Sun *et al. GigaScience* (2016) 5:18
DOI 10.1186/s13742-016-0124-7

GigaScience

COMMENTARY Open Access

Fish-T1K (Transcriptomes of 1,000 Fishes) Project: large-scale transcriptome data for fish evolution studies

Ying Sun[1,2*†], Yu Huang[2†], Xiaofeng Li[2†], Carole C. Baldwin[3], Zhuocheng Zhou[4], Zhixiang Yan[5], Keith A. Crandall[6], Yong Zhang[2], Xiaomeng Zhao[2], Min Wang[2,7], Alex Wong[8], Chao Fang[2], Xinhui Zhang[2], Hai Huang[9], Jose V. Lopez[10], Kirk Kilfoyle[10], Yong Zhang[1], Guillermo Orti[6*], Byrappa Venkatesh[11*] and Qiong Shi[2,7,12*]

Abstract

Ray-finned fishes (Actinopterygii) represent more than 50 % of extant vertebrates and are of great evolutionary, ecologic and economic significance, but they are relatively underrepresented in 'omics studies. Increased availability of transcriptome data for these species will allow researchers to better understand changes in gene expression, and to carry out functional analyses. An international project known as the "Transcriptomes of 1,000 Fishes" (Fish-T1K) project has been established to generate RNA-seq transcriptome sequences for 1,000 diverse species of ray-finned fishes. The first phase of this project has produced transcriptomes from more than 180 ray-finned fishes, representing 142 species and covering 51 orders and 109 families. Here we provide an overview of the goals of this project and the work done so far.

Keywords: Fish-T1K, Fish, Transcriptome, RNA, Database, Biodiversity

Background

Ray-finned fishes (Actinopterygii) are the most diverse and abundant group of extant vertebrates. Thus far, approximately 32,900 fish species are recorded in FishBase [1]. Fishes encompass enormous variation in morphology, physiology and ecology. They are of great economic and medical significance as a primary source of protein for people worldwide, as a novel source of active ingredients in pharmaceuticals [2], and as evolutionary models for specific human diseases and conditions [3].

However, genomic resources for fishes are relatively underrepresented and published genetic data represent only a small fraction of extant fish species. So far, the whole genomes of only 38 fish species have been published (Additional file 1) and, although the number is growing (Additional file 2), searching the National Center for Biotechnology Information (NCBI)'s Sequence Read Archive (SRA) database for "fish AND transcriptome" yields 16,975 transcriptomes of only 242 fish species (Table 1). A lack of genomic resources for most fish species motivated us to generate large-scale fish transcriptome data and establish a database that may be used by scientists around the world. To this end, we initiated the "Transcriptomes of 1,000 Fishes" (Fish-T1K) project, an effort devoted to sequencing the transcriptomes of 1,000 different species of ray-finned fishes.

Fish-T1K

Fish-T1K is an international, collaborative and non-profit initiative officially launched by BGI and the China National Genebank (CNGB) in November 2013. The objective is to generate RNA-seq transcriptome sequences for 1,000 diverse fish species to help scientists unravel the mysteries of fish evolution, and pursue innovative approaches and strategies for addressing challenges in

* Correspondence: ying_sun09@icloud.com; gorti@email.gwu.edu; mcbbv@imcb.a-star.edu.sg; shiqiong@genomics.cn
†Equal contributors
[1]State Key Laboratory of Biocontrol, Institute of Aquatic Economic Animals and Guangdong Provincial Key Laboratory for Aquatic Economic Animals, School of Life Sciences, Sun Yat-Sen University, Guangzhou 510275, China
[6]Department of Biological Sciences, The George Washington University, Washington, DC 20052, USA
[11]Institute of Molecular and Cell Biology, A*STAR, Singapore 138673, Singapore
[2]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI, Shenzhen 518083, China
Full list of author information is available at the end of the article

Table 1 List of fish species with published transcriptome data in NCBI's SRA, and those generated by Fish-T1K

Order	No. of species in SRA	No. of species in Fish-T1K	No. of new species generated by Fish-T1K
Cypriniformes	42	5	3
Cyprinodontiformes	33	2	0
Perciformes	21	9	9
Cichliformes	15	2	2
Salmoniformes	14	0	0
Order-level *incertae sedis* in Eupercaria	9	2	2
Pleuronectiformes	9	2	1
Osteoglossiformes	8	4	2
Siluriformes	8	9	8
Clupeiformes	6	1	1
Syngnathiformes	6	5	5
Gymnotiformes	5	2	1
Acipenseriformes	4	1	1
Anabantiformes	4	4	4
Anguilliformes	4	4	3
Centrarchiformes	4	4	3
Scombriformes	4	2	2
Beloniformes	3	2	2
Characiformes	3	6	6
Gadiformes	3	1	1
Order-level *incertae sedis* in Ovalentaria	3	5	5
Tetraodontiformes	3	4	4
Carangiformes	2	2	1
Amiiformes	1	1	0
Batrachoidiformes	1	1	1
Blenniiformes	1	4	4
Esociformes	1	0	0
Labriformes	1	3	2
Lepisosteiformes	1	2	2
Ophidiiformes	1	2	2
Osmeriformes	1	0	0
Pempheriformes	1	2	2
Polypteriformes	1	3	3
Spariformes	1	2	2
Synbranchiformes	1	2	2
Argentiniformes	0	1	1
Atheriniformes	0	2	2
Aulopiformes	0	2	2
Chaetodontiformes	0	1	1
Elopiformes	0	1	1
Ephippiformes	0	2	2
Galaxiiformes	0	1	1
Gobiiformes	0	3	3
Holocentriformes	0	3	3
Kurtiformes	0	2	2
Lepidogalaxiiformes	0	1	1
Lobotiformes	0	1	1
Lophiiformes	0	2	2
Mugiliformes	0	2	2
Order-level *incertae sedis* in Carangimorphariae	0	3	3
Order-level *incertae sedis* in Percomorpharia	0	12	12
Percopsiformes	0	1	1
Uranoscopiformes	0	1	1
Zeiformes	0	1	1
others (Chondrichthyes and Sarcopterygii)	17	/	/
All	242	142	128

fish breeding, disease control and prevention, seafood safety, and biodiversity conservation.

Through this project, an integrated biobank will be established, incorporating a high-level bio-repository and a large-scale transcriptome database. The biobank will collect and store fish genetic resources including vouchers and frozen tissues, DNA and RNA nucleotides, together with related sample information documented according to standard operating procedures (SOPs). A companion database, committed to being the world's largest database of fish transcriptomes, has already been established and provides access to the sequences via BLAST search.

The Fish-T1K consortium

More than 40 scientists from 25 institutions across seven countries are active members of the Fish-T1K project (Fig. 1; Additional file 3). The Steering Committee consists of six core consortium members who are recognized experts in ichthyology, taxonomy, bioinformatics, phylogenetics, and evolution. In addition to the head office at BGI in Shenzhen, China, we have also established a hub at the Smithsonian National Museum of Natural History (NMNH) in Washington DC, USA, to facilitate quality sample collection from North America.

Species selection

Fish-T1K proposes to sequence 1,000 different ray-finned fish species representing all the orders and major families [4], and filling important gaps in the phylogenetic tree.

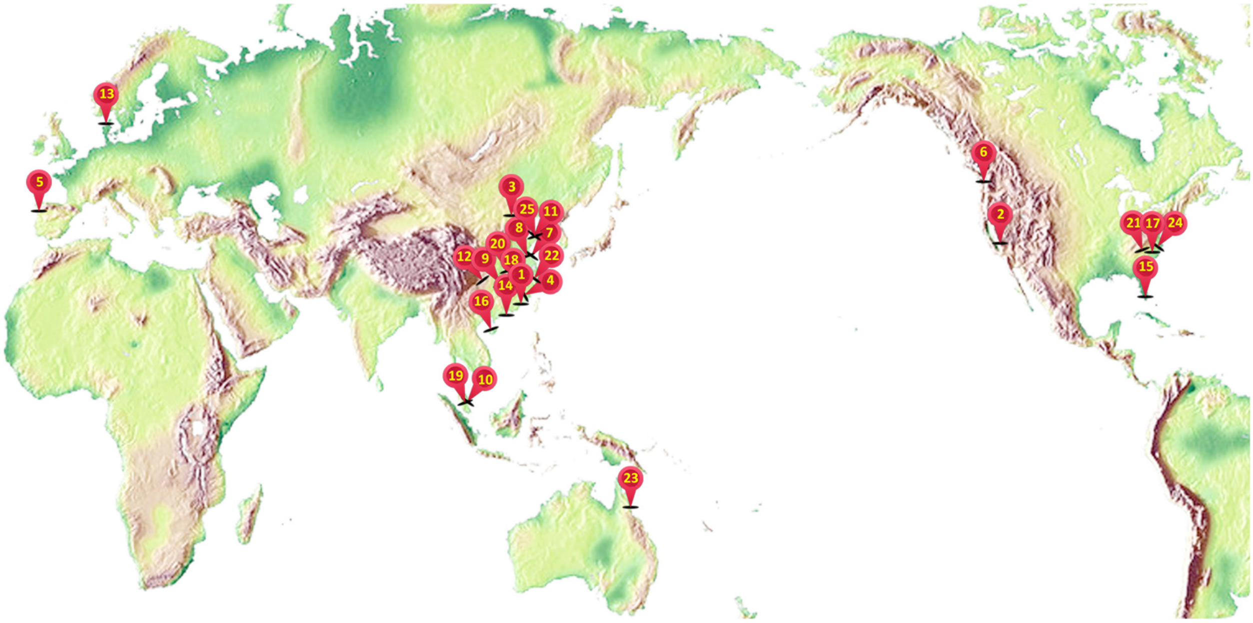

Fig. 1 Distribution of Fish-T1K Consortium members. See detailed information of these numbered institutions in Additional file 3

Species that are endangered, of great economic and medical significance, or exhibit extreme phenotypes will also be targeted. Candidate species will be decided based on their importance and availability, while the target number will be a compromise between scientific needs and practical limitations such as financial constraints and availability of specimens.

Subprojects

To maximize usage of these transcripts, Fish-T1K has launched several subprojects to address specific questions in fish evolution. The major research goal of Fish-T1K is to reconstruct a comprehensive molecular phylogeny of ray-finned fishes to further resolve and test existing phylogenetic hypotheses. Additional subprojects include analysis of the evolutionary genomics of fish venoms, evolution of the annual life cycle in killifishes, and adaptations related to marine-to-freshwater transitions/migration.

SOPs and best practices

In the past two years, the Fish-T1K Team has established a series of SOPs, approved by BGI's Institutional Review Board on Bioethics and Biosafety (No. BGI-IRB 15139), to ensure high quality sampling is achieved. Adhering to these SOPs means that all of our genetic resources, data and associated metadata are appropriately obtained, documented, and stored, which is helpful in establishing and optimizing standards common to large-scale transcriptome and genome sequencing projects.

Transcriptome data from multiple tissues of five fishes were generated as a pilot quality control test (Additional file 4). Accordingly, total RNA is now routinely extracted from gills and other tissues of interest, and approximately 3.5 Gb of raw data are generated for each sample. Clean reads are assembled *de novo* into contigs with SOAPdenovo-Trans (v1.3) [5], and the final assembled transcripts are used for annotation, ortholog prediction and other analyses.

Current RNA sequencing progress

The Fish-T1K team has established a collaborative global network for collecting specimens. As of January 2016, 7,000 high quality fish samples were collected from Australia, the Caribbean, Denmark, Singapore, the UK, USA, and many places in China such as the Tibetan Plateau, Sanya, and the Yellow Sea. From these 7,000 samples, RNA samples were extracted from 142 ray-finned species covering 51 orders and 109 families, and around 180 transcriptomes have been produced (Table 1; Additional file 5). Meanwhile, more RNA samples from other species are being isolated and sequenced.

Website and database

The official Fish-T1k website [6] is equipped with a database for BLAST search. The website provides detailed information about the Fish-T1K project, and particular sample information (RNA quality, sample provider, etc.) and data quality (raw data size, scaffold size and number, etc.) are presented in the database. Users can access the BLAST tool and download sequences of interest. Data will be uploaded periodically as sample collection and transcriptome sequencing progresses.

Data sharing policy and data availability

All sequences generated from Fish-T1K will be deposited in NCBI and GigaDB in addition to the Fish-T1K database, following the Fort Lauderdale rules [7] and Toronto International Data Release Workshop guidelines [8], and will be released at least in the time of publication of any resulting papers. We plan to peer review and publish the SOP and method papers, will be published and we're expecting publications for some of the ongoing subprojects are also expected in one the coming year or sooner.

Fish-T1K membership

All are welcome to participate in Fish-T1K and to propose new subprojects; these should address a major question in fish evolution and lead to (a) significant publication(s). Interested researchers can email fisht1k@genomics.cn with a brief proposal. The significance, question(s) to be addressed and fishes/tissues to be sequenced and analyzed should be included. On acceptance of a proposal, the lead scientist(s) will be asked to collect any fish tissues that are not already in our list, and to be in charge of analyzing and publishing the generated data.

Conclusions

Similar initiatives already exist to sequence the transcriptomes of large numbers of plants (1KP [9]) and insects (1KITE [10]). They have been well received and have been useful in establishing Fish-T1K. Although some progress has already been made, the Fish-T1K is at an early stage. We will continue to expand the scope of the project: in the first phase we aim to cover all orders, and all families in the second phase. More species will be added as required by subprojects. As the world's first large-scale transcriptome database exclusively for fish, Fish-T1K will greatly enhance the study of fish biology, and eventually contribute efforts towards global fish biodiversity conservation and the sustainable utilization of natural fish resources.

Additional files

Additional file 1: List of fishes with published genome data. (DOCX 30 kb)

Additional file 2: Number of fish species with newly published transcriptomes in SRA of the NCBI from 2009 to 2015. (TIF 4045 kb)

Additional file 3: List of the current Fish-T1K Consortium members. (DOCX 19 kb)

Additional file 4: Transcriptome data of five species for quality control. (DOCX 20 kb)

Additional file 5: List of fishes with transcriptome data generated by Fish-T1K. (XLSX 87 kb)

Abbreviations

1KITE: 1 K Insect Transcriptome Evolution; 1KP: 1000 Plants; Fish-T1K: transcriptomes of 1,000 Fishes; NGS: next-generation sequencing; RNA-seq: RNA sequencing; SOPs: standard operating procedures. SRA: Sequence Read Archive.; SRA: Sequence Read Archive.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YS drafted the original text with detailed input from YH, XL, GO, BV and QS. All authors have read and approved the final manuscript and participated in Fish-T1K.

Acknowledgements

We wish to acknowledge the Fish-T1K Consortium members, partners, advisors and supporters who have made Fish-T1K possible. We would also like to thank Junxing Yang from the Kunming Institute of Zoology of the Chinese Academy of Sciences, Qiang Lin from South China Sea Institute of Oceanology of the Chinese Academy of Sciences, Carol Stepien from the University of Toledo, Luiz Rocha from the California Academy of Sciences, Donald Stewart from the State University of New York College of Environmental Science and Forestry, and Andrew Thompson from The George Washington University for their tremendous support with sample collection.

Funding

This work was supported by China 863 Projects (No. 2012AA10A407 & 2014AA093501), Shenzhen and Hong Kong Innovation Circle Project (No. SGLH20131010105856414), Shenzhen Special Program for Future Industrial Development (No. JSGG20141020113728803), Shenzhen Special Program for Bio-industry Development (Nos. HY20130205008 and NYSW20130326010014) and Special Project on the Integration of Industry, Education and Research of Guangdong Province (No. 2013B090800017). A grant from the Smithsonian Institution (Biogenomics/GGI) to C. Baldwin, G. Orti, and R. Betancur provides partial funding for RNA extractions at the NMNH (Washington, DC, USA).

Author details

[1]State Key Laboratory of Biocontrol, Institute of Aquatic Economic Animals and Guangdong Provincial Key Laboratory for Aquatic Economic Animals, School of Life Sciences, Sun Yat-Sen University, Guangzhou 510275, China. [2]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI, Shenzhen 518083, China. [3]National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA. [4]China Fisheries Association, Beijing 100000, China. [5]China National Genebank, Shenzhen, 518083, China. [6]Department of Biological Sciences, The George Washington University, Washington, DC 20052, USA. [7]BGI-Zhenjiang Institute of Hydrobiology, Zhenjiang 212000, China. [8]BGI-Hong Kong, Hong Kong 999077, China. [9]Sanya Science and Technology Academy for Crop Winter Multiplication, Hainan 572000, China. [10]Oceanographic Center, Nova Southeastern University, Fort Lauderdale 33004, USA. [11]Institute of Molecular and Cell Biology, A*STAR, Singapore 138673, Singapore. [12]College of Life Sciences, Shenzhen University, Shenzhen 518060, China.

Received: 10 December 2015 Accepted: 14 April 2016
Published online: 03 May 2016

References

1. Froese R, Pauly D. FishBase. 2015. Available at: http://www.fishbase.org/. Accessed 12 Apr 2016.
2. Han S, Sun X, Ritzenthaler JD, et al. Fish oil inhibits human lung carcinoma cell growth by suppressing ILK. Mol Cancer Res. 2009;7(1):108–17.
3. Albertson RC, Cresko W, Detrich HW, et al. Evolutionary mutant models for human disease. Trends Genet. 2009;25(2):74–81.
4. Betancur-R R, Wiley Ed, Bailly N, et al. Phylogenetic classification of bony fishes (version 3). DeepFin. 2015; Available at: http://deepfin.bio.ou.edu/. Accessed 12 Apr 2016.
5. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1:18.
6. Transcriptomes of 1,000 fishes. Available at: www.fisht1k.org. Accessed 12 Apr 2016.

7. The Wellcome Trust (Fort Lauderdale January 14–15, 2003). Sharing data from large-scale biological research projects: a system of tripartite responsibility. 2003; Available at: https://www.genome.gov/Pages/Research/WellcomeReport03.03.pdf. Accessed 12 Apr 2016.
8. Birney E, Hudson TJ, Green ED, et al. Prepublication data sharing. Nature. 2009;461(7261):168–70.
9. Matasci N, Hung LH, Yan ZX, et al. Data access for the 1,000 Plants (1KP) project. GigaScience. 2014;3:17.
10. Misof B, Liu S, Meusemann K, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346(6210):763–7.