Data Article

Comprehensive data and workflow for mapping global proteome and post-translational modifications in Indian Major Carp, *Labeo rohita*

Mehar Un Nissa [a], Anwesha Banerjee [b], Mukunda Goswami [c], Sanjeeva Srivastava [a,*]

[a] *Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India*
[b] *Presidency University, 86/1 College Street Kolkata 700073, West Bengal, India*
[c] *Central Institute of Fisheries Education, Indian Council of Agricultural Research, Versova, Mumbai, Maharashtra 400061, India*

ARTICLE INFO

ABSTRACT

We present the data for the global proteome and post-translational modification mapping of *Labeo rohita* (Rohu) which consists of mass-spectrometric (MS) data for 8498 proteins at 1% false discovery rate, which constitutes 26% of the total protein-coding sequences in Rohu. This data consists of deep proteomics of 17 normal tissues including eye, spinal cord, brain, male gonad, female gonad, gill, air bladder, gall bladder, gut, liver, heart, kidney, skin, scales, muscle, fin, spleen, as well as blood plasma and embryo of Rohu. The data from SRM-based targeted analysis to validate the presence of few key proteins is also included. Global post translational modification-based analysis (global PTM) was also performed in the studied tissues and its background data is also publicly accessible. This data and the web-based proteome map may aid applied and basic research endeavors in aquaculture to meet the food demands and nutritional security challenges of an increasing

* Corresponding author.
  E-mail address: sanjeeva@iitb.ac.in (S. Srivastava).

world population. The data here is related to the research article "Organ-based proteome and post-translational modification profiling of a widely cultivated tropical water fish, *Labeo rohita*" in the Journal of Proteome Research [1].

Specifications Table

Subject	Omics: Proteomics (Biological sciences)
Specific subject area	Proteome map of *Labeo rohita*.
Type of data	Table, Figure
How the data were acquired	Data was acquired using liquid chromatography-tandem mass spectrometry (LC-MS/MS) through an Easy-nLC 1200 nano-flow liquid chromatography system coupled with Orbitrap Fusion Tribrid mass spectrometer.
	SRM/MRM data was acquired using an HPLC system (Thermo Vanquish) connected with Triple Quadrupole Mass spectrometer (TSQ Altis Thermo)
	Comparative protein expression data was obtained through Proteome Discoverer analysis of the mass spectrometry raw data. The expression data is presented in a web-based portal www.fishprot.org/ The Django framework was used in designing this portal. Currently it allows the data visualization as heatmap across the studied tissue.
	The data for PTMs (Phosphorylation, Methylation and Acetylation) was obtained using PTMProphet tool in the Trans-Proteomic Pipeline.
Data format	Analyzed
Description of data collection	For discovery based proteomic data, 19 normal tissues were taken for protein, peptide and PTM profiling where FDR of 1% was considered. For targeted proteomic comparison, peak intensities obtained from MRM analysis were compared across 9 tissues.
Data source location	Institution: Indian Institute of Technology, Bombay.
	City/Town/Region: Mumbai 400076.
	Country: India.
	Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: 19.1334° N, 72.9133° E
Data accessibility	Dataset 1: MS raw data and the protein database (.FASTA) is available at the Proteome-Xchange Consortium identifier PXD026377.
	All msf files (.msf) obtained from Proteome Discoverer analysis are available at PRIDE under the identifier PXD027141.
	Dataset 2: Datasets for selected/multiple reaction monitoring (S/MRM) experiment at Panorama public-
	https://panoramaweb.org/rohuorganwisesrm.url
	All the data is freely available at
	https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00759.
	Fish Proteome Map portal www.fishprot.org/
Related research article	Nissa, Mehar Un, Nevil Pinto, Arijit Mukherjee, Panga Jaipal Reddy, Biplab Ghosh, Zhi Sun, Saicharan Ghantasala, Chetanya Chetanya, Sanjyot Vinayak Shenoy, Robert L. Moritz, Mukunda Goswami, and Sanjeeva Srivastava. 2022. "Organ-Based Proteome and Post-Translational Modification Profiling of a Widely Cultivated Tropical Water Fish, *Labeo Rohita*." *Journal of Proteome Research* 21(2):420–37. doi:10.1021/acs.jproteome.1c00759.

Value of the Data

- This data is of great significance to the scientific community and researchers who focus on basic biological or industrial (fisheries) research.
- The provided data on organ-wise proteome, methylome, acetylome, and phospho-proteome will help in exploring the role of proteins and post-translationally modified proteins in this food fish.
- This dataset provides a map of proteins expressed by a tissue which gives clues to their function and organization in the cell.
- The provided data can be further analyzed and compared with other studies to identify the protein targets for accessing other aspects such as eco-toxicological monitoring.
- Our data extends a thorough understanding of commercially and ecologically important fish Rohu, and can benefit researchers involved in basic research, on a fish model as well as researchers associated with aquaculture and food industry.

1. Rationale and Objectives

Aquaculture is one of the food industries with the fastest growth rates. However, progress in aquaculture research has been significantly hampered by the scarcity of multi-omics data for the majority of the cultivated aquaculture species. One of the economically significant aquaculture species is Rohu. The recent release of Rohu's complete genome sequence [2] has increased the demand for creating an equivalent proteome map. In order to achieve this goal, this dataset was created to offer a thorough organ-based protein and PTM map of different tissue samples for this species.

2. Data Description

We describe two datasets acquired to develop a global proteome and PTM map for Rohu, a significant aquaculture species. Dataset 1 consists of discovery proteomic data obtained using Orbitrap Fusion based LC-MS/MS through data dependent acquisition mode. Dataset 2 consists of targeted proteomic data obtained using Multiple reaction monitoring approach of mass spectrometry. Before performing these LC-MS/MS analyses, protein was extracted from different tissue samples, followed by SDS-PAGE based fractionation. Protein was digested using in-gel method. The peptides were taken first to generate the discovery proteomic data which was acquired through 289 mass spectrometry runs by injecting one microgram of peptide each time. This raw data (dataset 1) obtained was analyzed in two different tools to perform the comparative protein expression and PTM identification in all the tissue samples taken (Fig. 1). MRM data consisting of 54 raw files was acquired for a set of 45 proteins targets. The data presented here consists of three figures and three tables (given in supplementary). Table 1 (.docx) has the details of all the reagents and equipment used to acquire or analyze this dataset. Fig. 2 and Table 2 (.xls) shows the distribution of proteins in each tissue based on number of unique peptide identifications. Number of peptide identifications with zero missed cleavage are represented in Fig. 3 and Table 3 (.xls).
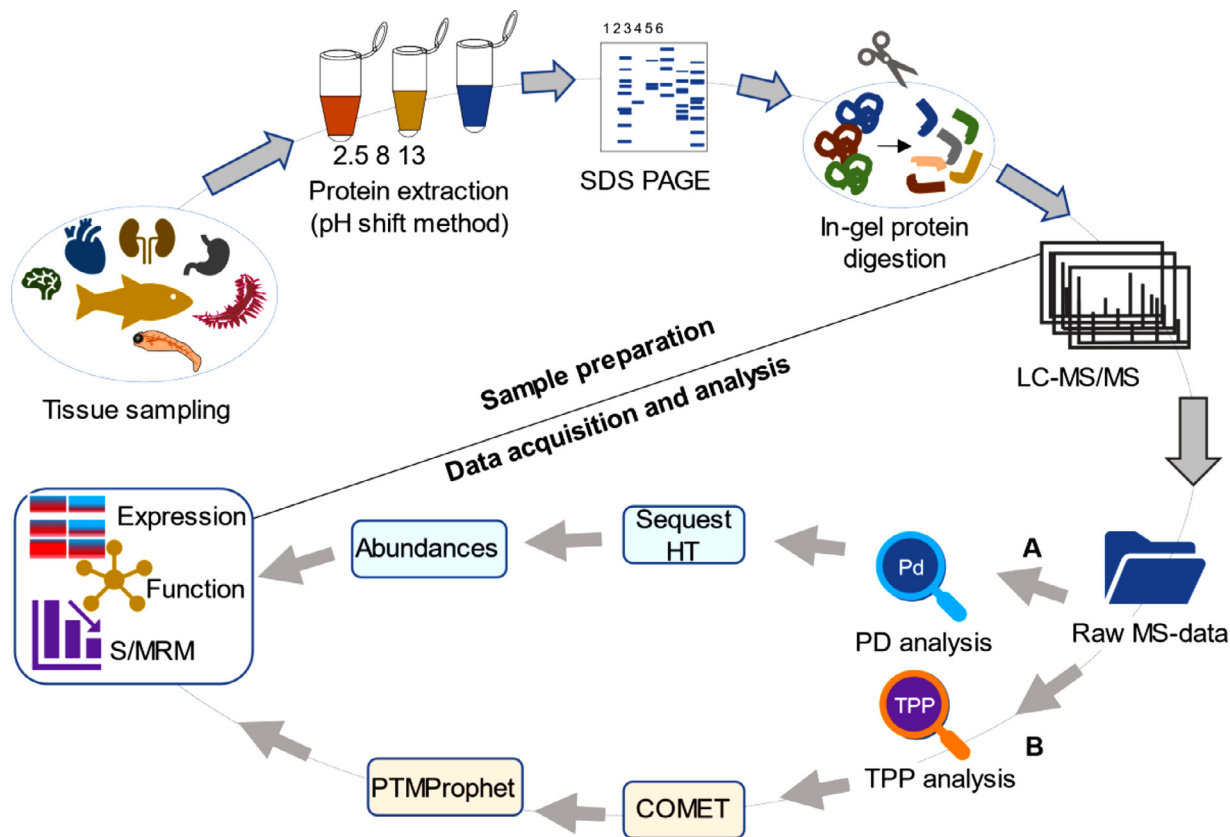
Fig. 1. Schematic representation of workflow for sample preparation and data analysis: Tissue samples were collected from fish and proteins were extracted using different methods as pH shift method. SDS-PAGE was performed to fractionate the sample. Gel pieces were processed for in-gel digestion to obtain peptides for Liquid chromatography tandem mass spectrometry (LC-MS/MS). Raw data obtained was analyzed using two different pipelines; **A**. Proteome Discoverer (PD) tool using Sequest HT for protein search to finally obtaining protein abundances and, **B**. Trans-Proteomic Pipeline (TPP) with COMET and PTM prophet tools to identify the post-translational modifications (PTMs) across the tissue samples. Data was further analysed to obtain comprehensive picture of protein expression and function. Validation experiment for trends of protein expression was performed using Selected/Multiple reaction monitoring (S/MRM).
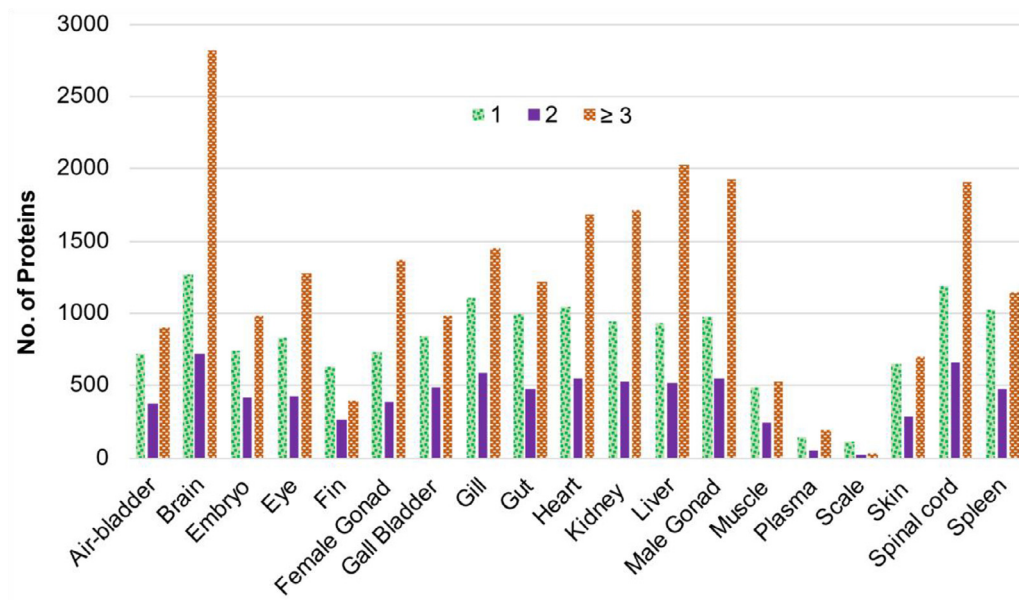
Fig. 2. Bar plot showing organ wise distribution of proteins based on number of identified unique peptides (1, 2 or >=3).
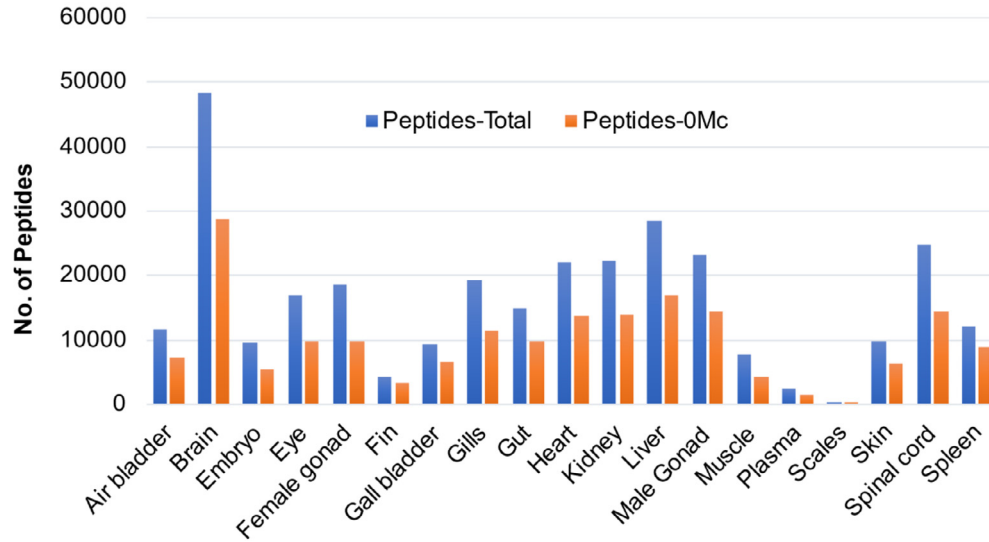
Fig. 3. Bar plot showing organ wise distribution of peptides based on missed cleavage. Peptides-Total –Number of total identified peptides, Peptides-0Mc- Peptides with 0 missed cleavage.

3. Experimental Design, Materials and Methods

3.1. Proteomic Sample Preparation

3.1.1. Acclimation of Fish and sample collection

Healthy fingerlings were acclimatized to laboratory conditions for a week. At a 2% body weight, fish were fed two times daily. These acclimation conditions included 24 h aeration and 12 h daylight. Water temperature was maintained at 28−30°C. Following this acclimation, five fingerlings were sacrificed for sample collection including heart, kidney, liver, air bladder, fin, eye, brain, gall bladder, spinal cord, skin, muscle, gut, skin, gill. Gonad tissues from male and female and blood plasma were sampled from adult fish weighing ∼1 kg. Additionally, embryo tissue at 4th day of fertilization was also taken for proteomic analysis.

3.1.2. Extraction of proteins from the collected samples

The pooled sample for each tissue was taken for protein extraction. For fifteen of the tissues (male gonad, female gonad, heart, kidney, liver, air bladder, eye, brain, gall bladder, spinal cord, muscle, gut, skin, gill, spleen), proteins were extracted using urea lysis buffer through a pH shift method [3] in 3 different pHs of 2.5, 8, and 13. Lysis buffer contained 8 M urea, 1 mM MgCl2, 50 mM Tris−HCl, and 75 mM NaCl. This method covers the proteome in-depth from the pooled samples while reducing the number of mass spectrometric runs. Tissue sample (25-100 mg) was immersed in lysis buffer (300 μL) followed by homogenization using bead beating with silica/zirconium beads. It was followed by centrifugation at 4°C with 8000 rpm for 15 min. Clear supernatant was collected and stored at −80°C. Trizol method was used to extract the embryonic proteins. Blood plasma from the female fish was taken directly for protein quantification and digestion.

3.1.3. Peptide Preparation using in-gel protein digestion

From each tissue, 30 μg of protein sample was run in SDS-PAGE and fractionated by excising at least six separate bands per lane. After in-gel digestion of all bands from all the samples, 302 fractions were obtained. Prior to protein digestion, gel bands were detained using a solution of two parts of acetonitrile (ACN) mixed with one part of ammonium bicarbonate (ABC). Proteins were in-gel reduced and alkylated with dithiothreitol (DTT) and iodoacetamide (IAA), respectively. Following this, digestion with trypsin ∼1:30 w/w trypsin-to-protein ratio was performed. Extraction of the digested peptides utilized an uphill gradient of acetonitrile and the peptides were vacuum dried before desalting. Peptides were quantified by the Scopes method and 1μg of the peptide for each fraction was taken for mass spectrometry.

3.2. Liquid chromatography tandem mass spectrometry (LC-MS/MS)

3.2.1. Discovery proteomic data acquisition using Data-Dependent Acquisition (DDA) mode

Following in-gel protein digestion, mass spectrometric based data was acquired in data dependent acquisition mode using nano Liquid chromatography linked with Orbitrap Fusion mass spectrometer. For all the samples, each fraction peptides were run individually except for spleen, skin, scales and plasma. For these tissues, 2-3 fractions of the same sample were pooled owing to low peptide quantity in individual fraction. For first three tissues, out of initial 18 fractions, 16, 13 and 15 fractions, respectively were run in mass spectrometer. Similarly, for plasma sample, mass spectrometric data was acquired for 8 fractions. Consequently, raw data was obtained for 289 mass spectrometric runs from all the samples. Peptides were run at a flow rate of 5 μL/min through the pre-analytical column. At a flow rate of 300 nL/min, over a 120 min analytical column, the sample was monitored. Solvents used include buffer B (80% Acetonitrile and 0.1% formic acid (FA)) and buffer A (MS Grade water and 0.1% formic acid) full scan range of peptides were subjected to liquid chromatography at a flow rate of 5 μL/min onto pre-analytical

column (Details in Table 1), followed by resolution on an analytical column (Details in Table 1) in which flow rate was 300 nL/min, over a 120 min gradient in solvent B. MS spectra were acquired where scan range of 375−1700 m/z was employed in DDA mode. The peptides were fragmented using High-energy collision dissociation (HCD) method. Value of 400,000 and 10,000 were set for AGC target in MS1 and MS2, respectively.

3.2.2. Protein Identification for global mapping of proteins

All raw files obtained after DDA based mass spectrometry were analyzed using Proteome Discoverer (PD) (Version 2.2, Thermo). NCBI database for *Labeo rohita* (Bio project: PRJNA437789) proteome was used for protein identification. This database consisted of translated coding sequences (CDs) based on gene prediction, locus tag IDs (prefix ROHU) and DDBJ CSS IDs (prefix RXN). This database corresponds to the *Labeo rohita* UniProt database where each of these CD are provided with UniProt ID (ProteomeID: UP000290572). The analysis was performed in label-free quantification (LFQ) mode and proteins were identified keeping a maximum of 2 missed cleavages. Static modification included cysteine carbamido-methylation whereas variable modifications included N-terminus acetylation and oxidation of methionine. False discovery rate (FDR) of 1% was kept for both peptides and proteins.

3.2.3. Identification of Post-translational Modifications (PTM)

The PTM based analysed was performed using PTMprophet tool in Trans Proteomic Pipeline (TPP) software (Version 5.2.0 Flammagenitus). The PTM search was performed to map three PTMs including phosphorylation at serine, threonine, and tyrosine (STY), methylation at the protein N-terminus, lysine, and arginine (nKR), and acetylation on lysine and N-terminus of the peptide. Firstly, all MS raw (.RAW) were converted into mzML followed by Comet search using NCBI database along with decoy sequences generated using decoy algorithm and the contaminant sequences from the common Repository of Adventitious Proteins (cRAP) database (http://www.thegpm.org/crap/). Along with the modifications as oxidation (+15.994915 Da) and cysteine carbamidomethylation (+57.021464 Da), selected PTM were added. It included methyl modification on nKR (+14.015650), phosphoryl modification on STY (+79.966 Da) and acetylation modification on nK (+42.010 Da: N-terminus and lysine). Comet outputs were taken for PTMprophet analysis. MS1 peak tolerance was 20 ppm and high accuracy MS2 was performed. FDR value of 0.0008 was selected for PTM peptide spectral match (PSM) for final filtration.

3.2.4. Targeted proteomic data acquisition using Selected/multiple reaction monitoring (S/MRM) mode

S/MRM data was acquired for 45 proteins across nine tissue samples including spleen, spinal cord (SC), female gonad (FG), brain, liver, heart, male gonad, eye and embryo. This data was compared with the protein expression trend obtained using DDA data. Transition list was created using Skyline software (version 21.1.1) [4] where the target list of proteins was imported. The final list consisted of 2166 transitions corresponding to 280 peptides run in six transition list (SRM methods). The list also includes nine transitions of heavy labelled synthetic peptide DIFT-GLIGPMK spiked in all the samples to monitor the consistency across the runs. The data was acquired in Triple Quadrupole Altis (Thermo) instrument coupled with an HPLC system (Vanquish, Thermo). Sample was run over a gradient of 10 min (Column details in Table 1) at a flow rate of 0.450 mL/min using reverse phase separation. A cycle time of 2 sec was set. Solvent A has milli Q water with 0.1% FA and solvent B has 20% ACN with 0.1% FA. Column temperature was set to 45°C.

3.3. Data analysis

3.3.1. Comparative protein expression analysis across the tissues

The abundance values acquired after PD search were considered as a measure of protein expression and utilized for comparative expression analysis across different tissues. The heatmaps,

to represent a landscape of protein expression, were plotted by hierarchical clustering, performed via Hierarchical Clustering Explorer software (Version 3.5). Log2 transformed abundances for all the quantified proteins excluding those with missing abundance values assorted into four datasets- 8127 total proteins, 26 shared proteins, 3960 total gonad and 1756 shared gonad proteins, were used as input list, normalized by scaling between 0 and 1, range set as column-by-column for the initial dendrogram. Both rows and columns for the total protein dataset were clustered based on Pearson Correlation Coefficient, UPGMA (average linkage) and no data partitioning to generate the hierarchically clustered protein expression heatmaps. For the shared proteins and the gonadal expression heatmaps, the distance measure was kept as Euclidean distance, the other parameters being the same as that of the total protein expression.

3.3.2. Gene Ontology and Functional analysis

Gene ontology (GO) based annotation was performed in PANTHER tool (Version15.0) [5] to obtain functional classification of identified proteome based on protein class, cellular components, biological processes and molecular function. For background proteome, LABRO was selected. Using Fisher exact test, an overrepresentation test was also performed in PANTHER where a p value threshold of 0.05 was used. For each tissue, pathway analysis was performed in KEGG mapper [6] where preferred gene names obtained from eggNOG analysis [7] were used and *Danio rerio* as reference. For eggNOG analysis, FASTA sequences (downloaded from UniProt [8]) were taken as input. Seed ortholog detection criterion was 0.001 and taxonomic scope was selected as "Actinopterygii". Under ortholog restriction, "transfer annotation from any ortholog" was chosen. Protein-protein interaction (PPI) networks for selected pathways of tissues (FDR ≤ 0.01) was performed in STRING tool [9] using *Danio rerio* as reference. For PPI enrichment, p value of ≤ 0.05 was selected.

3.3.3. Selected/multiple reaction monitoring (S/MRM) data analysis

Obtained data was analyzed in Skyline against spectral libraries obtained from DDA data (.msf files) and Prosit-based library [10] . Peaks were manually annotated based on peak shape, co-elution of product ions and match with spectral library. Peak area and intensity of peptides /proteins were compared across the samples.

3.3.4. Development of the Fish Proteome Map (FPM) Portal

The Fish Proteome Map (FPM) is a web-based portal that has been designed on the Django framework. It is made available as www.fishprot.org/ and allows the visualization of data obtained. Currently it has the data for all the organs with all identified proteins, peptides. For each query protein, comparative protein expression data can be visualized across the tissues. This is a dynamic platform and has the provisions of being extended to add more information on Rohu or even encompass other fishes.

Ethics Statement

CRediT Author Statement

Mehar Un Nissa: Conceptualization, Methodology, Data curation, Visualization, Writing – original draft, Writing – review & editing, Investigation; **Anwesha Banerjee:** Visualization, Writ-

ing – review & editing; **Mukunda Goswami:** Writing – review & editing, Investigation, Supervision; **Sanjeeva Srivastava:** Writing – review & editing, Investigation, Supervision.

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.

Data Availability

Fish Proteome Map (Reference Data) (Web based portal).
Proteomic profiling of Labeo Rohita; a widely cultivated fish (Reference Data) (PRIDE).
Organ wise proteomic profiling of Indian major carp, Labeo rohita (Reference Data) (PRIDE).
Organ wise targeted mass spectrometric analysis of proteins in Labeo rohita using SRM/MRM approach (Reference Data) (Panorama).

Acknowledgments

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2022.108746.

References

[1] M.U. Nissa, N. Pinto, A. Mukherjee, P.J. Reddy, B. Ghosh, Z. Sun, S. Ghantasala, C. Chetanya, S.V. Shenoy, R.L. Moritz, M. Goswami, S. Srivastava, Organ-Based Proteome and Post-Translational Modification Profiling of a Widely Cultivated Tropical Water Fish, Labeo rohita, J. Proteome Res. 21 (2022) 420–437, doi:10.1021/acs.jproteome.1c00759.

[2] P. Das, L. Sahoo, S.P. Das, A. Bit, C.G. Joshi, B. Kushwaha, D. Kumar, T.M. Shah, A.T. Hinsu, N. Patel, S. Patnaik, S. Agarwal, M. Pandey, S. Srivastava, P.K. Meher, P. Jayasankar, P.G. Koringa, N.S. Nagpure, R. Kumar, M. Singh, M.A. Iquebal, S. Jaiswal, N. Kumar, M. Raza, K. Das Mahapatra, J. Jena, De novo Assembly and Genome-Wide SNP Discovery in Rohu Carp, Labeo rohita, Front. Genet. 11 (2020) 386, doi:10.3389/fgene.2020.00386.

[3] V.K.R. Surasani, A. Tyagi, T. Kudre, Recovery of Proteins from Rohu Processing Waste Using pH Shift Method: Characterization of Isolates, J. Aquat. Food Prod. Technol. 26 (2017) 356–365, doi:10.1080/10498850.2016.1186130.

[4] B. MacLean, D.M. Tomazela, N. Shulman, M. Chambers, G.L. Finney, B. Frewen, R. Kern, D.L. Tabb, D.C. Liebler, M.J. MacCoss, Skyline: an open source document editor for creating and analyzing targeted proteomics experiments, Bioinforma. Oxf. Engl. 26 (2010) 966–968, doi:10.1093/bioinformatics/btq054.

[5] H. Mi, D. Ebert, A. Muruganujan, C. Mills, L.-P. Albou, T. Mushayamaha, P.D. Thomas, PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API, Nucleic Acids Res 49 (2021) D394–D403, doi:10.1093/nar/gkaa1106.

[6] M. Kanehisa, Y. Sato, KEGG Mapper for inferring cellular functions from protein sequences, Protein Sci 29 (2020) 28–35, doi:10.1002/pro.3711.

[7] J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S.K. Forslund, H. Cook, D.R. Mende, I. Letunic, T. Rattei, L.J. Jensen, C. von Mering, P. Bork, eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses, Nucleic Acids Res 47 (2019) D309–D314, doi:10.1093/nar/gky1085.

[8] R. Apweiler, UniProt: the Universal Protein knowledgebase, Nucleic Acids Res 32 (2004) 115D–1119, doi:10.1093/nar/gkh131.

[9] D. Szklarczyk, A.L. Gable, K.C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N.T. Doncheva, M. Legeay, T. Fang, P. Bork, L.J. Jensen, C. von Mering, The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets, Nucleic Acids Res 49 (2021) D605–D612, doi:10.1093/nar/gkaa1074.

[10] S. Gessulat, T. Schmidt, D.P. Zolg, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghe, A. Huhmer, U. Reimer, H.-C. Ehrlich, S. Aiche, B. Kuster, M. Wilhelm, Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning, Nat. Methods. 16 (2019) 509–518, doi:10.1038/s41592-019-0426-7.