeletal	483	99.44
Unigene4288041	Q32QL6 [Callithrix jacchus] Phospholipid hydroperoxide glutathione peroxidase, mitochondrial	1242	99.36
Unigene4290162	Q4R502 [Macaca fascicularis] Isocitrate dehydrogenase [NADP], mitochondrial	3658	99.18
Unigene4219974	P52175 [Bos taurus] Nucleoside diphosphate kinase A 2	625	99.17
Unigene4240745	P02140 [Carassius auratus] Hemoglobin subunit beta	776	99.10
Unigene4147468	Q64640 [Rattus norvegicus] Adenosine kinase	1604	99.08

Fig. 1 Aquatic animal transcriptomes maps database. **a** The database collects RNA-seq data for more than 20 different aquatic animals. The database consists of three parts: **b** Detailed information for individual genes such as gene name and description, length, expression (FPKM), and sequence. **c** Functional annotation of all gene groups in individual species. **d** Evolutionary studies of all species transcriptomic data by constructing a comparative analysis system for homologous genes



Gene ID	10010 [Primer design] [Oligo analyzer]
Expression Value	100 (FPKM)
Sequence	ATCTAATGCGCATG...... [NCBI Blast]
Swiss-Prot	Q2XNF5 : Erythropoietin [Danio rerio]
KEGG	K05437: erythropoietin
CDD	pfam00758: Erythropoietin/thrombopoietin

Fig. 2 dbATM database system and analysis. First, publicly available RNA-seq data were collected and combined with additional data produced for this study. Next, the NGS raw data were trimmed and filtered to remove low quality reads. Third, NGS reads were assembled based on the de Bruijn graphs and clustered into transcripts or unigenes. Finally, assembled reads were represented to genes and gene functions by BLAST against external database

Table 1 Gene annotation statistics in dbATM

Species	No. of transcripts	No. of proteins	No. of UniGene	No. of homologous genes
Anguilla japonica	99,696	38,815	13,600	9893
Astyanax mexicanus	59,723	23,085	10,443	7838
Clupea harengus	28,360	15,975	7713	5773
Crassostrea gigas	47,472	33,491	1149	377
Fundulus grandis	32,405	17,620	7388	4707
Gasterosteus aculeatus	77,510	43,170	14,035	6394
Ictalurus punctatus	57,435	28,850	13,400	10,452
Lateolabrax japonicus	27,075	15,130	6502	4227
Microphysogo biobrevirostris	50,827	31,923	14,839	11,295
Mytilus galloprovincialis	10,738	5944	815	141
Neocaridina denticulate	59,513	23,033	6593	1406
Pandalus latirostris	39,062	11,894	4127	1065
Penaeus monodon	36,979	14,740	4796	1162
Planorbarius corneus	61,364	20,951	3802	1004
Plecoglossus altivelis	43,524	22,836	9115	6544
Poecilia formosa	47,579	22,728	8271	3836
Poecilia mexicana	39,598	27,485	10,322	6423
Protopterus annectens	26,187	14,498	6050	2463
Pundamilia nyererei	19,205	13,762	5287	2749
Sinocyclocheilus angustiporus	51,104	41,391	11,505	8168
Sinocyclocheilus anophthalmus	81,704	48,526	13,583	10,452
Tetraodon nigroviridis	53,394	32,410	11,781	4238
Total	1,023,379	533,127	185,166	19,363[a]

[a]The total number of homologous gene groups show here is the real homologous gene groups in the database. Not the summation from individual species

on homologous evidence, KEGG orthology pathway [28], COG [29] and Pfam [30]. An introduction page presents assembled transcriptomic information for each species, including statistics for raw sequence data, assembly quality and annotated genes (Fig. 4a, c). The embedded NCBI BLAST page allows users to input sequences or upload a text file containing sequences in FASTA format to search all genes sequences in the dbATM based on nucleotide or protein similarity (Fig. 4b). dbATM also allows users to browse homologous genes in homologous groups (Fig. 4d), and provides an interface for comparing KEGG pathway mapping results with the average FPKM of genes for a given pathway in different species, allowing users to select a maximum of 10 species for expressional profiling comparison.

Investigation of aquatic animals' homologous genes

Most current transcriptomic studies focus on single species, and these "standalone" transcriptomic databases do not allow for the exploration of homologous gene identities and their relative expression levels. To overcome this constraint, we provide a novel function to comparatively analyze potential homologous genes among 22 aquatic animal species. In total, 19,369 homologous genes were identified by OrthoMCL [31] and their gene name. Using homologous search function in dbATM, we can discover that several homologous genes (like *yars*, *cars*, *nop14*, *acsbg2*, *trip12*, *rab3gap2*, and *herc2*) are distributed from invertebrates (mollusca and arthropoda) to vertebrates (ostariophysi, euteleostei, actinopterygii, and chordate). The results shown that the essential genes such as the tRNA synthetase *yars* and *cars*, are evolutionary conserved. We also found there 21 homologous genes shared across at least 17 species: *yars, cars, nop14, acsbg2, trip12, rab3gap2, herc2, ddost, vwa8, cdk11b, ascc3, aplp2, loc563777, nup98, nup188, sdad1, nup205, dlat, acadm, rtn4ip1, kansl3*. The number of homologous genes in each clade were shown in Table 2. The homologous gene lists in each clade were shown in Additional file1: Table S2. Information on homologous genes across different animal phyla will provide good material for future evo-devo studies.

Investigation of gene expression profiles

To profile the gene expression level among different aquatic animals, we compare the gene expression value (displayed by FPKM) across different KEGG pathway

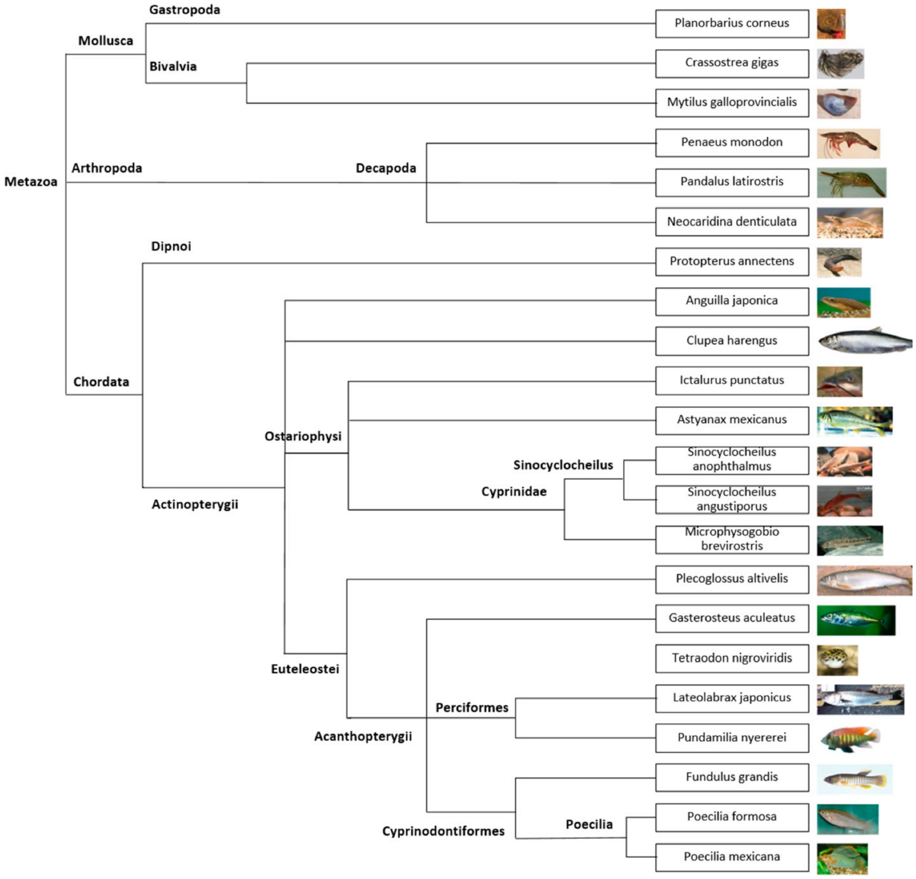
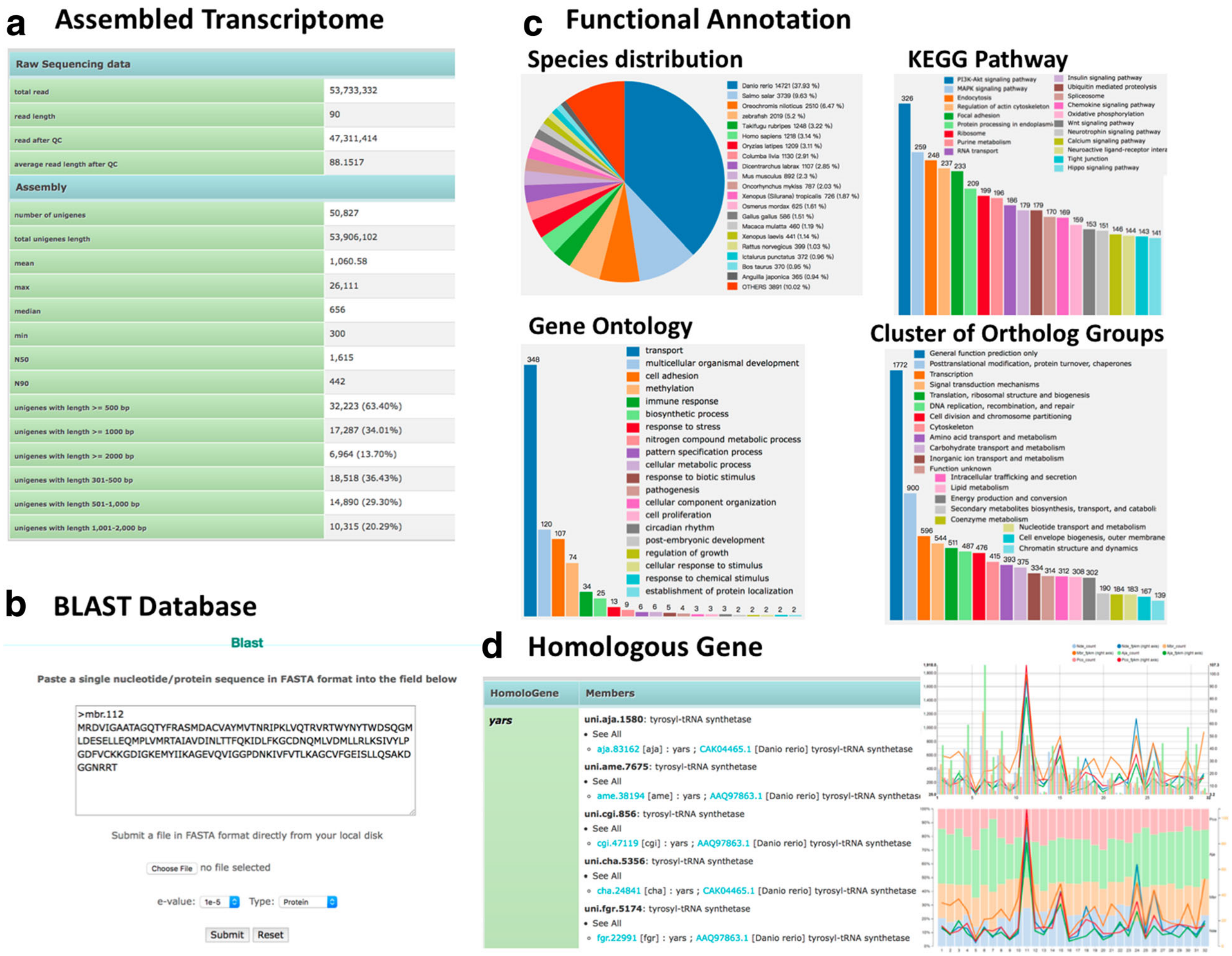


Fig. 3 Species in dbATM and their biological classification. We categorize organisms into sub-groups and represent their rough evolutionary distances for further evolutionary and phylogenetic investigation

categories. Results show the gene expression profile of a particular KEGG pathway stands out in some species. Distinct gene expressional profiling provides a good entry point to compare the gene expression level across different species or different tissue/organs among the same species. For example, the condition of the three species *Sinocyclocheilus angustiporus*, *Sinocyclocheilus anophthalmus*, and *Tetraodon nigroviridis* are from the brain tissue, and show similar expression profile trends (Fig. 5a). However, tissue for another four different condition species *Clupea harengus*, *Gasterosteus aculeatus*, *Protopterus annectens*, *Tetraodon nigroviridis* are whole organism, muscle, liver, and brain, respectively. The expression profiles panel of the four conditions (Fig. 5b) demonstrate quite shock and distinct.

Discussion

We present a new systematic approach for de novo RNA-seq dataset analysis and annotation including improvement of de novo assembly quality and construct a new resource for the transcriptomic map aquatic animals for evolution and phylogenetic study. The system flow is illustrated in Fig. 2 and further detail is provided in the Materials and Methods section. First, we optimized the de novo assembly pipeline of RNA-seq data [32] by combining Oases, SOAPdenovo-Trans, and Trinity assemblers. To obtain complete transcriptomes from various assemblers [33, 34], we used the CD-HIT-EST cluster tool to merge Oases, SOAPdenovo-Trans, and Trinity assembling results. This combinational approach significantly reduce the transcript number and keeping a representative sequence for each unigene. Second, the gene annotation by BLASTX shows that some transcripts are mapped to the same gene or gene family. By following NCBI UniGene approach, we clustered the same gene to a group to reduce redundancy. Third, to facilitate the use of transcriptomes data for evolution and phylogenetic investigations, unigenes derived from all collected species were used to construct a homologous gene group. Besides, we provided the nucleotide (unigene) and protein sequence of all the 22 species on “Download Page”. The sequence data may apply the machine learning approach that could extract new features to develop gene



Raw Sequencing data	
total read	53,733,332
read length	90
read after QC	47,311,414
average read length after QC	88.1517
Assembly	
number of unigenes	50,827
total unigenes length	53,906,102
mean	1,060.58
max	26,111
median	656
min	300
N50	1,615
N90	442
unigenes with length >= 500 bp	32,223 (63.40%)
unigenes with length >= 1000 bp	17,287 (34.01%)
unigenes with length >= 2000 bp	6,964 (13.70%)
unigenes with length 301-500 bp	18,518 (36.43%)
unigenes with length 501-1,000 bp	14,890 (29.30%)
unigenes with length 1,001-2,000 bp	10,315 (20.29%)

Paste a single nucleotide/protein sequence in FASTA format into the field below

```
>mbr.112
MRDVIGAATAGQTYFRASMDACVAYMVTNRIPKLVQTRVRTWYNYTWDSQGM
LDESELLEQMPLVMRTAIAVDINLTTFQKIDLFKGCDNQMLVDMLLRLKSIVYLP
GDFVCKKGDIGKEMYIIKAGEVQVIGGPDNKIVFVTLKAGCVFGEISLLQSAKD
GGNRRT
```

Submit a file in FASTA format directly from your local disk

c
Functional Annotation
Species distribution
Danio rerio 14721 (37.93 %)
Salmo salar 3739 (9.63 %)
Oreochromis niloticus 2510 (6.47 %)
zebrafish 2019 (5.2 %)
Takifugu rubripes 1248 (3.22 %)
Homo sapiens 1218 (3.14 %)
Oryzias latipes 1209 (3.11 %)
Columba livia 1130 (2.91 %)
Dicentrarchus labrax 1107 (2.85 %)
Mus musculus 892 (2.3 %)
Oncorhynchus mykiss 787 (2.03 %)
Xenopus (Silurana) tropicalis 726 (1.87 %)
Osmerus mordax 625 (1.61 %)
Gallus gallus 586 (1.51 %)
Macaca mulatta 460 (1.19 %)
Xenopus laevis 441 (1.14 %)
Rattus norvegicus 399 (1.03 %)
Ictalurus punctatus 372 (0.96 %)
Bos taurus 370 (0.95 %)
Anguilla japonica 365 (0.94 %)
OTHERS 3891 (10.02 %)
KEGG Pathway
PI3K-Akt signaling pathway
MAPK signaling pathway
Endocytosis
Regulation of actin cytoskeleton
Focal adhesion
Protein processing in endoplasmi
Ribosome
Purine metabolism
RNA transport
Insulin signaling pathway
Ubiquitin mediated proteolysis
Spliceosome
Chemokine signaling pathway
Oxidative phosphorylation
Wnt signaling pathway
Neurotrophin signaling pathway
Calcium signaling pathway
Neuroactive ligand-receptor intera
Tight junction
Hippo signaling pathway
326
259
248
237
233
209
199
196
186
179
179
170
169
159
153
151
146
144
143
141
Gene Ontology
transport
multicellular organismal development
cell adhesion
methylation
immune response
biosynthetic process
response to stress
nitrogen compound metabolic process
pattern specification process
cellular metabolic process
response to biotic stimulus
pathogenesis
cellular component organization
cell proliferation
circadian rhythm
post-embryonic development
regulation of growth
cellular response to stimulus
response to chemical stimulus
establishment of protein localization
348
120
107
74
34
25
13
9
6
6
5
4
3
3
3
2
2
2
2
2
Cluster of Ortholog Groups
General function prediction only
Posttranslational modification, protein turnover, chaperones
Transcription
Signal transduction mechanisms
Translation, ribosomal structure and biogenesis
DNA replication, recombination, and repair
Cell division and chromosome partitioning
Cytoskeleton
Amino acid transport and metabolism
Carbohydrate transport and metabolism
Inorganic ion transport and metabolism
Function unknown
Intracellular trafficking and secretion
Lipid metabolism
Energy production and conversion
Secondary metabolites biosynthesis, transport, and cataboli
Coenzyme metabolism
Nucleotide transport and metabolism
Cell envelope biogenesis, outer membrane
Chromatin structure and dynamics
1772
900
596
544
511
487
476
415
393
375
334
314
312
308
302
190
184
183
167
139
d
Homologous Gene
Nde_count
Nde_fpkm (right axis)
Mbr_count
Mbr_fpkm (right axis)
Aja_count
Aja_fpkm (right axis)
Pco_count
Pco_fpkm (right axis)

HomoloGene	Members
yars	**uni.aja.1580**: tyrosyl-tRNA synthetase • See All ○ aja.83162 [aja] : yars ; CAK04465.1 [Danio rerio] tyrosyl-tRNA synthetase **uni.ame.7675**: tyrosyl-tRNA synthetase • See All ○ ame.38194 [ame] : yars ; AAQ97863.1 [Danio rerio] tyrosyl-tRNA synthetas **uni.cgi.856**: tyrosyl-tRNA synthetase • See All ○ cgi.47119 [cgi] : yars ; AAQ97863.1 [Danio rerio] tyrosyl-tRNA synthetase **uni.cha.5356**: tyrosyl-tRNA synthetase • See All ○ cha.24841 [cha] : yars ; CAK04465.1 [Danio rerio] tyrosyl-tRNA synthetase **uni.fgr.5174**: tyrosyl-tRNA synthetase • See All ○ fgr.22991 [fgr] : yars ; AAQ97863.1 [Danio rerio] tyrosyl-tRNA synthetase

Fig. 4 dbATM web interface. The dbATM provides various query interfaces and graphical visualizations to facilitate access to aquatic transcriptomic data. **a** Summary table of the reads, assembly quality, and annotation statistics. **b** A web interface to facilitate unigene mining by entering either sequence or keyword. **c** The functional annotation interface including annotated species distribution, KEGG pathway, Gene Ontology, and cluster of ortholog groups. **d** The homologous gene search and browsing interface for comparative analysis

prediction tools for aquatic animals. Finally, we develop a comparative analysis system by incorporating KEGG pathway analysis and gene expression profiles. This system assesses species and conditional diversity using the gene expression profile in specific biological pathways. dbATM not only improves de novo assembly quality of RNA-seq data but also constructs a database of the homologous genes of aquatic animals to allow for comparative study.

Conclusions

dbATM is the first database to contain comprehensive transcriptomes annotations for more than 20 aquatic

Table 2 Homologous genes statistics in each clade

	Mollusca[a]	*Arthropoda*[b]	*Ostariophysi*[c]	*Euteleostei*[d]	*Actinopterygii*[e]	*Chordate*[f]
Numbers	27	582	3027	85	19	8

[a] *Mollusca* clade: *Planorbarius corneus, Crassostrea gigas, Mytilus galloprovincialis*

[b] *Arthropoda* clade: *Penaeus monodon, Pandalus latirostris, Neocaridina denticulate*

[c] *Ostariophysi* clade: *Ictalurus punctatus, Astyanax mexicanus, Sinocyclocheilus angustiporus, Sinocyclocheilus anophthalmus, Microphysogo biobrevirostris*

[d] *Euteleostei* clade: *Plecoglossus altivelis, Gasterosteus aculeatus, Tetraodon nigroviridis, Lateolabrax japonicas, Pundamilia nyererei, Fundulus grandis, Poecilia Formosa, Poecilia mexicana*

[e] *Actinopterygii* clade: *Ictalurus punctatus, Astyanax mexicanus, Sinocyclocheilus angustiporus, Sinocyclocheilus anophthalmus, Microphysogo biobrevirostris, Plecoglossus altivelis, Gasterosteus aculeatus, Tetraodon nigroviridis, Lateolabrax japonicas, Pundamilia nyererei, Fundulus grandis, Poecilia Formosa, Poecilia Mexicana. Anguilla japonica, Clupea harengus*

[f] *Chordate* clade: *Ictalurus punctatus, Astyanax mexicanus, Sinocyclocheilus angustiporus, Sinocyclocheilus anophthalmus, Microphysogo biobrevirostris, Plecoglossus altivelis, Gasterosteus aculeatus, Tetraodon nigroviridis, Lateolabrax japonicas, Pundamilia nyererei, Fundulus grandis, Poecilia formosa, Poecilia mexicana. Anguilla japonica, Clupea harengus, Protopterus annectens*

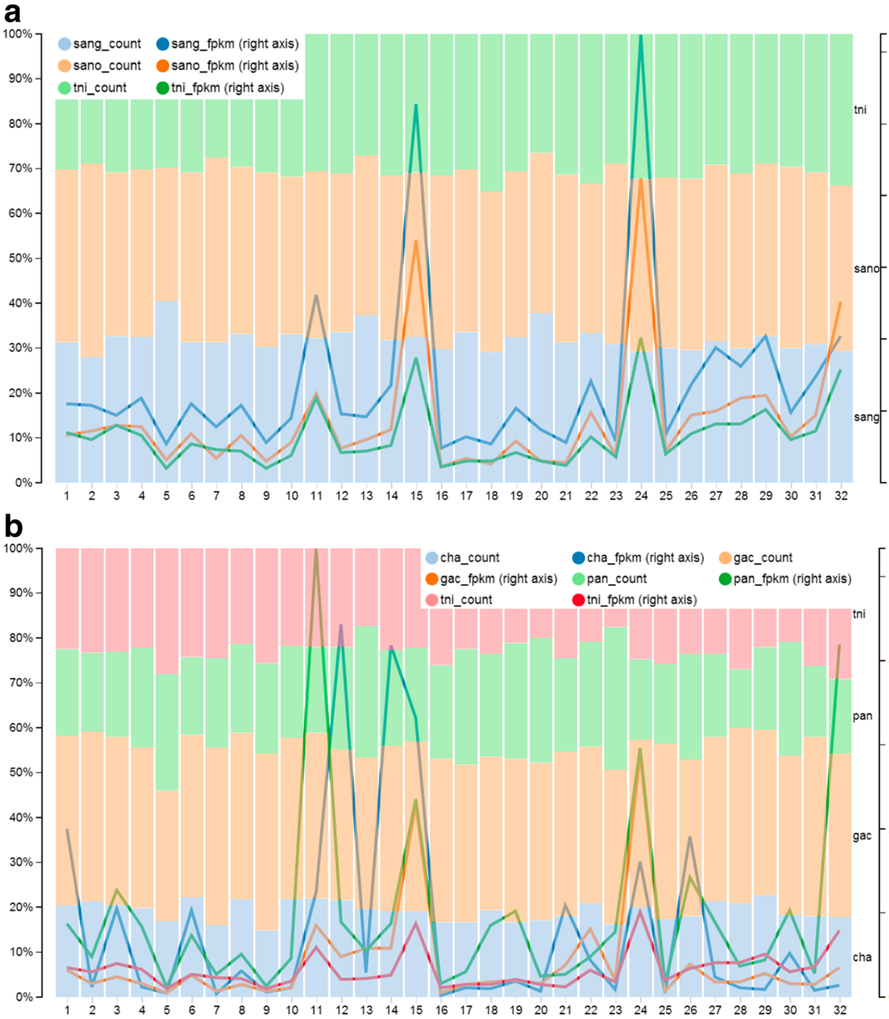


Fig. 5 Gene expression profiles panel. The dbATM provides a function for comparative analysis of different tissues, conditions or species. **a** All three species samples are taken from the brain, presenting similar expression profile trends. **b** All four species are taken from different tissue, showing distinct expression profiles. Abbreviation code for species: *cha, Clupea harengus*; gac, *Gasterosteus aculeatus;* pan, *Protopterus annectens;* sang, *Sinocyclocheilus angustiporus;* sano, *Sinocyclocheilus anophthalmus; tni, Tetraodon nigroviridis*

animals. The user-friendly web interface allows public access to valuable information for transcriptome assembly of each organism. The newly constructed homologous genes database and comparative analysis system also give aquatic researchers insights into the fields of eco-toxicity, animal physiology, comparative genomics and phylogenetic. dbATM also serves as an important repository for the aquatic animal transcriptomes by analyzing RNA-seq data. The assembly gene information provided could be a potentially valuable resource for designing expression microarrays to detect gene expression profiles across many conditions, thus facilitating studies in ecology and molecular biology. Given recent advances in NGS technologies, more non-model aquatic animal RNA-seq experiments will be deposited in the public domain.

Methods

System overview

First, RNA-seq data for 18 species of aquatic animals that were not well-annotated were collected from public resources such as SRA, ENA and DRA (the data from Illumina paired-end NGS sequencing platform were collected), with an additional four datasets produced for this study (Fig. 3 and Additional file 1: Table S1). Second, the quality of the raw reads from NGS was validated to trim and filter low quality reads. Third, the trimmed reads were then assembled through the designed combinational approach. Finally, genes functionally annotated by BLAST were compared against external databases (Fig. 3). All of the annotation results were soundly provided in the dbATM database.

RNA-seq dataset collection

We collected eighteen RNA-seq datasets from the NCBI Sequence Read Archive (SRA), European Nucleotide Archive (ENA), and DDBJ Sequence Read Archive (DRA). RNA-seq data deposited to these datasets were generated from Illumina paired-end NGS sequencing platform. All RNA-seq data collected in dbATM were reference genome-free (including scaffolds or contig status) in the NCBI taxonomy database. We also performed RNA-seq on four additional aquatic species. Additional file 1: Table S1 present detailed information for the 22 aquatic animal species profiled in our dbATM database.

Cleaning NGS read data

The NGS raw data for the RNA-seq in SRA format were converted to FASTQ format using the SRA-Toolkit v2.2.2 [35] and the FASTQ format reads were cleaned to increase read quality by FASTX-Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/). First, reads with adapters were removed. Second, the sequencing reads were scanned for quality at the reads tails. If the Phred quality scores were below 20 at the nucleotide end, the nucleotide sequences were removed. Third, if a read's average Phred quality score was below 20, the read was discarded. Fourth, reads were discarded if their length was less than 70% nucleotides after tail trimming and reads filtering. Finally, for paired-end datasets, if one paired read had been removed in a previous step, the other read was also removed to synchronize the read pairs.

De novo assembly

To ensure more complete assembly, we used three de novo assembly tools: Oases (v0.2.08, requiring Velvet v1.2.09) [36, 37], SOAPdenovo-Trans (release 1.02) [38], and Trinity (release 2013–02-25) [24] that based on the de Bruijn graphs. Previous reports suggest that coverage in transcriptomes assembly is not uniform. Higher *k*-mer lengths could obtain good quality assemblies for highly expressed transcripts. Lower *k*-mer lengths are used for poorly expressed transcripts [39]. This has been experimentally verified in model organisms [40]. We used multiple *k*-mer assemblies in Oases and SOAP-denovoTrans (for all species, $k = 21$, 29, 37, 45, 53, and 61, except for read lengths in *Fundulus grandis* smaller than 60, where $k = 21$, 27, 33, 39, 45, and 51). The insertion length of paired-end RNA-seq data was estimated using the observed-insert-length.pl program (included in the Oases package). Trinity was applied with default parameters. After the first step assembly in three assembler, we selected high read coverage and relatively long length as the representative transcript and conducted further assembly using CAP3 Version Date: 2007–10-15 [41] with an overlap length cutoff of 200 and overlap percent identity cutoff of 99 (–o 200 -p 99) [32]. To obtain comprehensive transcriptomes from various assemblers [33, 34], we used the CD-HIT-EST v4.6 [42] cluster tool with a sequence identity cutoff of 90% to merge results from Oases, SOAPdenovo-Trans, and Trinity.

Abundance estimation

To determine the expression level in each species, we calculated the relative abundance of previously assembled results by RSEM (v1.2.3) [43] requiring bowtie (v1.0.0) [44] to accurately quantify transcript expression with or without reference genomes. The relative abundances of transcripts were measured in unit of normalized reads count aligned on de novo assembly, FPKM (Fragments Per Kilobase of transcript per Million mapped reads) [45]. After de novo assembly and expression calculation, we select FPKM > = 1 and gene length > = 300 as a cutoff for further analysis such as functional annotation and comparative analysis.

Functional annotation

BLASTX (v2.2.28+) [46] with E-value cutoff of 1.0e-5 were used to homology search the NCBI non-redundant protein database (nr, v2012–11-19) for the assembled genes. The best blast hit protein was annotated to the representative gene and protein name. The gene annotation by BLASTX indicates that some transcripts were mapped to the same gene or gene family. To reduce redundancy, we used the NCBI UniGene approach which clusters the same gene to a group. The results for the BLASTX annotation genes were mapped to the Gene Ontology database. We also extracted the open reading frames (ORFs) from all genes according to the best BLASTX match hit results. To identify genes involved in KEGG pathway, we used the KEGG Automatic Annotation Server, KAAS [47]. Finally, we identified the putative protein domains on translated protein sequences extracted from genes by aligning to CDD (conserved domain and protein classification) [48] using RPS-BLAST.

Comparative analysis

Based on a comparative analysis, different species were grouped into three main categories. First, we used OrthoMCL [31] with an E-value cutoff of 1.0e-10 [49] to perform a similarity search orthologous by computational approach, and then defined the orthologous groups using the representative gene name or description, and called each group a homologous gene group, indicating that the different species shared a common ancestor. Second, we compared the abundant gene expressions among the different species using the functional annotation KEGG results and the D3: Data-Driven Documents [50] to provide enhanced graphical representations. Considerable amounts of annotation information was obtained from the KEGG pathway category [28] with average of gene expressions (FPKM) in different species. In Fig. 5, the bar is the read count aligned to each KEGG pathway category, and the line is the average FPKM of the genes mapped to this pathway. In the KEGG pathway category, we excluded Human Diseases, Drug Development, and Global Maps from the metabolic category. The graph visualization of comparative analysis represents the gene expression profile in different tissues, conditions or species using KEGG pathway. For example, the Fig. 5a illustrates the samples from the brain, so the expression profile trends (gene count and average FPKM) are similar. In contrast, the Fig. 5b displays the samples from different tissue, so the expression profile trends are distinct. Third, a web interface of BLAST against dbATM to allow the user using blastp/blastn to search protein/nucleotide data by pasting the query sequence or by submitting a file in FASTA format.

Additional file

Additional file 1: Figure S1. The cumulative publications of related articles of aquatic animal in PubMed. **Table S1.** RNA-seq datasets from twenty-two aquatic animals were analyzed in dbATM. **Table S2.** Statistics of homologous genes in each clade and their gene lists. (PDF 280 kb)

Abbreviations

COG: Clusters of Orthologous Groups; DRA: DDBJ Sequence Read Archive; ENA: European Nucleotide Archive; FPKM: Fragments per kilobase of transcript per million mapped fragments; KEGG: Kyoto Encyclopedia of Genes and Genomes; NGS: Next-generation sequence; RNA-seq: RNA sequencing; SRA: Sequence Read Archive

Acknowledgements

Not applicable.

Funding

This work is supported by the Ministry of Science and Technology, Taiwan (MOST) [MOST 106–2627-M-009-002, MOST 106–2319-B-400-001, MOST 106–2633-B-009-001]. This work is particularly supported by "Aiming for the Top University Program" of the National Chiao Tung University and Ministry of Education, Taiwan [MOHW106-TDU-B-212-144005 and MMH-HB-10602]. Funding for the publication of this article was provided by MOST grant [MOST-106-2627-M-009-002].

Availability of data and materials

The RNA sequencing data sets generated in this study have been deposited in the NBCI SRA database under accession SRR1185326 for *Anguilla japonica*, SRR1185341 for *Microphysogobio brevirostris*, SRR118532 for *Neocaridina denticulate*, and SRR1185333 for *Planorbarius corneus*. Others accession codes of RNA-seq data download from public domain database were listed in Additional file 1: Table S1. The database is now publically accessible through the URL http://dbATM.mbc.nctu.edu.tw/.

About this supplement

This article has been published as part of *BMC Genomics* Volume 19 Supplement 2, 2018: Selected articles from the 16th Asia Pacific Bioinformatics Conference (APBC 2018): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-19-supplement-2.

Authors' contributions

CHC, CDH and HDH conceived and designed the experiments. CHC, HYH, WCH, SDH, CYL, YHC, YCL, WYH analyzed the datasets. CHC, HYH, SDH, CYL, YHC establish the database. CHC, HYH, SDH, CYL, YHC, YCL, WYH, WCH, MLL, YCC, CDH and HDH wrote or revised the paper. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan. [2]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan. [3]Department of Bioscience Technology, Chung Yuan Christian University, Chungli 320, Taiwan. [4]Institute of Molecular Medicine and Bioengineering, National Chiao Tung University, Hsinchu 300, Taiwan.

Published: 9 May 2018

References

1. Balmford A, Bruner A, Cooper P, Costanza R, Farber S, Green RE, Jenkins M, Jefferiss P, Jessamy V, Madden J, et al. Economic reasons for conserving wild nature. Science. 2002;297(5583):950–3.
2. Schwartsmann G. Marine organisms and other novel natural sources of new cancer drugs. Ann Oncol. 2000;11(Suppl 3):235–43.
3. Naylor RL, Goldburg RJ, Primavera JH, Kautsky N, Beveridge MC, Clay J, Folke C, Lubchenco J, Mooney H, Troell M. Effect of aquaculture on world fish supplies. Nature. 2000;405(6790):1017–24.
4. Cabello FC. Heavy use of prophylactic antibiotics in aquaculture: a growing problem for human and animal health and for the environment. Environ Microbiol. 2006;8(7):1137–44.
5. Lu XJ, Hang XY, Yin L, He YQ, Chen J, Shi YH, Li CH. Sequencing of the first ayu (Plecoglossus Altivelis) macrophage transcriptome and microarray development for investigation the effect of LECT2 on macrophages. Fish Shellfish Immunol. 2013;34(2):497–504.
6. Ji P, Liu G, Xu J, Wang X, Li J, Zhao Z, Zhang X, Zhang Y, Xu P, Sun X. Characterization of common carp transcriptome: sequencing, de novo assembly, annotation and comparative genomics. PLoS One. 2012;7(4):e35152.
7. Christie AE, Roncalli V, Lona PB, McCoole MD, King BL, Bucklin A, Hartline DK, Lenz PH. In silico characterization of the insect diapause-associated protein couch potato (CPO) in Calanus Finmarchicus (Crustacea: Copepoda). Comp Biochem Physiol Part D Genomics Proteomics. 2013;8(1):45–57.
8. Bai Z, Zheng H, Lin J, Wang G, Li J. Comparative analysis of the Transcriptome in tissues secreting purple and white nacre in the pearl mussel Hyriopsis cumingii. PLoS One. 2013;8(1):e53617.
9. An HS, Lee JW. Development of microsatellite markers for the Korean mussel, Mytilus Coruscus (Mytilidae) using next-generation sequencing. Int J Mol Sci. 2012;13(8):10583–93.
10. Minegishi Y, Henkel CV, Dirks RP, van den Thillart GE. Genomics in eels–towards aquaculture and biology. Marine Biotechnol. 2012;14(5):583–90.
11. Hu H, Bandyopadhyay PK, Olivera BM, Yandell M. Elucidation of the molecular envenomation strategy of the cone snail Conus geographus through transcriptome sequencing of its venom duct. BMC Genomics. 2012;13:284.
12. Roberts SB, Hauser L, Seeb LW, Seeb JE. Development of genomic resources for Pacific herring through targeted transcriptome pyrosequencing. PLoS One. 2012;7(2):e30908.
13. Garcia TI, Shen Y, Crawford D, Oleksiak MF, Whitehead A, Walter RB. RNA-Seq reveals complex genetic response to deepwater horizon oil release in Fundulus Grandis. BMC Genomics. 2012;13:474.
14. Kawahara-Miki R, Wada K, Azuma N, Chiba S. Expression profiling without genome sequence information in a non-model species, Pandalid shrimp (Pandalus Latirostris), by next-generation sequencing. PLoS One. 2011;6(10):e26043.
15. Kelley JL, Passow CN, Plath M, Arias Rodriguez L, Yee MC, Tobler M. Genomic resources for a model in adaptation and speciation research: characterization of the Poecilia Mexicana transcriptome. BMC Genomics. 2012;13:652.

16. Meng F, Braasch I, Phillips JB, Lin X, Titus T, Zhang C, Postlethwait JH. Evolution of the eye transcriptome under constant darkness in Sinocyclocheilus cavefish. Mol Biol Evol. 2013;30(7):1527–43.
17. Gross JB, Furterer A, Carlson BM, Stahl BA. An integrated Transcriptome-wide analysis of cave and surface dwelling Astyanax Mexicanus. PLoS One. 2013;8(2):e55659.
18. Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K, Takeshima H, Miya M, et al. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol Biol Evol. 2013;30(11):2531–40.
19. Werner T. Next generation sequencing in functional genomics. Brief Bioinform. 2010;11(5):499–511.
20. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671–82.
21. Pipes L, Li S, Bozinoski M, Palermo R, Peng X, Blood P, Kelly S, Weiss JM, Thierry-Mieg J, Thierry-Mieg D, et al. The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. Nucleic Acids Res. 2013;41(Database issue):D906–14.
22. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
23. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8.
24. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
25. Coordinators NR. Database resources of the National Center for biotechnology information. Nucleic Acids Res. 2013;41(Database issue):D8–D20.
26. Zeng V, Extavour CG. ASGARD: an open-access database of annotated transcriptomes for emerging model arthropod species. Database. 2012;2012:bas048.
27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000;25(1):25–9.
28. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32(Database issue):D277–80.
29. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997;278(5338):631–7.
30. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40(Database issue):D290–301.
31. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
32. Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics. 2013;14:328.
33. Garg R, Patel RK, Jhanwar S, Priya P, Bhattacharjee A, Yadav G, Bhatia S, Chattopadhyay D, Tyagi AK, Jain M. Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. Plant Physiol. 2011;156(4):1661–78.
34. Ghangal R, Chaudhary S, Jain M, Purty RS, Chand Sharma P. Optimization of de novo short read assembly of seabuckthorn (Hippophae Rhamnoides L.) transcriptome. PLoS One. 2013;8(8):e72516.
35. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database C. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40(Database issue):D54–6.
36. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086–92.
37. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.
38. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1(1):18.
39. Surget-Groba Y, Montoya-Burgos JI. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. 2010;20(10):1432–40.
40. Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, Rokas A. Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. Mol Biol Evol. 2009;26(12):2731–44.
41. Huang X, Madan A. CAP3: a DNA sequence assembly program. Genome Res. 1999;9(9):868–77.
42. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
43. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011;12:323.
44. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
45. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5.
46. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC bioinformatics. 2009;10:421.
47. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35(Web Server issue):W182–5.
48. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ, et al. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. 2013;41(Database issue):D348–52.
49. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res. 2001;11(12):2120–6.
50. Bostock M, Ogievetsky V, Heer J. D(3): Data-Driven Documents. IEEE Trans Vis Comput Graph. 2011;17(12):2301–9.