The Proteome Map and PeptideAtlas of a widely cultivated tropical water fish Labeo rohita: A resource for the Aquaculture Community

Mehar Un Nissa
Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

Jaipal Reddy
Indian Institute of Technology Bombay

Arijit Mukherjee
Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore, 119077, Singapore

Nevil Pinto
Central Institute of Fisheries Education, Indian Council of Agricultural Research, Versova, Mumbai, Maharashtra 400061

Biplab Ghosh
Regional Centre for Biotechnology, Faridabad, 121012, India

Zhi Sun
Institute for Systems Biology, Seattle, WA, 98109, USA

Saicharan Ghantasala
Indian Institute of Technology Bombay

Chetanya Chetanya
Indian Institute of Technology Bombay

Sanjyot Shenoy
Indian Institute of Technology Bombay

Trinica George
Central Institute of Fisheries Education, Indian Council of Agricultural Research, Versova, Mumbai, Maharashtra 400061

Robert Moritz
Institute for Systems Biology https://orcid.org/0000-0002-3216-9447

Mukunda Goswami
Central Institute of Fisheries Education, Indian Council of Agricultural Research, Versova, Mumbai, Maharashtra 400061

Sanjeeva Srivastava (✉ sanjeeva@iitb.ac.in)
Indian Institute of Technology Bombay https://orcid.org/0000-0001-5159-6834
Original Research Article

The Proteome Map and PeptideAtlas of a widely cultivated tropical water fish *Labeo rohita*: A resource for the Aquaculture Community

Mehar Un Nissa\(^1\), Panga Jaipal Reddy\(^2\#\), Arijit Mukherjee\(^{18}\#\), Nevil Pinto\(^{3}\#\), Biplab Ghosh\(^4\),
Zhi Sun\(^2\), Saicharan Ghantasala\(^1\), Chetanya Chetanya\(^1\), Sanjyot Vinayak Shenoy\(^5\), Trinica George\(^3\), Robert L. Moritz\(^2\), Mukunda Goswami\(^{3}\#\), and Sanjeeva Srivastava\(^{1}\#\)

\(^1\)Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India
\(^2\)Institute for Systems Biology, Seattle, WA, 98109, USA
\(^3\)Central Institute of Fisheries Education, Indian Council of Agricultural Research, Versova, Mumbai, Maharashtra 400061
\(^4\)Regional Centre for Biotechnology, Faridabad, 121012, India
\(^5\)Department of Chemistry, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India
\(^6\)Current address: Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore, 119077, Singapore

\#Contributed equally

\#\#Correspondence for proteomics work: Dr. Sanjeeva Srivastava, E-mail: sanjeeva@iitb.ac.in, Phone: +91-22-2576-7779, Fax: +91-22-2572-3480

\*\*Correspondence for fish work: Dr. Mukunda Goswami, E-mail: mukugoswami@gmail.com
Abstract

With the global consumption of fish outpacing population growth, aquaculture sector is facing challenges to address the rising demand of food and nutritional security. Integrative omics research provides a strong platform to understand the basic biology and translate this knowledge into sustainable solutions in tackling disease outbreak, increasing productivity thus ensuring food security. To further understand the complex biology of host-pathogen response and support the aquaculture effort, genome and proteome reference maps moving beyond simple sequence information of cultivated fish species will accelerate research and translation of quality products for food industries. Towards this end, we have performed an extensive proteomics-based investigation of *Labeo rohita*, one of the economically important fish species produced in world aquaculture. Deep proteomic profiling of 17 histologically normal tissues, plasma and embryo provided mass-spectrometric evidence for 6015 high confident canonical proteins at 1% false discovery rate. Tissue enriched expression of several biologically important proteins was validated using targeted proteomics with high quantitative accuracy. We characterised the global post translational modifications (PTMs) in terms of acetylation (n-terminus and lysine), methylation (n-terminus, lysine and arginine) and phosphorylation (serine, threonine and tyrosine) to present a comprehensive proteome resource. An interactive web-based portal was developed to support the *Labeo rohita* PeptideAtlas, (www.peptideatlas.org), a unique community resource for mass spectrometry-based peptide/protein evidence in fish. This draft proteome map of *Labeo rohita* would advance basic and applied research in aquaculture to meet the most critical challenge of providing food and nutritional security to an increasing world population.

Keywords

*Labeo rohita*, proteomics, mass spectrometry, PeptideAtlas, aquaculture, Indian major carp
**Abbreviations**: AB- Air bladder, DDA- Data dependent acquisition, FDR- False Discovery Rate, FG- Female gonad, GB- Gall bladder, MG- Male gonad, PTM- Post translational modification, SC- Spinal cord, SDS-PAGE- Sodium dodecyl sulphate polyacrylamide gel electrophoresis, SRM-Selected Reaction Monitoring
**Introduction**

The average annual increase in global consumption of fish has outpaced population growth. Of the global animal protein consumption, 20% is met by fish suggesting the importance of fisheries in global food security and nutrition. India ranks second in global aquaculture production and Indian major carps (IMCs) contribute 90% of its aquaculture economy. *Labeo rohita* (*L. rohita*) is an IMC and among the top twelve finfish species produced in world aquaculture. Proteomic approaches have been applied in diverse areas to investigate developmental biology, physiology, disease mechanisms, impact of stress inducers and effects of dietary supplements on overall physiology of fish. Application of proteomics studies in zebrafish and *Xiphophorus sp.* revealed the role of phosphorylated Ezrin in gastrulation and peroxiredoxins in human melanoma. Proteomics can identify and explore sensitive and specific markers for assessing the quality of fish or fishery related product. The effect of pesticide mixtures and temperature have also been explored in goldfish (*Carassius auratus*). All these findings suggest the importance of proteomic characterization of fish would help to address basic biology to ecological, environmental and food related issues.

Proteome reference maps for many organisms have been generated using high resolution mass spectrometry such as for human and zebrafish. Recent publication of rohu genome reported a prediction of 26,400 protein coding genes. However, proteomics studies in *L. rohita* (rohu) are rare with most studies focusing on only a particular tissue in isolation. Therefore, to develop a draft map of rohu proteome; we performed an in-depth proteomic profiling of 17 histologically normal rohu organs, embryo and plasma using high-resolution high-mass accuracy mass-spectrometry. We identified a total of 6015 canonical proteins with 1% FDR, which is the first such extensive map for rohu proteome.

Post-translational modifications (PTMs) play an important role in governing many biological processes by affecting the dynamics and structure of proteins. Protein conformation changes
and PTM changes occurring during food processing or storage can sometimes affect overall allergenicity. Thus, making it highly relevant to study PTMs as many signalling cascades of proteins are activated through PTM. Few known PTMs include acetylation, phosphorylation, glycosylation, methylation, lipidation and ubiquitination. Both phosphorylation and acetylation have been found to regulate several meat quality attributes such as tenderness, firmness, colour stability and meat flavour in different animals. Significant changes in myofibrillar and sarcoplasmic phosphoproteome have been observed in low quality and normal bovine meat suggesting the role of phosphorylation in determining downstream meat quality. Lysine acetylation have been reported to regulate the effect of pre-slaughter stress on development of post-mortem flesh quality in porcine and mice muscle. Also, acetylation of Calcium signalling proteins and contractile proteins may help in regulation of rigor mortis. Along with phosphorylation and acetylation, some studies have reported the role of methylation in food quality. PTM site localization information is very limited in fish analyses especially in carps as only few studies are available. To address this gap, we performed global PTM analysis for acetylation (N-terminus and lysine), methylation (lysine and arginine) and phosphorylation (Serine, Threonine and Tyrosine) across all studied organs of L. rohita. Further, we developed a freely available L. rohita PeptideAtlas repository, for all observed peptides and PTM peptides that would be helpful for researchers for designing hypothesis driven or targeted experiments. Additionally, a web portal is prepared for the expression analysis of the identified proteome across the organs (www.fishproteome.org). This study could serve as a reference to advance search for specific gene products associated with muscle growth and fertility improvement as well as investigate system level alteration in any diseased or stressed conditions which can collectively determine the health of fish. We believe this extensive proteomic sequence and PTM information along with its genomic characterization
would serve as a community resource for the food and aquaculture industry to accelerate basic research and applications in industrial aquaculture.
Results

*L. rohita* organ based proteomic profiling

An extensive proteome catalogue of *L. rohita* was generated through in-depth proteomic profiling of 19 different tissue samples including 17 histologically normal tissues, blood plasma and four-day post fertilisation embryo (Fig. 1a). Proteins were extracted from the various tissues using urea lysis buffer or Trizol method (see methods) and run on 1D SDS-PAGE. Following the gel electrophoresis, fractionated proteins were sliced from the 1D gel into 2 mm slices and then proteolytically cleaved into peptides using trypsin by in gel digestion followed by mass spectrometry (Fig. 1b). Representative SDS-PAGE profile for all organs is shown in Fig. S1a. The workflow for protein identification and analysis has been depicted in Fig. 1c.

We employed the pH shift method\(^{26}\) for fifteen organs, using the Urea lysis buffer in order to obtain higher coverage (i.e., 30% increment as compared to only pH 8 buffer) (Fig. S1b-c). However, the frequency distribution remained same in terms of molecular weight (Fig. S1d).

A total of 9286 proteins (6015 canonical proteins, 667 indistinguishable representative proteins, 671 marginally distinguished proteins, 768 representative proteins and 1165 others) were identified using the combined non-redundant Uniprot database (downloaded on 16\(^{th}\) August, 2019) and NCBI protein sequence (Bio project: PRJNA437789) of *Labeo rohita* (Table S1). We considered only canonical proteins for further analysis. The highest number of proteins were identified from the brain while scales being the lowest (Table S1, Fig. 1d).

For functional annotation of identified proteins, ortholog analysis was performed using eggNOG database\(^{27}\) for all identified proteins across the organs, where the canonical proteins were mapped against orthologs of wide range of cellular processes and metabolic functions. Around 97\% of the mapped orthologs belong to *Actinopterygii*, the class of ray finned fishes and majority of them were linked to signal transduction mechanism (Table S1, Fig. S2).
Landscape of protein expression across the organs

Spectral count-based label-free quantification has been shown to be reproducible and particularly applicable for analysing large number of samples, as it is in our case\(^\text{28}\). Unsupervised-hierarchical clustering was performed to visualise the protein expression pattern across all the organs. These analyses revealed a large number of proteins showing broad range of expression among tissues while few proteins were found expressed in one particular tissue (Fig. 2a, Table S2). Proteins shared among all the samples were found to be mainly involved in cell cycle regulation, cellular respiration (e.g., glycolysis), protein folding and structural components of cell (e.g., cytoskeletal components) (Fig. S3 and Table S2).

Moreover, Actin cytoplasmic 1 and Glyceraldehyde-3-phosphate dehydrogenase reported as housekeeping genes in zebrafish\(^\text{29}\) have been identified in all the tissues of rohu in this study. This cohort could be considered as “housekeeping proteins”, given their broad expression among the tissues. Expression data is also provided through a web-based portal where expression of a particular protein can be visualised as a heatmap for the selected organs using accession IDs as input list (www.fishproteome.org). Nevertheless, one should note that some proteins may indeed be expressed in all tissues but may not appear in mass spectrometry for all organs owing to the sample quality, difficulty in extraction or inherent detection limits.

Further, we detected several proteins with tissue-enriched expression. Muscle tissue showed expression of many proteins belonging to calcium signalling pathway and regulation of actin cytoskeleton. Also, a few proteins involved in photo-transduction and retinol metabolism were identified in the eye; analogously proteins involved in bile acid biosynthesis and heart muscle contraction were detected in liver and heart respectively (Fig. 2b and Table S3). Organ wise enriched pathways are enlisted in Table S3. Many other proteins also showed tissue-enriched expression and these can be used as a starting point for validation and hypothesis generation in rohu and other related fishes.
Several popular plasma proteins such as apolipoproteins (Apo A and B), complement factors (C3, C5-9), coagulation factors (factor VIII, IX, X, XIII) and ceruloplasmins were detected in our data. In 96 hours post fertilization embryo, we were able to detect Fatty acid binding proteins (FABPs) which have been reported to play an important role in yolk-sac absorption. Few more proteins detected in rohu embryo which were previously reported in early embryonic stages in zebrafish such as adenylate kinase isoenzyme 1, ATP synthase, calsequestrin, creatine kinase B-type, creatine kinase M-type, beta crystallin A2, beta-crystallin B1, periostin isoform X1, actin cytoplasmic 1, Tubulin alpha and Tubulin beta.

Several hundred proteins mapped exclusively in gonads of either sex. For example, Zona pellucida sperm-binding 2-like protein (zp2), Zona pellucida sperm-binding 3-like protein (zp3) and Maternal DNA replication licensing factor mcm3-like isoform X1 (mcm3) were identified in female gonad (Fig. S4 and Table S4). Mcm3 is an important for DNA replication and cell cycle progression and zona pellucida sperm binding proteins, play significant role in oogenesis and fertilization. In parallel, we could map several proteins like Testis-expressed sequence 10-like protein (tex10), Tudor domain-containing 5-like protein (tdrd5), F-box only 22-like protein, Glomulin and ATP-dependent RNA helicases predominantly in male gonad. Tex10 is a membrane anchored protein reported to be involved in male fertility and acrosome reaction. Tdrd5 is reported to be involved in male reproductive functions. Tudor domain containing proteins play role in germ cell development and F-box only like protein has been reported to have biased expression at transcriptomic level in testis of humans; that might be an indication of conserved functions of respective genes in various animal groups.

Interestingly, we could link differential expression of certain proteins present in both the sexes to particular biological functions (Fig. 2c, Table S4). Overall, pathway analysis revealed that there is active protein synthesis and energy production in female gonad as majority of the proteins mapped to ribosome and oxidative phosphorylation. Importin subunit alpha-1-like
protein has shown comparatively higher expression in female gonad. The Importin family of proteins are responsible for transporting proteins from cytoplasm to nucleus and have been reported to be expressed predominantly in zebrafish ovaries; indicating the resemblance of its expression in gonads between rohu and zebrafish\textsuperscript{34}. Moreover, male gonad proteins were majorly mapped to spliceosome and RNA transport that are essential for RNA splicing required for maturation of spermatogonia in males\textsuperscript{40}. These findings supported the validity of our approach and the repertoire of information generated could drive hypothesis generation and research for exploring novel molecular mechanism.

**Peptide Atlas of *L. rohita***

The peptide and protein information acquired for all the organs was considered for building *L. rohita* PeptideAtlas (LPA), the first public repository for any industrial food fish. The LPA build has extensive information detailing observed peptides and respective proteins, spectral information for each peptide identified, and many link out resources to understand the identified proteins obtained from 19 different samples of rohu. The overall summary for LPA is shown in table S5. In brief, the current build contains more than 2.96 million identified peptide MS/MS spectra with additional information for a selection of PSMs at FDR level less than or equal to 0.08\% (i.e., 150781 distinct peptides at 0.18\% peptide level FDR). This peptide information corresponds to all identified proteins at less than 1\% protein level FDR. All tissues except muscle, fin, scales and plasma have contributed ~15,000-20,000 peptides and ~2000-3000 canonical proteins each to the build. Majority of the identified peptides are 11-20 amino acid long (Fig. S5). PeptideAtlas is a user-friendly portal for researchers who can access network-based information of related proteins from the rohu species of fish. An overview of the pipeline for building the LPA, including the peptide and protein summary can be found in Fig. 3a-c. The LPA provides a platform for obtaining detailed information of all identified
proteins and peptides that can be helpful for discovery experiments as well as designing SRM assays for *L. rohita*.

**Protein and peptide search in *L. rohita* PeptideAtlas**

For each protein entry, a dynamic page can be obtained to provide mass spectral interpretation and peptide modification information about the protein such as total observed peptides and a graphical representation of coverage of protein by each observed peptide. Additionally, all observed peptides are represented in a tabular format and ranked according to their empirical suitability score (ESS) observability score (EOS) (Fig. 4a). ESS is a measure of incidence of observing a protein/peptide in a given sample while EOS represents how much the observed peptide is suitable for the significant proteotypic detection of protein from which it was obtained. The protein view page also gives the information of all the tissues/sample in which the particular protein was detected.

For any observed peptide, a peptide view page presents all available information of respective peptide including its alignment to particular protein, genome mapping, modification sites. It also presents the peptide spectra in each sample where the peptide was observed (Fig. 4b).

**Extensive PTM catalogue of *L. rohita***

After analysing the high mass accuracy data-dependent analysis (DDA) mass spectrometry data using the PTMProphet tool\(^4\), we could detect 6870 acetylated peptides (1834 K acetyl, 3733 n Term acetyl and 1303 both K and nTerm), 3679 methylated peptides (1535 K methyl, 1180 R methyl, 471 n-Term methyl and 493 with 2-3 modified sites) and 826 phosphopeptides (524 S phospho, 138 T phospho, 41 Y phospho, 123 with 2-3 modified sites) (Table S5). Representative spectra for phosphorylation, methylation and acetylation are shown in figure S6a-c respectively.
Included are acetylated peptides and canonical acetyl proteins varies for each tissue as highest in kidney (2500, 758) followed by liver (1841, 729), brain (1136, 653), male gonad (1187, 543), spinal cord (994, 525), eye (1364, 508), heart (1039, 490), gill (679, 431), gut (694, 405), skin (622, 327), female gonad (529, 321), air bladder (564, 288), muscle (573, 254), gall bladder (334, 252), spleen (337, 242), embryo (302, 213), fin (203, 151), plasma (87, 68) and scales (44, 26). Further distribution of acetylated peptides for N-term acetylation and lysine acetylation is shown in figure 5a.

In case of methylation, highest number of modified peptides were identified in liver (639, 322) followed by kidney (629, 319), heart (530, 284), brain (502, 319), gut (495, 303), gill (480, 310), eye (475, 183), muscle (448, 165), male gonad (445, 193), spinal cord (387, 185), skin (319, 147), air bladder (267, 124), female gonad (267, 124), spleen (197, 161), plasma (195, 93) embryo (182, 127), gall bladder (149, 116), fin (118, 108) and scales (34, 20). Distribution of methylated peptides for lysine and arginine and N-Terminus can be seen in figure 5b.

However, number of phosphorylated peptides and canonical phosphoproteins were found highest in brain (227, 139) followed by spinal cord (132, 69), heart (115, 81), male gonad (103, 52), kidney (102, 77), eye (91, 76), muscle (89, 40), liver (87, 63), skin (54, 42), female gonad (54, 33), air bladder (53, 27), gill (38, 35), gut (34, 28), gall bladder (16, 18), spleen (15, 13), fin (13, 11), embryo (11, 10), plasma (5, 2) and scales (4, 3). Overall, Serine phosphorylation was found to be predominant across the organs (Fig. 5c). Several phosphoproteins identified in muscle tissue were earlier reported as markers for different quality parameters in fish or other animals. For example, we could map seven phosphoproteins that are related to muscle firmness\textsuperscript{42}, five related to quality change due to post-mortem storage temperature\textsuperscript{43}, six related to color stability in fleshPMI\textsuperscript{44}, four related to preslaughter stress effects\textsuperscript{19}, and three for muscle tenderness\textsuperscript{45}. Many of these could be potential markers in case of rohu as well. The mapped
phosphoproteins in muscle were majorly myofibrillar and sarcoplastic belonging to energy metabolism, glucose metabolism and muscle contraction (Fig. 5d, Table S5).

Additionally, in female gonad, three serine phosphorylation sites have been identified in Vitellogenin Ab (vtgab) which binds to vitellogenin receptor and further taken up by oocytes. Vitellogenins are precursors for egg yolk proteins involved in oocyte growth and early embryonic development. Insects are more likely to be phosphorylated at the polyserine region in vitellogenin. Another serine phosphoprotein Zygote arrest 1-like protein (zar1l) is identified in female gonad which corresponds to a vertebrate conserved oocyte-specific maternal-effect gene and have significant role during the oocyte-to-embryo transition. Two serine sites identified in ZP4 which plays significant role in oogenesis and fertilisation.

On the other hand, in the male gonad, phosphorylation is found in cytoplasmic polyadenylation element-binding 1 (CPEB1) isoform X3 which is a sequence-specific mRNA binding protein and reported to have role in translational control by cytoplasmic polyadenylation in gametogenesis and memory. A serine phosphoprotein ELAV-like protein identified in male gonad is reported to have role in spermiogenesis in mouse. ELAV family of proteins have been implicated in the regulation of mRNA stability, pre-mRNA splicing, and translational activation. Multiple serine phosphorylation sites were identified in another nucleic acid binding protein, Y-box-binding 2-A-like isoform X2 (ybx2). Ybx2 has shown restricted expression in testis in some studies and reported to be involved in stability and translation of germ cell mRNAs. Another Serine phosphoprotein ATP-dependent RNA helicase DDX19A identified in male gonad belongs to DEAD-box RNA helicase family of proteins that help in spermatogenesis.

Validation of tissue enriched proteins
Our DDA proteomics data revealed the comparative expression of 5702 proteins across the samples conforming to specialized pathways of the respective organs. Targeted proteomic approaches e.g., Selected Reaction Monitoring (SRM) generally used for validating specific peptides for a particular target protein (Fig. 6a), was employed to validate the enriched expression of a set of nineteen proteins among randomly selected three organs; eye, male gonad and female gonad and their comparative intensities across these three organs was found at par with the DDA data. Zona-pellucida sperm binding protein 2 like-isoform X2 has shown to have enriched expression in female gonad as compared to other two organs (Fig. 6b). Similarly, TUDOR domain containing-5 protein and cone cGMP-specific 2’-5’ cyclic phosphodiesterase subunit alpha showed maximum intensity in male gonad and eye, respectively (Fig. 6c and d). Comparative intensity of individual peptides for these proteins is shown in figure S7-S12. Details of the remaining proteins measured by SRM based validation can be found in Table S6.
Discussion

Providing food security to the burgeoning world population is one of the most critical challenges. The Aquaculture industry is one of the fastest growing food sectors to cater the rising demand of nutrition and food security. However, the lack of information at the multi-omics level for most of the cultivated aquaculture species has been a major hindrance for progress in aquaculture research\textsuperscript{52}. Recent availability of the whole genome sequence of \textit{L. rohita}, an economically important aquaculture species, has raised the need for generating an equivalent proteome map. To this end, we developed for the first time, an extensive draft of rohu proteome comprising 6015 canonical proteins at 1\% FDR.

Also, the peptide information obtained from our DDA data is presented as freely accessible web repository for proteomics called the PeptideAtlas (www.peptideatlas.org), which can be used by the scientific community to review and explore protein identifications for hypothesis-driven studies and obtaining peptide fragment coordinates for performing targeted proteomic experiments. For example, detection of allergens in a food product can be performed using targeted approaches such as SRM once the best peptides and peptide fragment coordinates are known for a desired targeted protein detection\textsuperscript{53,54}. With the availability of PeptideAtlas, optimal target peptides for a particular experiment can be determined without running many DDA experiments, thereby saving time and resource\textsuperscript{55}. Hence, the \textit{L. rohita} PeptideAtlas will be helpful for any such targeted quantitative experiments related to allergen detection, disease biomarker validation or any other aspect in multiple tissues of rohu as well as other related food fishes.

There are several downstream benefits of our findings that can address key issues like nutritional, environmental and health management for safety and quality enhancement of aquaculture species. Fish quality is determined by firmness, tenderness and cohesive flesh with good water holding capacity and many of these traits are determined by proteins’ nature and
properties. Therefore, proteomics appears to be of special interest to study fish and fish quality.\(^{56}\)

Pazos et al., investigated susceptibility of Cod muscle proteins to \textit{in vitro} metal catalysed oxidation and pointed out candidate proteins that play a major role in the determination of fish quality.\(^{57}\) Known candidate proteins like Cathepsin and Calpain involved in post-mortem tenderization\(^ {58,59}\) and parvalbumin-beta known as a major food allergen\(^ {60,61}\) have also been identified in our study. These could be further investigated for quality assessment and safety measures using targeted proteomics-based assays. Proteomic analyses have shown high potential in studying the effects of dietary components or supplements on fish physiology and/or overall growth\(^ {62,63}\) and such findings can be considered during feed formulation as well.

Quality control of food products and safety of human nutrition are important things to consider during any industrial application. Proteins are important component of food and changes in protein characteristics can occur during production and processing. Proteomic technology can be utilised for monitoring and characterising these protein changes and hence their post effects on human health can be evaluated. Food proteomics can be helpful for deciding quality markers so that safer and better food can be designed.\(^ {64}\) An approach in which supplementation of fish diets is done with the aim to reduce the expression of major fish allergens, parvalbumins, can be helpful in terms of human health and food safety\(^ {65}\). Hence, proteomics can be utilized to design optimal feed for fish stocks and hence better food for humans in turn. The rohu proteome, acetylome, methylome and phosphoproteome catalogue will fuel research in improving feed formulation through artificial feed with better protein supplements so that fish can attain maximum growth in the shortest possible time.
Tissue specific proteins identified across the organs conformed to specialized biological pathways that could be investigated for assay development and designing targeted proteomics experiments. Proteomic studies have been reported to be useful in sex determination in farmed sturgeon\textsuperscript{66}. We identified and validated several proteins - ZP2 like isoform X2, TDRD5 to be sex-specific, thus, broadening the possibility of easy sex determination in rohu samples. The popular proteins involved in specialized pathways of various organs in well-characterized vertebrates have also been mapped in \textit{L. rohita}. For example, proteins involved in primary bile acid biosynthesis, photo-transduction and cardiac muscle contraction and regulation of actin cytoskeleton were highly expressed in rohu liver, eye, heart and muscle respectively. For plasma and embryo, several known proteins have been detected which were earlier reported in other fishes or human.

PTMs such as phosphorylation and non-histone protein acetylation and methylation has been found to be involved in various biological processes\textsuperscript{67-69}. Phosphorylation have been related to maintenance of tenderness during post-mortem ageing which determine the quality of muscle or meat\textsuperscript{70,71}. Several proteins such as fast skeletal muscle light chain 2, myosin heavy fast skeletal muscle protein, enolase and titin reported to have role in muscle tenderness and overall flesh quality\textsuperscript{19,72} have been identified in our data as well. Specifically, in muscle tissue, we could identify a few phosphoproteins that can be potential markers for various quality parameters in fish for flesh quality. PTM characterization is also important for identification and characterization of allergenic proteins in food products. Some allergenic proteins have been identified in different food products, such as egg, milk, soybean, peanut and wheat etc\textsuperscript{15}. We could also identify several phosphoproteins involved in growth and reproductive functions such as those in gonads. Our protein methylome, acetylome and phosphoproteome data for various organs of \textit{L. rohita} can further help in characterising the role of PTMs in this food fish.
There is an increasing interest among the researchers to exploit rohu in eco-toxicological monitoring\textsuperscript{73,74}. However, lack of appropriate proteome database restricted the analysis to conserved and ubiquitous proteins\textsuperscript{75}. Our extensive proteome database and PeptideAtlas will bridge this gap to facilitate research on rohu in investigating physiological and adaptive responses to eco-toxicological challenges. Additionally, the study of proteome level alterations due to the climate change can be exploited to trace evolution\textsuperscript{76}. Glia maturation factor beta (GMFB), reported to have role in angiogenesis\textsuperscript{77} has been detected in rohu embryo. Several proteins earlier detected reported in zebrafish plasma have been identified in rohu, highlighting the possibility for rohu to be utilized in biomedical research.

Several non-canonical proteins identified in this study can further be brought into canonical category in future with additional studies on rohu and database improvement. With the availability of whole genome and transcriptome of rohu, this proteomic catalogue is a step forward to understand the physiology of this food fish and also to accelerate basic and applied research in relative fish species. An interactive web-based portal is made to support the \textit{L. rohita} PeptideAtlas, the first such community resources for an aquaculture fish extending the scope for validating protein targets without antibodies and data mining for hypothesis generation. Besides that, the expression of identified proteins across the organs can be analysed using the developed web based portal (www.fishproteome.org). Our findings thus corroborate the utility of proteomics to complement genome annotation and would serve as a blueprint to augment the need for fisheries research aimed at improving nutritional and food security.
Methods

Collection and acclimation of fish

Three-month old healthy *L. rohita* fingerlings weighing 10 ± 2 g were collected from Powarkheda Regional Centre of ICAR-CIFE, Madhya Pradesh, India. Fingerlings were acclimated to laboratory condition (12 h daylight, 24 h aeration, 28-30 °C water temperature, 10% daily water exchange and fed twice at 2% of body weight). After a week of acclimation, five healthy fingerlings were kept in an aquarium and starved for a day. All five fingerlings were euthanized and 19 different samples including 17 tissues (air bladder, brain, eye, fin, gall bladder, gill, gut, heart, kidney, liver, muscle, scales, skin, spinal cord, spleen, male and female gonad) were collected. Plasma and gonadal tissues were collected from healthy adult male and female rohu weighing 1000 ± 100 g. For analysis of embryo, four-day post fertilized embryos were taken. All tissue samples, embryo and plasma were stored at -80 °C until further use.

Optimization of protein extraction protocol for in-depth proteomic profiling

For performing protein extraction pooled sample for each tissue was taken forward. In order to get an in-depth proteome coverage from different organs of fish, we employed pH shift solubilization process (pH 2.5, 8 and 13)\(^78\) for fifteen tissues using Urea-lysis buffer (8M Urea, 50mM Tris-HCl, 1mM MgCl\(_2\) and 75 mM NaCl). In brief, 75-100 mg of sample was immersed in 300 µl of lysis buffer, sonicated thrice and homogenized using Zirconium/Silica beads (Cat. No. 11079110z). This extract was centrifuged at 8000 rpm at 4°C for 15 min and clear supernatant was taken into a fresh tube, stored immediately at -80°C until further use. Hard tissues such as scales, skin and fins were first ground to fine powder and processed similarly as stated above. Embryo was processed using Trizol method\(^79\) and plasma (female) was processed for in-gel digestion directly.
In gel digestion and peptide preparation

Around 30 µg protein each from all the samples was resolved onto SDS-PAGE, fractionated into six separate bands followed by in-gel digestion (Fig. S13). For plasma 11 fractions were done. The gel bands were de-stained, reduced and alkylated prior to trypsin (Pierce) digestion (~1:30 w/w trypsin to protein ratio). Digested peptides were extracted using an increasing Acetonitrile (ACN) gradient and vacuum dried prior to desalting (C18-sep packs column, Merck Millipore). Peptide quantification was performed using Scopes method and 1 µg of peptide was subjected to data-dependent analysis (DDA) using a nano scale RP-HPLC with high-resolution Orbitrap-Fusion mass-spectrometry.

Dada-dependant Analysis by Liquid Chromatography Tandem Mass-spectrometry (LC-MS/MS)

In-gel digested peptides were analysed by DDA on an Orbitrap-Fusion Tribrid mass-spectrometer connected to an Easy-nLC nano-flow liquid chromatography 1200 system. One µg of desalted peptides was loaded onto the pre-analytical column (100 µm x 2 cm, nanoViper C18, 5 µm, 100A; Thermo Fisher Scientific) at a flow rate of 5 µl/min. Peptides were resolved on analytical column (75 µm × 50 cm, 3 µm particle, and 100 Å pore size; Thermo Fisher Scientific) at a flow rate of 300 n/min over 120 min gradient in solvent B (80% ACN with 0.1% Formic acid (FA). Mass spectrometric acquisition was performed in DDA mode in full scan range of 375-1700 m/z of the Orbitrap mass analyser at a mass resolution of 60,000. Mass window was set to be 10 ppm with a dynamic exclusion duration of 40 s. All MS/MS spectra were acquired by HCD i.e., High energy Collision Dissociation method of fragmentation. AGC target was set as 400000 and 10000 for MS1 and MS2 respectively. Positive internal calibration was done using a lock mass of 445.12003 m/z.
Protein identification, TPP analysis and PeptideAtlas assembly

MSconvert 3.0.5533 tool\textsuperscript{81} was used to convert mass spectrometry data generated from the Orbitrap Fusion mass spectrometer (.RAW into .mzML file). Comet (2019.01 rev. 1)\textsuperscript{82} was used to search the converted mzML against \textit{L. rohita} protein database (ProteomeID-UP000290572) having 32,379 protein sequences from unassembled WGS sequence. An equal number of decoy sequences generated using “randomize sequences and interleave entries” decoy algorithm and contaminant sequences from common Repository of Adventitious Proteins, cRAP, database (http://www.thegpm.org/crap/) were added to the protein database. The following parameters were used for the data analysis using TPP suite: 20 ppm peptide mass tolerance, 0.05 m/z fragment ions bins tolerance and 0.0 m/z monoisotopic mass offset, semi-tryptic peptides, two allowed missed cleavages, carbamidomethylation of cysteine (+57.021464 Da) as static modification and oxidation of methionine and tryptophan (+15.994915 Da) as variable modifications. Protein identification was carried out using Trans-Proteomic Pipeline (TPP) V 5.2.0 Flammagenitus\textsuperscript{83}. PeptideProphet and iProphet tools were used to score for peptide spectral match (PSM) for individual files and the score unique peptides in combined PeptideProphet files. Finally, ProteinProphet tool was used for protein identification based on iProphet input and true identifications were selected at less than 1% FDR\textsuperscript{84-86}. The chimeric spectra were accessed by reanalysing the iProphet files using reSpect algorithm\textsuperscript{87}. In brief, reSpect search was performed on iProphet files by increasing the precursor mass tolerance to 3.0 Da. TPP analysis was performed as mentioned earlier and the process of reSpect and TPP analysis was repeated once. A minimum iProphet probability ≥ 0.0 was used for the reSpect search. PeptideAtlas processing pipeline was used to build PeptideAtlas by combining the iProphet results from regular TPP, reSpect search and the
PTM search results described in below section. The spectrum was filtered at variable probability to get constant peptide spectrum match (PSM) FDR of 0.08% for each experiment. The statistically significant results were organized in the “ROHU PeptideAtlas”, which is built and maintained by ISB at

https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=500

PTM (Post translational modification) analysis

PTM (Post translational modification) analysis was performed using PTMProphet tool, one of the integrated tools in Trans-proteomics pipeline. The proteomic data used for building PeptideAtlas was used for the identification of three PTMs; methylation at nKR (N-terminal, lysine and arginine), phosphorylation at STY (serine, threonine, tyrosine) and acetylation on lysine and N-terminus of the protein. The comet search parameters were same as used for the L. rohita PeptideAtlas build, except for the addition of methyl modification on nKR (+14.015650), phospho modification on STY (+79.966 Da) and acetylation modification on nK (+42.010 Da: n-N terminal) as variable modification. PTMProphet was performed on comet results using PTMProphet tool, where same modifications were set as done for comet search, MS1 peak PPM tolerance set as 20 ppm and pre-fill for selected MS2 accuracy as high. The methylation, phosphorylation and acetylation events were filtered with PTM peptide spectrum match (PSM) FDR of 0.08%.

Selected reaction monitoring (SRM) assay

In order to validate the tissue enriched expression of proteins using SRM assay (Fig. 6a), a total of 19 proteins from three organs- female gonad, male gonad and eye were selected. Skyline software was used for method optimization and final data analysis. Initially, all the proteins
showing unique expression in DDA data to either of these three tissues were taken for SRM validation of which only 19 proteins showed consistency in results. For generation of the transition list, a background proteome consisting of NCBI rohu database was used (Table S6). Initially Following optimization of parameters for SRM, a scheduled method containing 456 transitions corresponding to 86 peptides for the 19 selected proteins was used for the final run (Table S6).

An EASY-nLC 1200 LC set-up coupled with TSQ Altis was used for acquiring SRM data. The Acclaim PepMap 100 (75µm x 2cm nanoviper, C18 pre-column- Thermo Fisher Scientific) was equilibrated with 2 µL of the sample at a flow rate of 4 µL/min. The EASY-nLC system was equipped with PepMap RSLC C18- analytical column of 75 µm x 15 cm for reverse phase separation of the peptide mixture at a flow rate of 300 nL/min. The solvent system comprised of solvent A (0.1% FA), and solvent B (0.1% FA in 80% ACN, v/v) and the column temperature was maintained at 45°C throughout the run. A 45 min long gradient was set with the following conditions- start at 3% B followed by a linear increase 35% B for 35 min, 35% to 95 % B for 3 min, wash with 95% B for 2 min, 95% to 3% B for 3 min, 3% B for 2 min. The MS parameters were optimized using Skyline and a cycle time of 3 seconds was used to acquire the transitions in a scheduled run.

**Data analysis**

**Expression analysis**: The raw MS data was searched against the NCBI L. rohita protein sequence database (Bio project: PRJNA437789) using Proteome Discoverer 2.2 (Thermo Scientific, USA). Briefly, the parameters included maximum two missed cleavages, oxidation of methionine and N-terminus acetylation as variable modifications, Carbamido-methylation of Cys as static modification, FDR 0.01, mass error tolerance of 0.05 Da and 10 ppm for
fragment and precursor ions respectively. Abundance values obtained for each tissue were taken as measure of protein expression. For the tissues extracted using buffers of three different pHs, average abundances were considered for each protein (Table S2). Hierarchical clustering was performed using Hierarchical Clustering Explorer (Version 3.5) where unsupervised clustering was performed. Abundances values acquired after PD analysis were log2 transformed and scaled in 0-1 range for plotting. Expression analysis was performed for the canonical proteins and those proteins which have missing abundance values across the samples were excluded (and abundance values for 5702 canonical proteins was acquired). Dot plots for SRM were plotted in R taking peak area intensity for different peptides across the organs.

**Functional annotation, pathway analysis and interaction network:** FASTA sequences of total canonical and organ wise proteins were acquired from UniProt\(^88\) and taken as input for functional annotation in EGGNOG-mapper genome-wide functional annotation tool\(^89\) (http://eggnog-mapper.embl.de/) (Table S5). Various parameters filled in include taxonomic scope selected as Actinopterygii, seed ortholog detection criteria were set as 0.001, orthology restrictions selected as transfer annotation from any ortholog. For pathway analysis, preferred gene names obtained against query IDs for respective organ were taken as input in KEGG Mapper (https://www.genome.jp/kegg/tool/map_pathway2.html) taking *Danio rerio* as reference organism. For protein interaction networks, preferred gene names of respective pathways were taken as input in STRING tool (https://string-db.org/).
Data availability

All MS raw files are available in peptide atlas repository. Datasets for Selected reaction monitoring (SRM) based validation have been submitted to PASSEL with the identifier PASS01479.

Author’s Contributions

Concept and design: M.N, S.S, R.M and M.G

Maintenance and sampling: N.P, M.N

Method development and Data acquisition: M.N, N.P, S.G, A.M


Acknowledgments

This work was supported by Department of Biotechnology (BT/PR15285/AAQ/3/753/2015) Govt. of India to S.S and G.M. M.N was supported by University Grants Commission (UGC) and the US National Institutes for Health, National institute for General Medical Sciences under grant No. GM087221.), the Office of the Director IS10OD026936, the National Institute on Aging grant U19AG023122, and by NSF award 1920268 (R.L.M.). We thank Director General, Indian Council of Agricultural Research; Director, ICAR-Central Institute of Fisheries Education, Mumbai for the support and facility. We acknowledge MASS-FIITB at IIT Bombay supported by the Department of Biotechnology (BT/PR13114/INF/22/206/2015) for Mass-spectrometric data acquisition. Additionally, we would like to thank Mr. Deeptarup Biswas, Mr. Lakhanlal Meena and Mr. Aththiyan for worthy suggestions and support.
Competing interests:

The authors have declared that no competing interests.

Main Figures Legends

Figure 1 | Overall experimental workflow and description of obtained proteomic data

a. Diagrammatic representation of total organs included for proteomic analysis, b. Schematic workflow of sample preparation for mass spectrometric protein identification using in-gel digestion protocol, c. Workflow for protein identification in which proteins were identified using rohu Uniport and NCBI databases with two different pipelines (Proteome Discoverer and Tran Proteomic Pipeline), Data obtained from TPP was taken for PTMprophet analysis and building peptide Atlas. Confident proteins were taken forward for further analysis, d. Bar plot depicting unique and total proteins identified in each tissue sample.

Figure 2 | Landscape of protein expression across the organs

A-b. Heatmap showing total protein expression profile across all organs in which some proteins seen only in one tissue are highlighted for liver, eye, heart and muscle b. Protein interaction network for enriched pathways in liver, eye, heart and muscle b. Differential expression pattern of shared proteome in gonads (FG-Female gonad, MG-Male gonad).

Figure 3 | Overview of Peptide Atlas build

a. Workflow for building rohu Peptide Atlas, in which raw MS data is passed through several steps of trans-proteomic pipeline, b-c. Plots showing cumulative number of canonical proteins and peptides respectively contributed by each experiment. Height of the blue bar represents number of proteins/peptides identified in each experiment, height of red bar represents cumulative number of proteins/peptides and width of the bar (x-axis) represents the number of spectra identified (PSMs) for each experiment.

Figure 4 | Post-translational modification profile across the organs
a-c. Frequency of acetylation, methylation and phosphorylation across the organs respectively showing the number of modified peptides; acetylation sites include n-Terminal acetylation and lysine acetylation, methylation at lysine and arginine and phosphorylation sites include those on serine, threonine and tyrosine, d. Literature curated potential phosphoprotein markers for flesh quality traits identified in our proteomics data (muscle tissue)

Figure 5| Example of a protein search and peptide search in Rohu peptide Altas a. Out of several collapsible sections for protein, three are shown representing the overview of protein information, observed peptides highlighted in red font, additional information for each observed peptide respectively, b. Under peptide view, two sections for one of the observed peptides of the same protein are shown representing general information about peptide and respective spectra.

Figure 6| Targeted proteomics validation of tissue enriched proteins a. Schematic representation of experimental workflow for targeted validation using selected reaction monitoring (SRM) approach in which peptide separation by liquid chromatography, selection of a particular transition in triple quadrupole along with acquired peptide peak is shown b-d. comparative expression of selected proteins across female gonad (FG), male gonad (MG) and eye; x-axis showing the particular organ and y-axis showing the peak intensity value across the organ for each peptide (peptides of each protein are represented in different colors and expression for at least three peptides per protein is shown).

Supplementary figures legends:

Figure S1| Overall protein profile and protein frequency based on molecular weights a. SDS-PAGE profile of all organs showing different band pattern across (GB- gall bladder, FG- female gonad, MG- male gonad, SC- spinal cord, b. Number of proteins identified in fifteen organs where proteins were extracted with pH shift method with respective buffers, c-
d. Venn diagram showing the fraction of proteome identified using three pHs and frequency distribution of total number of identified proteins based on molecular weights.

**Figure S2** | An overview of phylogenetically annotated orthologs for the identified proteins, showing distribution of identified proteins mapped against each ortholog group.

**Figure S3** | Heatmap for shared proteome expression across all studied organs; highlighting few proteins with similar expression along with their description and functional details.

**Figure S4** | Heatmap for proteome expression for gonads highlighting few important proteins for male and female gonad.

**Figure S5** | Frequency distribution of identified peptides based on peptide length, where each node represents the experiment (sample).

**Figure S6** | Representative spectra for phosphorylation, methylation and acetylation.

**Figure S7-S8** | Comparative SRM intensity of individual peptides for ZP2 (Zona-pellucida sperm binding protein 2 like-isoform X2): Upper panel shows peak view for each transition of a peptide and lower bar graph shows the comparative peak area across three organs. Each color represents one transition.

**Figure S9-S10** | Comparative SRM intensity of individual peptides for TDRD5 (TUDOR domain containing-5 protein): Upper panel shows peak view for each transition of a peptide and lower bar graph shows the comparative peak area across three organs. Each color represents one transition.

**Figure S11-S12** | Comparative SRM intensity of individual peptides for GNAT2 (cone cGMP-specific 2’-5’ cyclic phosphodiesterase subunit alpha): Upper panel shows peak
view for each transition of a peptide and lower bar graph shows the comparative peak area across three organs. Each color represents one transition.

**Figure S13** Fractionation of samples for in gel digestion; a. Liver tissue sample, b. Plasma sample
References:


36


Figure 1

- **Figure 1a**: Diagram showing various fish organs such as spinal cord, kidney, muscle, skin, scale, fin, eye, gill, heart, liver, gall bladder, spleen, and gonad.

- **Figure 1b**: Flowchart illustrating tissue sampling, lysate preparation, SDS-PAGE, in-gel digestion, LC-MS/MS, and data analysis.

- **Figure 1c**: Diagram showing the process of analyzing raw MS data with TPP analysis, PD analysis, peptide atlas, PTMprophet, canonical proteins, PTM analysis, and expression analysis.

- **Figure 1d**: Bar chart showing the total proteins and uniquely identified proteins across different organs.
Figure 2

(a) Heatmap of gene expression across different tissues and conditions.

(b) Network diagrams for L-Liver, H-Heart, and E-Eye:
- Liver: Bile acid biosynthesis, Drug metabolism, Porphyrin metabolism.
- Heart: Cardiac muscle contraction, Adrenergic signaling in cardiomyocytes, Vascular smooth muscle contraction.
- Eye: Phototransduction, Retinol metabolism, Wnt signaling.

(c) Heatmap for FG and MG tissues with annotations for gene expression patterns.
Figure 4

a  Protein Search

PeptideAtlas Build: L. rohita 2020-07
Protein Name: A0A498MAL1
Show extra fields

Sequence:

Protein Coverage = 82% (66.8% of likely observable sequence)

Distinct Observed Peptides (95)

b  Peptide Search

Search Peptide Sequence for: IVAIDEIVHPK

Peptide Accession: PA04165733
Peptide Sequence: IVAIDEIVHPK
Best Probability: 1.00

File: 1612018_Mehar_Ingel_Heart_pH_13_F6.40323.40332.2.Scan.40323. Exp. m/z: 673.9057
Figures

Figure 1

Overall experimental workflow and description of obtained proteomic data. a. Diagrammatic representation of total organs included for proteomic analysis, b. Schematic workflow of sample preparation for mass spectrometric protein identification using in-gel digestion protocol, c. Workflow for protein identification in which proteins were identified using rohu Uniport and NCBI databases with two different pipelines (Proteome Discoverer and Tran Proteomic Pipeline), Data obtained from TPP was
taken for PTMprophet analysis and building peptide Atlas. Confident proteins were taken forward for further analysis, d. Bar plot depicting unique and total proteins identified in each tissue sample.

**Figure 2**

Landscape of protein expression across the organs A-b. Heatmap showing total protein expression profile across all organs in which some proteins seen only in one tissue are highlighted for liver, eye, heart and muscle b. Protein interaction network for enriched pathways in liver, eye, heart and muscle b. Differential expression pattern of shared proteome in gonads (FG-Female gonad, MG-Male gonad).
Figure 3

Overview of Peptide Atlas build a. Workflow for building rohu Peptide Atlas, in which raw MS data is passed through several steps of trans-proteomic pipeline, b-c. Plots showing cumulative number of canonical proteins and peptides respectively contributed by each experiment. Height of the blue bar represents number of proteins/peptides identified in each experiment, height of red bar represents cumulative number of proteins/peptides and width of the bar (x-axis) represents the number of spectra identified (PSMs) for each experiment.
Figure 4

Post-translational modification profile across the organs a-c. Frequency of acetylation, methylation and phosphorylation across the organs respectively 640 showing the number of modified peptides; acetylation sites include n-Terminal acetylation and lysine acetylation, methylation at lysine and arginine and phosphorylation sites include those on serine, threonine and tyrosine, d. Literature curated potential phosphoprotein markers for flesh quality traits identified in our proteomics data (muscle tissue)
Figure 5

Example of a protein search and peptide search in Rohu peptide Atlas a. Out of several collapsible sections for protein, three are shown representing the overview of protein information, observed peptides highlighted in red font, additional information for each observed peptide respectively, b. Under peptide view, two sections for one of the observed peptides of the same protein are shown representing general information about peptide and respective spectra.
Figure 6

Example of a protein search and peptide search in Rohu peptide Atlas. a. Out of several collapsible sections for protein, three are shown representing the overview of protein information, observed peptides highlighted in red font, additional information for each observed peptide respectively. b. Under peptide view, two sections for one of the observed peptides of the same protein are shown representing general information about peptide and respective spectra.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuppFiguresS1.pdf
- TableS1Totalproteins.xlsx
- TableS2Expression.xlsx
- TableS3Pathway.xlsx
- TableS4ExpressionGonads.xlsx
- TableS5PeptideatlasPTM.xlsx
- TableS6SRM.xlsx