Environment International 182 (2023) 108309

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Full length article

Multi-tissue proteogenomic analysis for mechanistic toxicology studies in non-model species

M.S. Lin [a,1], M.S. Varunjikar [b,1], K.K. Lie [b], L. Søfteland [b], L. Dellafiora [c], R. Ørnsrud [b], M. Sanden [b], M.H.G. Berntssen [b], J.L.C.M. Dorne [d], V. Bafna [e,*], J.D. Rasinger [b,*]

[a] Bioinformatics and Systems Biology Program, UC San Diego, San Diego, CA, United States
[b] Institute of Marine Research, Bergen, Norway
[c] Department of Food and Drug, University of Parma, Parco Area delle Scienze 27/A, 43124 Parma, Italy
[d] European Food Safety Authority, Methodological and Scientific Support Unit, Via Carlo Magno 1A, 43121 Parma, Italy
[e] Computer Science & Engineering and HDSI, UC San Diego, San Diego, CA, United States

ARTICLE INFO

Handling Editor: Adrian Covaci

Keywords:
Toxicogenomics
Proteomics, transcriptomics
Mechanistic toxicology
Non-model species
Chemical defensome
New approach methodologies

ABSTRACT

New approach methodologies (NAM), including omics and *in vitro* approaches, are contributing to the implementation of 3R (reduction, refinement and replacement) strategies in regulatory science and risk assessment. In this study, we present an integrative transcriptomics and proteomics analysis workflow for the validation and revision of complex fish genomes and demonstrate how proteogenomics expression matrices can be used to support multi-level omics data integration in non-model species *in vivo* and *in vitro*. Using Atlantic salmon as an example, we constructed proteogenomic databases from publicly available transcriptomic data and in-house generated RNA-Seq and LC-MS/MS data. Our analysis identified ~80,000 peptides, providing direct evidence of translation for over 40,000 RefSeq structures. The data also highlighted 183 co-located peptide groups that supported a single transcript each, and in each case, either corrected a previous annotation, supported Ensembl annotations not present in RefSeq, or identified novel previously unannotated genes. Proteogenomics data-derived expression matrices revealed distinct profiles for the different tissue types analyzed. Focusing on proteins involved in defense against xenobiotics, we detected distinct expression patterns across different salmon tissues and observed homology in the expression of chemical defense proteins between *in vivo* and *in vitro* liver systems. Our study demonstrates the potential of proteogenomic analyses in extending our understanding of complex fish genomes and provides an advanced bioinformatic toolkit to support the further development of NAMs and their application in regulatory science and (eco)toxicological studies of non-model species.

1. Introduction

New approach methodologies (NAM) including omics and *in vitro* approaches are currently contributing to the implementation of 3R (reduction, refinement and replacement) strategies in regulatory science and risk assessment (Krewski et al., 2020; Marx-Stoelting et al., 2023). Multi-level omics data from well-studied model-organisms such as rodents or zebrafish (*Danio rerio*), have frequently been used in mechanistic toxicity research (Bernhard et al., 2018; Mellingen et al., 2022, 2021; Rasinger et al., 2018, 2017, 2014). For non-model organisms such as (farmed) Atlantic salmon (*Salmo salar*), few studies exist which include NAMs when accessing the toxicity of xenobiotics (Bernhard et al., 2019; Hampel et al., 2015; Rasinger et al., 2022; Søderstrøm et al., 2022; Søfteland and Olsvik, 2022). The general lack of NAM derived data in these organisms limits the development and application of state-of-the-art quantitative adverse outcome pathways (qAOPs) and the associated design of quantitative structure–activity relationships (QSARs) for mechanistic toxicokinetic-toxicodynamic (TK-TD) modelling. It also hampers the efficient implementation of (quantitative) *in vitro – in vivo* extrapolations (QIVIVE).

* Corresponding authors.
E-mail addresses: msl043@eng.ucsd.edu (M.S. Lin), madhushri.shrikant.varunjikar@hi.no (M.S. Varunjikar), kaikristoffer.lie@hi.no (K.K. Lie), liv.softeland@hi.no (L. Søfteland), luca.dellafiora@unipr.it (L. Dellafiora), robin.ornsrud@hi.no (R. Ørnsrud), monica.sanden@hi.no (M. Sanden), marc.berntssen@hi.no (M.H.G. Berntssen), jean-lou.dorne@efsa.europa.eu (J.L.C.M. Dorne), vbafna@ucsd.edu (V. Bafna), josef.rasinger@hi.no (J.D. Rasinger).
[1] Both authors contributed equally.

https://doi.org/10.1016/j.envint.2023.108309
Received 25 January 2023; Received in revised form 15 August 2023; Accepted 4 November 2023
Available online 7 November 2023

Advanced toxicological modeling approaches are vital for advancing environmental risk assessments of chemicals in non-model species (Ashauer and Jager, 2018). For the last decade, fish physiology and toxicology have been predicted to move towards the concomitant analysis of genomic, proteomic, and metabolomic data sets (Martyniuk and Denslow, 2009). An uptick in sequencing efforts yielded draft genome assemblies for a wide range of teleost fish-species (Malmstrøm et al., 2017) and high quality assemblies for the Atlantic salmon (Davidson et al., 2010; Lien et al., 2016). In parallel, transcriptomic and proteomic data have increasingly become available providing multi-level omics information which, when integrated in proteogenomic analysis pipelines (Castellana and Bafna, 2010), can be used for additional validation and corrections of earlier genome annotations and for the prediction new genes previously missed in automated annotation pipelines (Ansong et al., 2008; Armengaud et al., 2014; Prasad et al., 2017). In the recent past, proteogenomic analyses have been demonstrated for many different species (Armengaud et al., 2014; Nesvizhskii, 2014; Ruggles et al., 2017) including, marine bivalves (Dumas et al., 2022) and zebrafish (Kelkar et al., 2014). For the zebrafish (Danio rerio), an important model organism widely used in biomedical and toxicology research, Kelkar et al. (2014) showed that proteogenomics can extend the understanding of even well-annotated genomes. Like zebrafish, Atlantic salmon is a well-studied fish species (Lien et al., 2016). However, an in-depth proteogenomic analyses to validate and revise its current reference genome annotation for use in mechanistic toxicology studies has yet to be performed.

In several non-model organisms, the recent advancements in genome sequencing, assembly, and annotation have strongly facilitated mechanistic analyses across different research areas, including ecology, toxicology, feed- and food-safety, and opened up the possibility for integrative analyses of global changes across the omics hierarchy (Armengaud et al., 2014; Dellafiora and Dall'Asta, 2017; Heck and Neely, 2020; Kelkar et al., 2014). For marine aquaculture and ecosystem research, mechanistic toxicity assessments are of interest when fish such as, Atlantic salmon are exposed to legacy (Lundebye et al., 2017; Nøstbakken et al., 2021) or emerging contaminants (Merel et al., 2019; Regueiro et al., 2017) and when information on contaminant-specific metabolism and toxic mode of action (MOA) are lacking (Rasinger et al., 2022). In model teleost fish species, genomic assessments of genes involved in the metabolism of foreign compounds identified integrative networks comprising key genes and proteins involved in the defense against xenobiotics (Eide et al., 2021). This includes xenobiotic receptors, transcription factors, biotransformation enzymes, transporters, antioxidants, and metal- and heat-responsive genes which collectively are known as the chemical defensome (Goldstone et al., 2006). Chemical defense proteins such as cytochromes P450 (cyp) for example, catalyze the metabolic conversion of xenobiotics in the liver (Sprenger et al., 2022) and are involved in the excretion of metabolites by hepatocytes (Hammer et al., 2021). Inter-species differences in the expression, abundances and activities of orthologs of different cyp genes related to xenobiotic metabolism including cyp1, cyp2, cyp3, and cyp4 were reported for different teleost fish species (Eide et al., 2021) as well as for different mammalian species and their respective cell model systems (Hammer et al., 2021). The toxicant-induced chemical defensome of Atlantic salmon has yet to be described.

In the present study, we apply proteogenomic strategies (Woo et al., 2015, 2014a, 2014b) to validate and revise the current Atlantic salmon reference genome annotation. We generate tissue-specific proteogenomics expression matrices to describe similarities and differences between Atlantic salmon tissues, and liver tissue and primary hepatocyte samples, respectively. To highlight the potential of proteogenomics for target tissue toxicity assessments in non-model species and their respective cell models, we provide an omics-based description of the chemical Atlantic salmon defensome *in vivo* and *in vitro*.

2. Material and methods

2.1. Samples

Atlantic salmon (*Salmo salar L.*) tissue samples of both genders (SalmoBreed strain) were obtained from a previously published trial (Berntssen et al., 2021). For proteogenomic analyses, 22 tissues samples comprising of liver, brain, eye, gill, gut, pyloric caecum, head kidney, heart, muscle, ovary, skin, and spleen were collected from two randomly chosen fish. In addition, liver samples were collected from 44 randomly selected fish, and a total of 60 3D salmon primary hepatocyte samples were included. In vitro samples were prepared as previously described (Olsvik and Søfteland, 2020). Following cell viability assessments (for which values between 82 and 98 % were obtained), hepatocytes were plated on 2 μg/cm^2 laminin (Sigma-Aldrich, Oslo, Norway) coated 3D Alvetex Scaffolds (200 μm cross-linked polystyrene membranes, 42 μm mean void size, REPROCELL, Glasgow, United Kingdom) in 12 well culture plates (FALCON, Corning, VWR, Bergen, Norway). Per well, 2.9 $\times$ 10^6 cells were used and suspended in 2.5 ml complete L-15 medium.

2.2. RNA-seq analyses

Total RNA was extracted from liver and primary hepatocyte samples using BioRobot® EZ1 and RNA Tissue Mini Kit (Qiagen, Hilden, Germany) including DNase treatment as instructed in the RNA Tissue Mini Kit manual (Qiagen, Hilden, Germany). RNA quality/purity of the samples was determined with a UV–vis Spectrophotometer (NanoDrop ND-1000, NanoDrop Technologies, Wilmington, USA). RNA integrity was analyzed using an Agilent 2100 Bioanalyzer and the RNA 6000 Nano LabChip kit (Agilent Technologies, Palo Alto, USA). The analysis showed satisfactory integrity and purity for both the *in vivo* samples (RIN = 9.6 ± 0.2, A260/280 = 2.10 ± 0.2, A260/230 = 2.16 ± 0.03) and the *in vitro* samples (RIN = 10.0 ± 0.0, A260/280 = 2.08 ± 0.03, A260/230 = 2.10 ± 0.15). Library preparation was done at the Norwegian Sequencing Centre (https://www.sequencing.uio.no). cDNA libraries from individual samples were constructed using TruSeq Stranded mRNA Library Prep Kit (Illumina) and standard Illumina adaptors for multiplexing as described by the manufacturer (Illumina). Individual libraries were sequenced using the NextSeq Illumina platform according to manufacturer instructions generating single end 75 bp read libraries. The TrimGalore 0.4.2 tool (https://github.com/FelixKrueger/TrimGalore) was applied for removing adaptors and for quality trimming using default parameters. Sequence quality for each sample was investigated using FastQC imbedded in TrimGalore. SAM files were generated by mapping to the above-mentioned salmon genome using the Hisat2 short read aligner version 2.0.4. Using FeatureCounts (Liao et al., 2014) of the Subread package (https://subread.sourceforge.net/), the abundance of transcripts was estimated for the individual libraries.

2.3. Proteomics mass spectrometry

Proteins were extracted from salmon livers and tryptic protein digestions were performed according to in-house standardized protocols as described in Bernhard et al. (2019). In short, employing an ultrasonication rod (Q55 Sonicator, Qsonica, CT, USA) at 30 % amplitude, liver tissues were lysed in 10 μL buffer (4 % SDS, 0.1 M Tris-HCl pH 7.6) per mg tissue. 3D hepatocytes cultured in Alvetex scaffolds were washed with PBS and proteins were extracted and solubilized using 600 μL 2x DIGE lysis solution (7 M urea, 2 M thiourea, 4 % w/v CHAPS, 2 % w/v DTT and 2 % v/v Pharmalyte™ pH 3–11) prepared as described in Rasinger et al. (2014). Following incubation for 30 min on a rotating platform (100 rpm) at room temperature, lysates were homogenized five times with a 20-gauge needle. Cell suspensions were transferred to centrifugation tubes and centrifuged at 12,000 rpm for 30 min at 4 °C. Supernatants were transferred to new centrifugation tubes, flash frozen and stored at −80 °C. Supernatants were collected and protein

concentrations were determined with a PierceTM BCA Protein assay kit (Thermo Scientific). To obtain a final concentration of 0.1 M, dithiotreitol (1 M) was added to the lysates, sample mixes were incubated (95 °C for 5 min) and then trypsin digested using a filter aided sample preparation (FASP) procedure as described by Wisniewski et al. (2009).

Mass spectrometry analyses followed standardized protocols at the Proteomics Unit of the University of Bergen (PROBE) as described in Bernhard et al. (2019). In short, between 0.5 and 1 ug tryptic peptides (dissolved in 2 % acetonitrile and 0.1 % formic acid) were injected into an Ultimate 3000 RSLC (Thermo Scientific) connected online to a linear quadrupole ion trap-orbitrap mass spectrometer equipped with a nano-spray ion source (LTQ-Orbitrap Elite, Thermo Scientific). Samples were desalted on a pre-column (Acclaim PepMap 100 C18, 2 cm × 75 μm ID, Thermo Scientific) and then, separated with a biphasic acetonitrile gradient (5–80 % for a total of 195 min) on an analytical column (Acclaim PepMap C18 100, 50 cm × 75 μm ID, Thermo Scientific). Peptides eluting from the LC-column were ionized, fragmented using collision-induced-dissociation (CID) and analyzed in data dependent acquisition (DDA) mode.

2.4. Proteogenomics analyses

To construct tissue-specific splice graph databases, raw RNA-Seq reads were retrieved from BioProjects PRJNA72713 and PRJNA260929, trimmed using TrimGalore (v.0.5.0), merged if from BioProject PRJNA260929, and mapped to the *Salmo salar* RefSeq genome (GCF_000233375.1_ICSASG_v2) using STAR (v.2.7.0.f) in 2-pass mode. Genome indexes for GCF_000233375.1_ICSASG_v2_genomic.fna were generated for reads from Illumina MiSeq and Illumina HiSeq 2000 instruments (genome index creation: hiseq_2x100: –sjdbOverhang 99 –sjdbGTFtagExonParentTranscript Parent –genomeChrBinNbits 14, miseq_2x250: –sjdbOverhang 249 –sjdbGTFtagExonParentTranscript Parent –genomeChrBinNbits 14). RNA-Seq read alignments in SAM format were then used to construct a splice graph database using the SpliceDB method (minimum number of reads: 3) (Woo et al., 2014a). Similarly, a splice graph database was constructed for the liver and 3D primary hepatocyte datasets using matched RNA-Seq data (minimum number of reads: 10).

The proteogenomics workflow searches MS/MS spectra against reference protein and proteogenomic databases using a multi-stage FDR approach similar to that described by (Woo et al. (2014b). Briefly, MS/MS spectra were first searched against the *Salmo salar* RefSeq proteome (GCF_000233375.1_ICSASG_v2_protein.faa) to identify reference proteins and peptides. Unidentified spectra not passing a 1 % PSM-level FDR are subsequently searched against a proteogenomic-based splice graph protein database using MS-GF+ (Kim and Pevzner, 2014). A final search against an enhanced database comprising RefSeq proteins, Augustus predictions, and Ensembl proteins supported by proteogenomic events was conducted using MS-GF+ for downstream analyses. For the *in vitro* and *in vivo* dataset, the enhanced database contained 97,555 RefSeq protein sequences, 71 Ensembl protein sequences, and one Augustus sequence. For the tissue dataset, the enhanced database held 97,555 RefSeq protein sequences, 107 Ensembl protein sequences, and eight Augustus sequences. In addition to sequences corresponding to the three novel Augustus gene predictions, additional Augustus sequences were included if proteogenomic evidence pointed to corrections in RefSeq or Ensembl annotations.

Genomic regions flanking identified proteogenomic events were submitted to Blastx, where high-scoring segment pairs (HSPs; *e*-value ≤ 1e−10) to proteins in the non-redundant (nr) database provided protein coding support for the genomic region. Specifically, sub-sequences of the *Salmo salar* genome (GCF_000233375.1_ICSASG_v2), defined as the nucleotides spanning the genomic coordinates of identified proteogenomic events and 350 bp of flanking sequences, were submitted to Blastx (online server) for a six-frame translation and comparison against protein sequences in the nr database. HSPs were included as hints to the gene prediction software Augustus. For predicted genes and transcripts, corresponding protein sequences were submitted to Blastp for comparison against protein sequences in the nr database. Blastp hits with significant alignments (*e*-value ≤ 1e−10, alignment ≥ 10 amino acids) were considered as orthologs in support of the predicted gene structure.

Along with identified proteogenomic events, HSPs from Blastx, and annotations from the *Salmo salar* genome (i.e., RefSeq GCF_000233375.1_ICSASG_v2 and Ensembl release-99), identified peptides from a MSGF + search against the RefSeq *Salmo salar* proteome along with Ensembl proteins supported by proteogenomic events were provided as hints to Augustus (v 3.3.3) for gene-prediction (parameters: –codingseq = on, –species = zebrafish, –alternatives-from-evidence = true, –allow_hinted_splicesites = atac, –extrinsicCfgFile = extrinsic.M.RM.E.W.P.cfg, –softmasking = on). Hints were limited to those within 60,000 bp of a given proteogenomic event group. Predicted genes with evidence from at least one proteomic hint were kept for further analysis.

To compute a protein-level FDR, proteins were scored based on precursors (combination of peptide sequence and charge) passing a 1 % level FDR, where the score of a precursor is defined as the SpecEValue of its best scoring PSM. Specifically, the score of a protein is initialized as the sum of the negative log of the SpecEValue of its precursors. After choosing the highest scoring protein, precursors to that protein are removed and protein scores are recalculated. This process is iterated until no precursors are left. A protein false discovery rate is then computed as the number of decoy proteins/target proteins.

2.5. Protein expression and chemical defensome analyses

MS/MS spectra were searched against the RefSeq *Salmo salar* proteome with added Ensembl and Augustus predicted proteins supported by proteogenomic and comparative genomic evidence (Tissue dataset: 107 Ensembl and 8 Augustus predictions; Liver and Hepatocyte dataset: 71 Ensembl and 1 Augustus predictions; Fig. 1B). Ensembl proteins are defined as translations resulting from Ensembl gene predictions (Salmo_salar.ICSASG_v2.pep.all.fa). Raw spectra counts, S_p, for identified proteins (1 % protein-level FDR) in each sample are based on spectra passing a 1 % PSM-level and 1 % precursor-level FDR, where c_p is the number of PSMs representing peptide, p, mapping to a protein, P, and N (p) is the number of proteins mapping to a peptide p.

$$S_p = \sum_{p \in P}^{P} \frac{c_p}{N(p)}$$

A normalized spectral count, PSK_p, is implemented to account for protein length.

$$SPK_p = \frac{S_p}{|P| \times 0.01}$$

$$SPK_{allP} = \sum_{P \in all\ P}^{all\ P} SPK_p$$

$$PSK_p = \frac{SPK_p}{\frac{SPK_{all\ P}}{1000}}$$

To generate a heatmap of the expression matrix, hierarchical clustering was performed separately on the spectrum files (columns) and proteins (rows) (ward.D2, Euclidean) of the normalized PSK expression matrix, and the matrix was further log transformed for visualization purposes.

A list of genes included in the chemical defensome of five model teleost fish was created based on the supplementary material provided in Eide et al. (2021). For sanity checking, files and codes created were first used to find chemical defensome related genes in the zebrafish (*Danio rerio*) reference proteome (UP000000437, March 2022) and the results obtained were compared with data presented in Eide et al. (2021). Following the successful completion of the sanity check (data

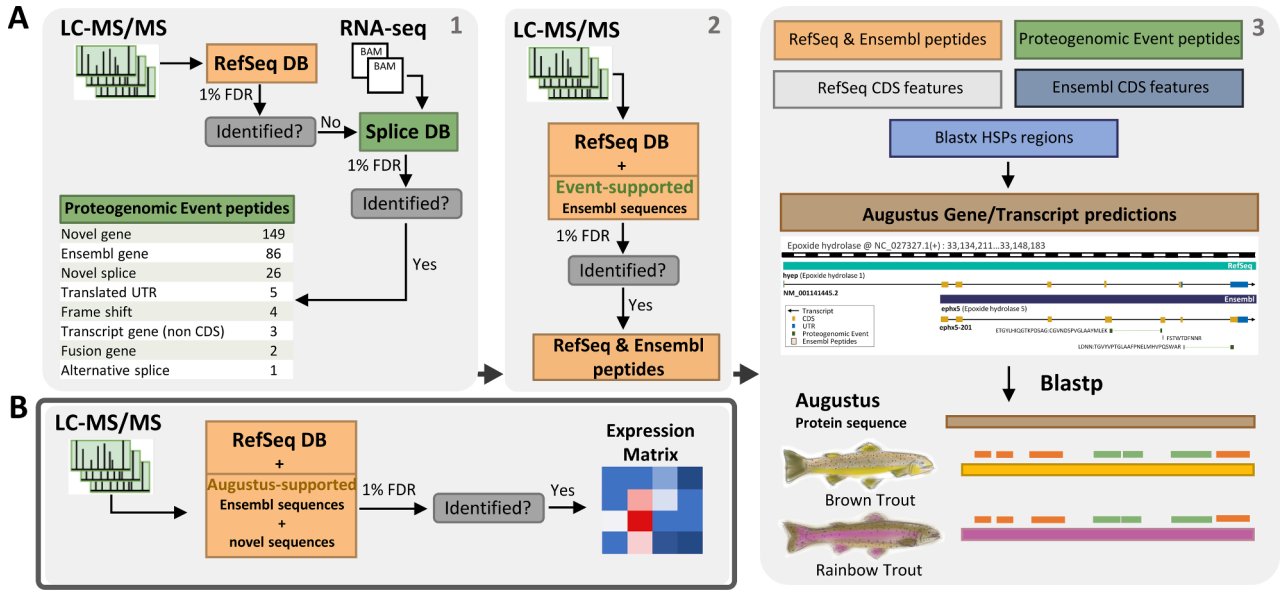


Fig. 1. Proteogenomics workflow, gene prediction, downstream analyses. (A) MS/MS spectra were searched against the reference *Salmo salar* proteome and a proteogenomic splice database in a multi-stage FDR approach (1). Proteogenomic databases were constructed using either proteomic sample-matched RNA-Seq reads or publicly available transcriptomic data. Of the 276 proteogenomic events identified in (1), 86 support gene annotations in the Ensembl assembly. Spectra were therefore searched against the reference proteome appended with proteogenomic supported Ensembl proteins to provide additional hints for gene prediction (2). Identified peptides from (1) and (2), RefSeq and Ensembl annotated coding regions (CDS), and Blastx high-scoring pair (HSP) regions were submitted as hints to Augustus for gene prediction (3). Corresponding protein sequences were subsequently submitted to Blastp for identification of homologs, providing additional comparative genomic support for predictions. (B) For downstream analyses, MS/MS spectra were searched against an enhanced reference proteome with additional novel proteins and Ensembl annotated proteins supported by Augustus predictions. Additional information on the workflow presented here is provided in Suppl. File 1.

not shown), the defensome gene list from Eide et al. (2021) was matched against the Atlantic Salmon reference proteomes (UP000087266; March 2022) and the protein expression matrix obtained in the present study using a proteogenomic strategy. For improving dentification rates of chemical defensome proteins in salmon, using PAW_BLAST db_to_db_blaster.py (https://github.com/pwilmart/PAW_BLAST) multi-tissue and liver specific proteogenomic expression matrices were matched against zebrafish reference proteome (UP000000437, March 2022). Outputs were recoded using tidyverse (v.1.3.0) functions (Wickham et al., 2019); plots were created using the packages ggplot2 (v.3.3.2), UpSetR (v.1.4.0) and ComplexHeatmap (v.2.4.3).

2.6. Data availability

To support the rapidly advancing development of omics tools for toxicological research in non-model organisms and in line with FAIR (findable, accessible, interoperable, and reusable) data principles in risk assessment and toxicology (Pineda-Pampliega et al., 2022), the proteogenomics workflow is provided in full in Suppl. File 1 and all supporting data are provided in Suppl. Figs. 1–7 and Suppl. Tables 1–10. RNA-Seq data and LC-MS/MS data generated in the present work were made available on NCBI (PRJNA987864/PRJNA987421) and MassIVE (MSV000089139, MSV000089141), respectively. Relevant scripts used were made available on https://github.com/miinslin/SalmonProteogenomics.

3. Results

3.1. Proteogenomic support for novel gene and Ensembl assembly annotations

For a systems-wide representation of the Atlantic salmon, samples from 12 salmon tissues, and salmon primary hepatocyte cultures, were subjected to liquid chromatography tandem mass spectrometry (LC-MS/MS). Spectra were searched against the reference salmon proteome (RefSeq assembly GCF_000233375.1) followed by a custom proteogenomic splice database using a multi-stage FDR approach (Woo et al., 2015) (Fig. 1A, panel 1). Our workflow identified 79,498 peptides which matched 40,590 RefSeq gene structures and included 19,132 spliced peptides providing a strong proteomic validation of these transcripts. In addition, we detected 274 peptides representing 183 distinct proteogenomic events which matched genomic loci where transcripts had been observed but were not accompanied by a RefSeq gene annotation. We next checked if these events were supported by other annotations of the genome and found that 86 of the 274 identified proteogenomic event peptides supported gene annotations in the Ensembl assembly, suggesting that reference proteomes based solely on the RefSeq assembly may not be comprehensive.

To detect novel gene structures and reference annotation corrections, we provided the gene prediction software Augustus (Stanke et al., 2004) with hints to coding sequence regions, including genomic coordinates for identified reference peptides and proteogenomic events, reference annotation features, and genomic regions that are also known to be protein-coding in other organisms (Fig. 1A, panel 3; methods). Additional hints from proteomic sources may increase the accuracy of Augustus gene predictions; we thus also included peptides that were identified from searching spectra a second time against the reference proteome with additional proteogenomic supported Ensembl proteins (Fig. 1A, panel 2). This resulted in Augustus predictions for 71 transcripts (Suppl. Table 1) supported by 681 peptides. The protein sequences of predicted transcripts were then submitted to Blast for the identification of homologs as further support for the prediction (Fig. 1A, panel 3).

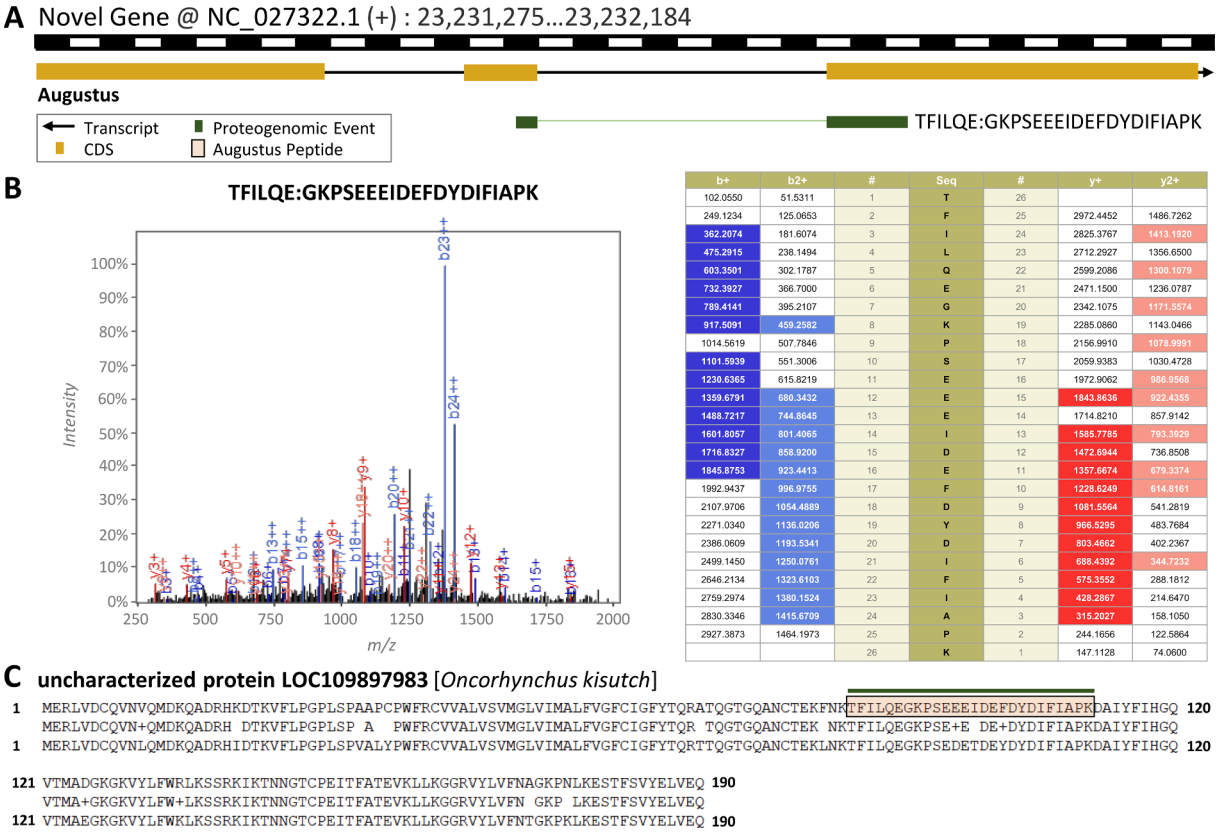
While many transcripts matched Ensembl annotations, others provided corrections to RefSeq annotations (Suppl. Table 2). Of the 71 predicted transcripts, 62 were highly matched to Ensembl annotations, with 20 of the 62 without a RefSeq annotation in the region. We identified three novel genes in previously unannotated regions. One such

example is a novel gene on chromosome 23 (NC_027322.1) that is supported by a proteogenomic peptide TFILQE:GKPSEEEIDEFDYDIFIAPK spanning the junction between exons 2 and 3 of the predicted transcript (Augustus_ID #64 in Suppl. Table 1; Fig. 2A). The peptide annotated multiple peaks corresponding to b/y ions and explaining 17 of the 50 highest intensity peaks (Fig. 2B). A Blastp search of the predicted protein sequence against the nr database resulted in a hit to an uncharacterized protein (LOC109897983; identity: 92.6 %; e-value: 7.33e-129) in the coho salmon (*Oncorhynchus kisutch*), a closely related salmonid species (Fig. 2C). Together, these results suggest strong proteogenomic and comparative genomic evidence in support of the novel gene.

Of the 43 Augustus predictions supporting corrections to reference annotations, the myosin heavy chain 7 (myh7) gene on the negative strand of chromosome three (NC_027302.1) is an example where the RefSeq gene prediction is truncated and missing coding regions 3′ downstream of the annotated stop codon (Augustus_ID #1; Fig. 3). Primarily expressed in the cardiac muscle, the myh7 gene may also be expressed in other skeletal muscles. Indeed, seventeen proteogenomic event peptides were identified in the salmon heart and eye tissues, with four being expressed in both. Mapping to CDS regions in the Ensembl gene (ENSSSAT00000137950.1), many of these peptides also spanned exon-exon junction sites (Fig. 3A).

To rule out the possibility that the Ensembl gene was mis-annotated and has been accounted for elsewhere by the reference annotation, the corresponding protein (ENSSSAP00000103327.1) was submitted to Blastp for a search against the nr database. This resulted in a hit to a myosin-like protein on chromosome nineteen (XP_014015175.1) of the reference annotation. Despite an overall high-scoring alignment between the two protein sequences (identity: 88.4 %; e-value: 0.0), in areas with poorer alignment, the Ensembl protein is supported by high quality proteogenomic event peptides (Fig. 3B). For example, the proteogenomic peptides LLTVLFQNYAGTDA:ADDSK and KKLESDTNQLQTEVEEAVQECR at positions 615 and 1733, respectively, of the Ensembl protein have high-quality peptide-spectrum matches (PSMs) with a high coverage of b- and y-ions (Suppl. Fig. 1A, 1B). Furthermore, a proteomic search against an enhanced reference database (methods) identified 198 peptides that span the length of the Ensembl protein (Fig. 3B), suggesting that a correction should be made to the current annotation.

As another example, the Ensembl epoxide hydrolase gene (ENSSSAT00000145609.1) on the positive strand of chromosome 28 (NC_027327.1) represents either a correction to the reference annotation or an additional isoform (Augustus_ID #66; Fig. 4A). Specifically, we identified 20 peptides mapping to the Ensembl protein (ENSSSAP00000110682.1), including three proteogenomic event peptides supporting the partial translation of a 3′ untranslated region (UTR) and the addition of novel splice junctions. As before, we submitted a Blastp search of the Ensembl protein against the nr database to verify that the predicted transcript has not been mapped elsewhere in the reference annotation. While there was a hit to the salmon epoxide hydrolase 1-like protein (XP_014000050.1) on chromosome 15, there was poor overall alignment (identity: 58.2 %; e-value: 4e-176), as also presence of high-quality proteogenomic peptides at regions with poorer alignment (Fig. 4B). These peptides, FSTWTDFNNR, LDNN:TGVYVPTGLAAFPNELMHVPQSWAR, and ETGYLHIQGTKPDSAG:CGVNDSPVGLAAYMLEK at positions 287, 344, and 254, respectively, of the Ensembl protein



b+	b2+	#	Seq	#	y+	y2+
102.0550	51.5311	1	T	26		
249.1234	125.0653	2	F	25	2972.4452	1486.7262
362.2074	181.6074	3	I	24	2825.3767	1413.1920
475.2915	238.1494	4	L	23	2712.2927	1356.6500
603.3501	302.1787	5	Q	22	2599.2086	1300.1079
732.3927	366.7000	6	E	21	2471.1500	1236.0787
789.4141	395.2107	7	G	20	2342.1075	1171.5574
917.5091	459.2582	8	K	19	2285.0860	1143.0466
1014.5619	507.7846	9	P	18	2156.9910	1078.9991
1101.5939	551.3006	10	S	17	2059.9383	1030.4728
1230.6365	615.8219	11	E	16	1972.9062	986.9568
1359.6791	680.3432	12	E	15	1843.8636	922.4355
1488.7217	744.8645	13	E	14	1714.8210	857.9142
1601.8057	801.4065	14	I	13	1585.7785	793.3929
1716.8327	858.9200	15	D	12	1472.6944	736.8508
1845.8753	923.4413	16	E	11	1357.6674	679.3374
1992.9437	996.9755	17	F	10	1228.6249	614.8161
2107.9706	1054.4889	18	D	9	1081.5564	541.2819
2271.0340	1136.0206	19	Y	8	966.5295	483.7684
2386.0609	1193.5341	20	D	7	803.4662	402.2367
2499.1450	1250.0761	21	I	6	688.4392	344.7232
2646.2134	1323.6103	22	F	5	575.3552	288.1812
2759.2974	1380.1524	23	I	4	428.2867	214.6470
2830.3346	1415.6709	24	A	3	315.2027	158.1050
2927.3873	1464.1973	25	P	2	244.1656	122.5864
		26	K	1	147.1128	74.0600

Fig. 2. Novel Gene. (A) Augustus predicted a gene on chromosome 23 (NC_027322.1) in a region previously unannotated by RefSeq and Ensembl. A peptide identified in the proteogenomic database search (green) maps to CDS annotations in the predicted transcript. (B) Peptide-spectrum match (PSM) of the proteogenomic event peptide TFILQE:GKPSEEEIDEFDYDIFIAPK identified in the *Salmo salar* ovary tissue showing b-ion (blue) and y-ion (red) matches of charge +1 and +2 (fragment mass tolerance 0.4 Da). (C) A Blastp search of the predicted protein against the non-redundant database resulted in a hit to an uncharacterized protein (LOC109897983) in the salmonid *Oncorhynchus kisutch* (Coho salmon) RefSeq gene build (XP_020348278.1; e-value: 7.33e-129; identity: 176/190, 92.6 %). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

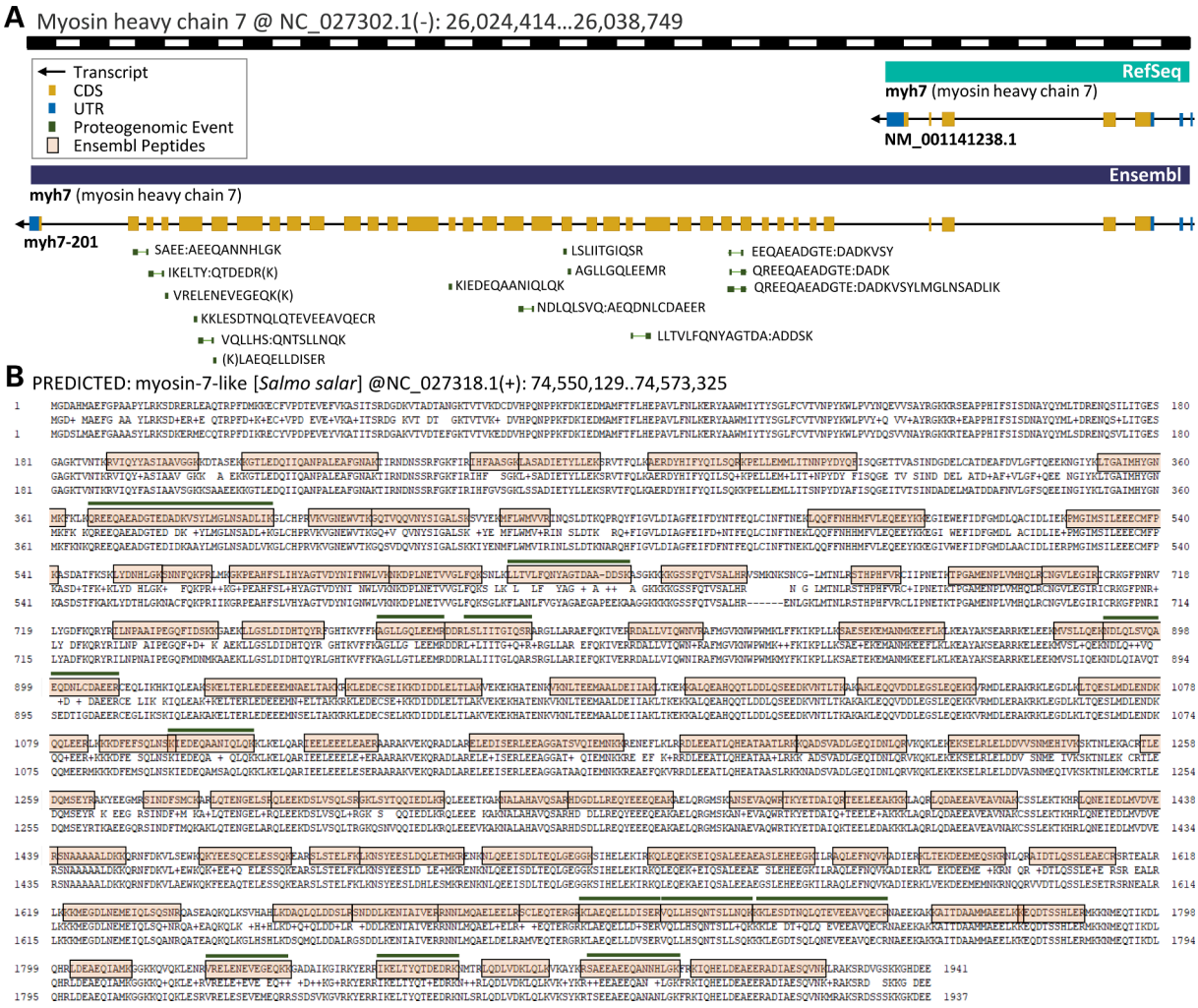


Fig. 3. Myosin heavy chain 7. (A) Reference (turquoise) and Ensembl (navy) annotations of the myosin heavy chain seven (myh7) gene on chromosome three (NC_027302.1). Seventeen peptides identified in the proteogenomic database search (green) map to CDS annotations in Ensembl. (B) A Blastp search of the Ensembl protein (ENSSSAP00000103327.1) against the non-redundant database resulted in a hit to a predicted myosin-7-like protein (XP_014015175.1) on chromosome 19 (NC_027318.1) of the *Salmo salar* genome (e-value: 0; identity: 1718/1943, 88.4 %). A proteomic search against an enhanced reference database identified 198 (orange) peptides mapping to ENSSSAP00000103327.1, including the previously identified proteogenomic event peptides (green lines). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

prediction, have high-quality peptide-spectrum matches (PSMs) with a high coverage of b- and y-ions (Suppl. Fig. 2). In contrast, the Ensembl protein prediction has a higher scoring alignment (identity: 98.7 %; e-value: 0.0) to the epoxide hydrolase 1-like isoform X1 (XP_029621434.1) from the brown trout (*Salmo trutta*), suggesting that the Ensembl annotation is supported by conserved coding regions (Fig. 4B).

3.2. Multi-tissue LC-MS/MS datasets: Tissue-specific protein profiles

The availability of tissue specific proteomic data provides an opportunity to study protein expression directly across different tissues. We generated 23 LC-MS/MS salmon tissue specific datasets as a resource for the community (MassIVE ID: MSV000089139 and Suppl. Table 3). Given the strong proteogenomic evidence in support of Ensembl and Augustus predicted transcripts, we searched spectra again; this time against an enhanced reference database containing RefSeq proteins, Augustus predictions, and Ensembl proteins (Fig. 1B; methods). We identified a total of 9,392 proteins (59,899 peptides) across all tissues, with brain, ovary, and liver samples expressing the highest number of proteins uniquely detected in a specific tissue (Fig. 5). Principal component analysis (PCA) of the normalized expression count values of the 600 most highly expressed proteins across all 23 LC-MS/MS samples showed strong clustering of tissue types, with principal component 1 and 2 explaining 29.2 % and 15.0 % of the observed variance, respectively (Fig. 5B).

Hierarchical clustering of the expression matrix (top 600 proteins) further revealed tissue-specific protein signatures (Suppl. Fig. 3; Suppl. Table 4). Cluster #1 was characterized by proteins found across all tissues, including those related to oxygen transportation in blood (Suppl. Fig. 4A). The gill, spleen, head kidney, and heart showed a particularly high expression of these proteins, consistent with the role that these organs play in the vascular system. Cluster #2 contained proteins related to maintaining the integrity and function of the intestine, and fittingly, were also expressed in the pyloric caecum, which is adjacent to the stomach and gut (Suppl. Fig. 4B). Cluster #3 contained proteins related to muscle anatomy and function, including actin, myosin, and those involved in the glycolytic process - these proteins are highly expressed in muscle tissue and not in other tissue types (Suppl. Fig. 4D). Cluster #4 consisted of proteins related to heart function and structure, including those involved in calcium ion binding, ventricular cardiac myofibril assembly, heart contraction, and oxygen transportation (Suppl. Fig. 4C). Cluster #6 is dominated by proteins found in the integrity of skin, including keratin and collagen related proteins (Suppl. Fig. 4E).

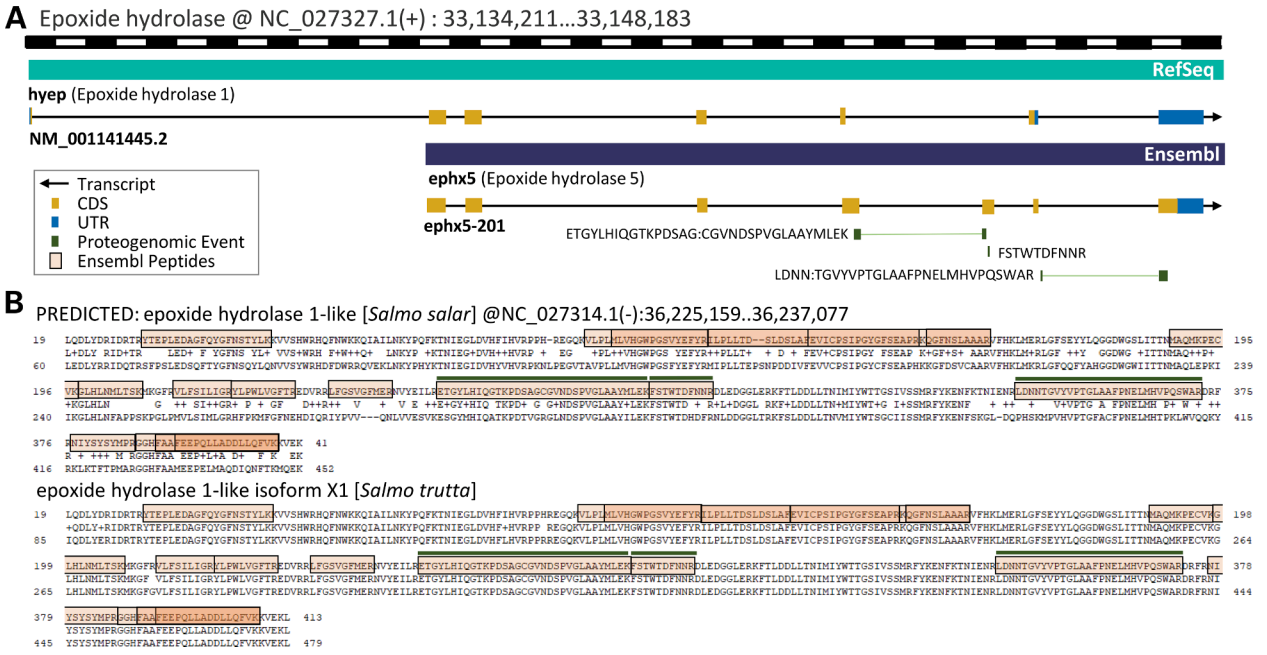


Fig. 4. Epoxide hydrolase. (A) Reference (turquoise) and Ensembl (navy) annotations of the epoxide hydrolase gene on chromosome 28 (NC_027327.1). Three peptides identified in the proteogenomic database search (green) map to CDS annotations in Ensembl. (B) A Blastp search of the Ensembl protein (ENSSSAP00000110682.1) against the non-redundant database resulted in hits to *Salmo salar* proteins, including XP_014000050.1 on chromosome 15 (NC_027314.1) (e-value: 4e-176; identity: 231/397, 58.2 %), and other salmonids proteins, including XP_029621434.1 from *Salmo trutta* (e-value: 0; identity: 390/395, 98.7 %). A proteomic search against an enhanced reference database identified 20 (orange) peptides mapping to ENSSSAP00000110682.1, including the previously identified proteogenomic event peptides (green lines). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

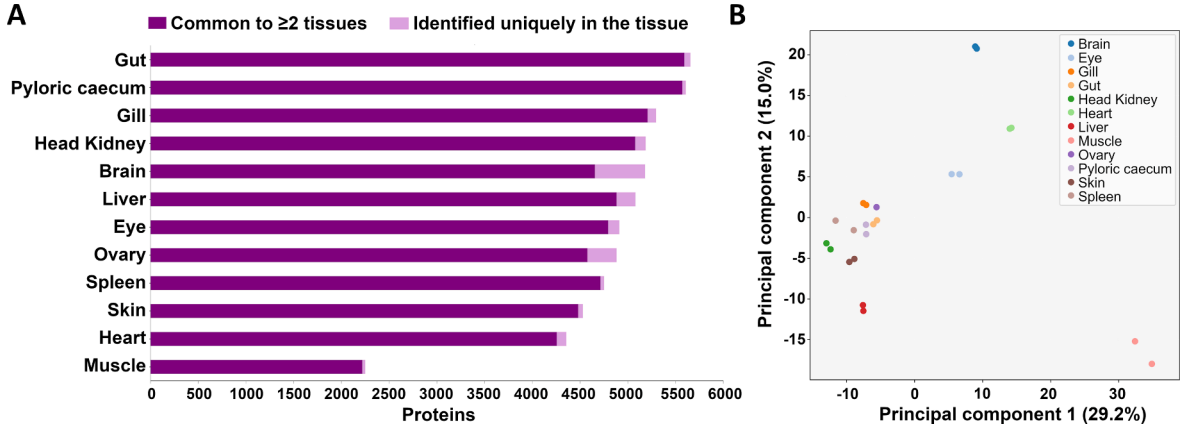


Fig. 5. Various Salmon tissues. (A) A total of 9,392 proteins (59,899 peptides) were identified in twelve tissue types. The highest number of proteins were identified in the Gut (5,675 proteins; 43,795 peptides), while the highest number of tissue-unique proteins were identified in the Brain (526 proteins; 1,467 peptides). (B) Principal component analysis (PCA) of the top 600 proteins across all twenty-three LC-MS/MS samples reveals grouping of samples based on the twelve tissue types, with principal component 1 and 2 explaining for 29.2 % and 15.0 % of variance, respectively.

3.3. Salmon primary hepatocytes as in vitro model for Atlantic salmon liver

LC-MS/MS analysis of liver samples collected from 44 randomly selected fish and a total of 60 3D salmon primary hepatocyte samples were performed and both datasets were made available on MassIVE (ID: MSV000089141). Of the 6,280 identified proteins, 5,389 proteins were detected in both primary hepatocytes (*in vitro*) and liver tissue (*in vivo*) samples (Suppl. Table 5). Of the remaining 891 proteins, 661 proteins were uniquely identified *in vitro*, and 230 proteins were uniquely identified *in vivo* (Fig. 6A).

A smaller number of spectra acquired for *in vivo* samples may account for the difference seen in the number of identified peptides and proteins (Suppl. Fig. 6A). Measured protein abundances spanned over six orders of magnitude (Fig. 6B). For the bulk of higher abundance (mean centered log abundance > 0) proteins from the PSK normalized proteogenomics expression matrix overall, a good global correlation (correlation coefficient: 0.84) between *in vitro* and *in vivo* samples was observed (Fig. 6C). Hierarchical clustering of the expression matrix (104 spectrum files, top 1,000 proteins) revealed a separation of *in vitro* samples from *in vivo* samples (Suppl. Fig. 6B). Proteins highly expressed *in vivo* but not *in vitro* (Cluster A, B) include proteins with functional and structural roles in blood, including hemoglobin alpha and beta subunit proteins and serum albumin (Suppl. Table 6; Suppl. Fig. 5).

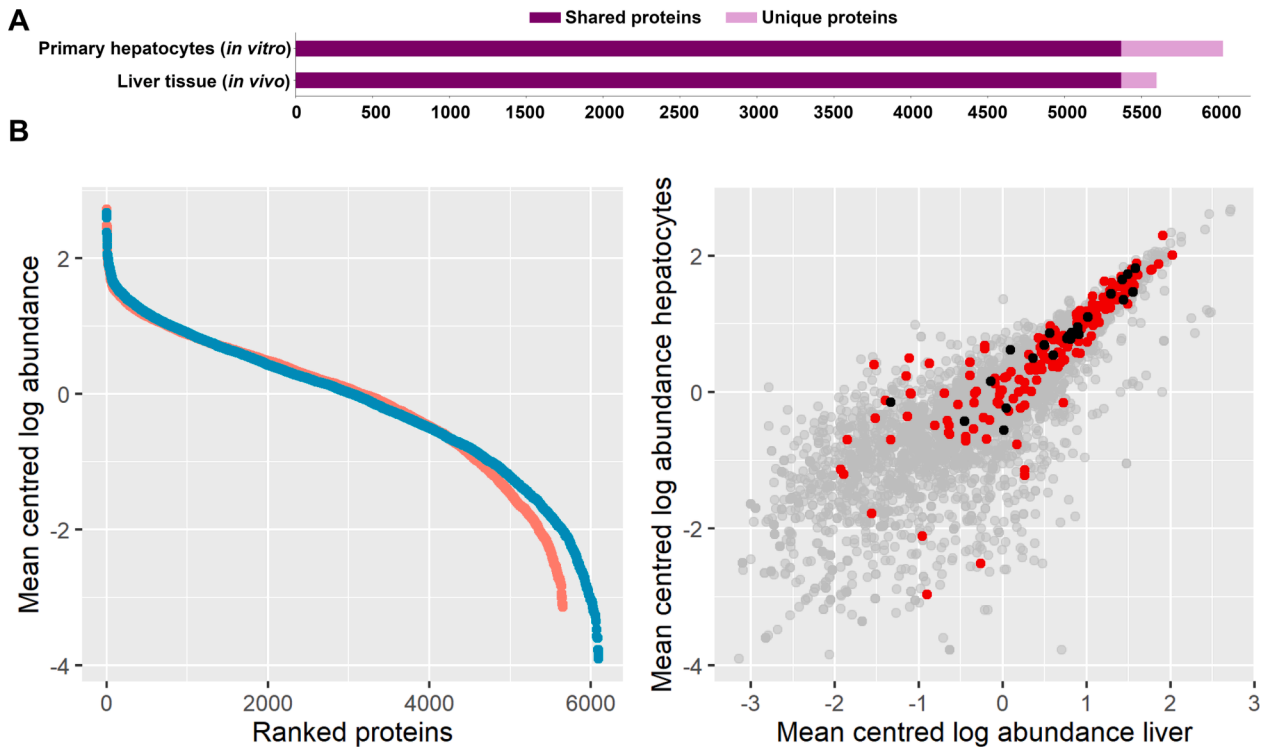


Fig. 6. Salmon primary hepatocytes and liver tissue. (A) Of the total of 6,280 identified proteins (39,676 peptides), 5,389 proteins (37,872 peptides) were detected in both primary hepatocytes and liver tissue, while 661 proteins (1,169 peptides) and 230 proteins (635 peptides) were identified uniquely, respectively. Ranked protein abundances (B) and correlation of protein abundances (B) in Atlantic salmon liver and primary hepatocytes. (B) Blue and red dots represent primary hepatocyte and liver samples; (C) grey, red, and black dots represent all, chemical defensome specific, and cyp proteins detected in Atlantic salmon liver and primary hepatocytes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4. The Atlantic salmon chemical defensome

Protein expression matrices of multi-tissue, liver and 3D primary hepatocyte samples were matched against the full complement of the putative chemical defensome of model teleost fish species (Suppl. Table 7) adapted from the supplementary material of Eide et al. (2021). Using the salmon-specific annotation of the multi-tissue and the liver/primary hepatocyte protein matrices, a total of 68 (Suppl. Table 3) and 59 (Suppl. Table 5), respectively of chemical defensome specific proteins were identified. Following a PAW_BLAST search for zebrafish orthologs (Suppl. Table 8 and Suppl. Table 9), the identification of chemical defensome specific proteins improved; a total of 306 and 248 proteins of the multi-tissue and liver/primary hepatocyte data, respectively were successfully mapped to gene patterns and ontology terms linked to proteins involved in the detoxification and clearance of xenobiotic compounds in Atlantic salmon (Fig. 7A; Fig. 7B). Hierarchical clustering of the 306 multi-tissue defensome proteins showed tissue-specific proteins signatures and clustering of samples based on tissue type (Suppl. Fig. 7). Distinct and overlapping defensome protein expression were observed across all salmon tissues analyzed (Fig. 7C), and a large degree of homology and between *in vivo* and *in vitro* liver systems was observed. Of 248 liver-specific defensome proteins present in the proteogenomics expression matrix, only two proteins were detected exclusively in the *in vitro* system (Fig. 7D).

As shown in Fig. 6C, a good global correlation (correlation coefficient: 0.86) between chemical defense related proteins was observed between *in vivo* and *in vitro* data. Taking cyp expression as an example (correlation coefficient: 0.94; Fig. 6C), we found that several paralogs of cyp1, cyp2, cyp3 and cyp4 were detected in Atlantic salmon liver and primary hepatocyte samples (Suppl. Table 10). We also observed that for many cyp paralogs mean PSK normalized abundance values were similar across samples. For several zebrafish orthologs multiple salmon paralogs were found. In addition, it was observed that for cyp1 and cyp2 paralogs multiple RefSeq entries were detected for individual Ensembl annotations (Suppl. Table 10).

4. Discussion

We provide a comprehensive proteogenomic characterization of the Atlantic salmon (*Salmo salar*) genome for mechanistic toxicology studies *in vivo* and *in vitro*. Using a custom proteogenomics workflow (Fig. 1, Suppl. File 1), we identified close to 80,000 peptides providing direct evidence of translation for over 40,000 RefSeq structures. In addition, our data highlighted differences between the two annotation platforms, RefSeq and Ensembl. Our data indicated that salmon reference proteomes based solely on RefSeq assemblies might not be as comprehensive as the ones based on Ensembl annotations. In the context of RNA-Seq mapping, it was found previously that Ensembl generally annotates more genes than RefSeq does (Zhao and Zhang, 2015). With functional annotation results potentially having a strong influence on the conclusions to be derived from experimental studies (McCarthy et al., 2014), the selection of one annotation over another can have a strong effect on the overall outcome of omics analyses (Frankish et al., 2015). For transcriptomic-focused studies, it was suggested previously to choose less complex genome annotations such as RefSeq, for reproducible and robust gene expression estimates; for exploratory research, more complex genome annotations such as Ensembl, were recommended (Wu et al., 2013).

Until experimentally validated, genome annotations remain hypothetical (Prasad et al., 2017). To validate and correct previous annotations, several proteogenomics strategies have been developed (Castellana and Bafna, 2010; Nesvizhskii, 2014; Ruggles et al., 2017) and applied to different species of scientific interest (Prasad et al., 2017; Ruggles et al., 2017) including fish such as, zebrafish (Kelkar et al.,

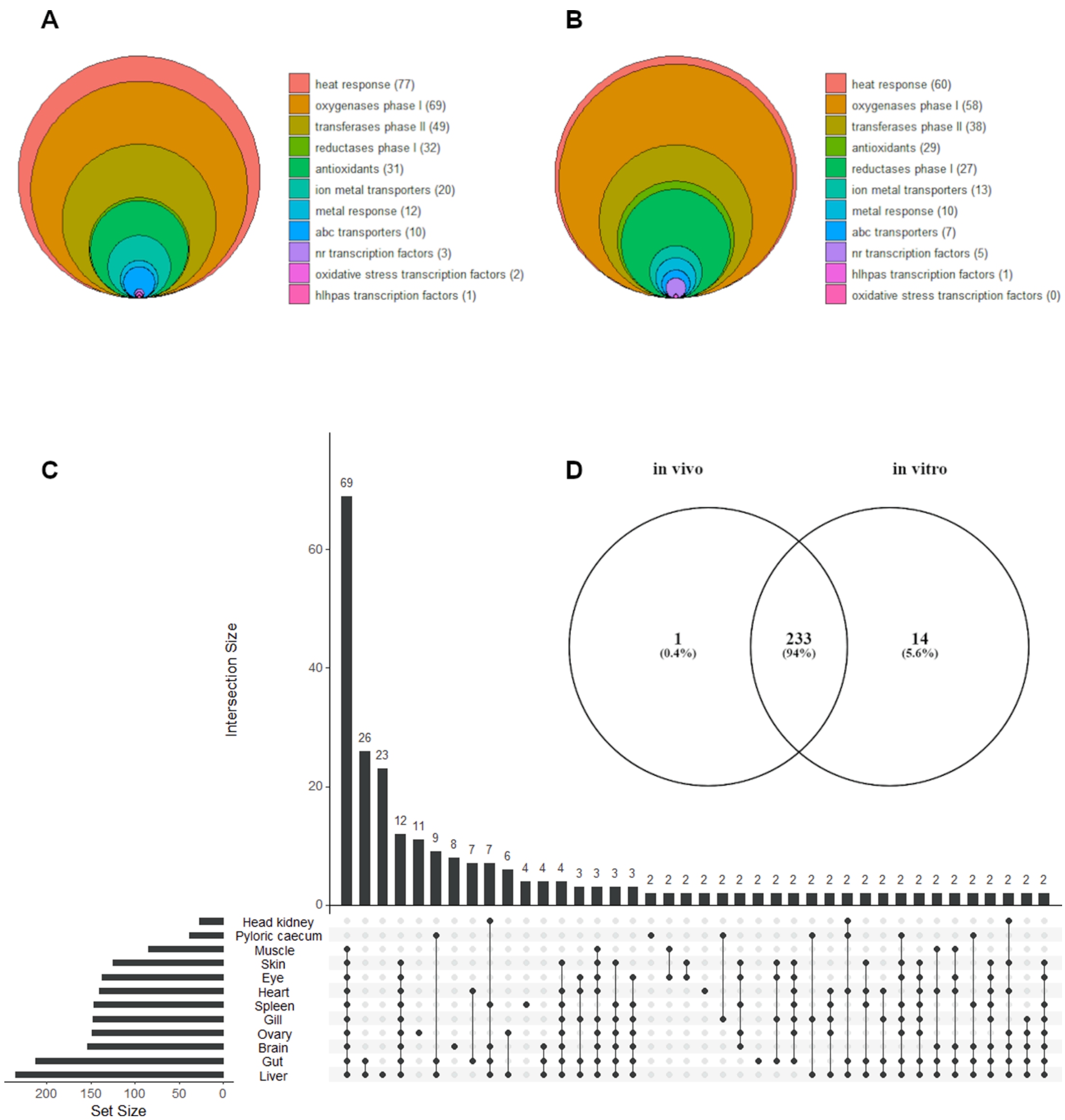


Fig. 7. Atlantic salmon chemical defensome. Bubble plot of chemical defensome specific proteins of multi-tissue (A) and liver/hepatocyte (B) proteogenomics expression matrices (C) UpSet plot visualizing unique and overlapping defensome protein expression across all salmon tissues analyzed (D) Venn diagram displaying distinct and overlapping defensome proteins detected *in vivo* and *in vitro* liver systems.

2014). Like zebrafish, Atlantic salmon is also a well-studied fish species (Lien et al., 2016). In addition, salmon is considered to be of high environmental, economic, and cultural value (Fedoroff et al., 2022; Gillson et al., 2022; Glover et al., 2020). Yet, a proteogenomic analysis on this important species had not previously been performed.

Our proteogenomic analyses of the Atlantic salmon suggested corrections to several annotations and identified novel genes that other annotation efforts had missed. This included genes such as epoxide hydrolase 1 (ephx1; Fig. 4A), a biotransformation enzyme crucial for xenobiotic detoxification *in vivo* and *in vitro* (Marowsky et al., 2017). A proteomic search against our enhanced reference database identified a total of 20 ephx1 peptides (Fig. 4A). In other words, in line with earlier proteogenomic studies on fish with well-annotated genomes such as the zebrafish (Kelkar et al., 2014), our data also indicate that proteomic evidence can efficiently be used to confirm translation and as such, should be considered to be included as an integral part of modern-day annotation pipelines (Castellana and Bafna, 2010; Nesvizhskii, 2014).

In contrast to genomic annotation, but similar to transcript-based annotation, proteogenomic annotation provides a dynamic picture of the proteome that displays species- and tissue-specific expression patterns of proteins; these can efficiently and cost-effectively be analyzed using standard shotgun proteomics approaches (Rasinger et al., 2016). The analysis of multi-tissue salmon data confirmed this. According to PCA of data derived from a single expression matrix, paired samples from the same tissue clustered together closely and separated from other tissues (Fig. 5B). The multi-tissue expression matrix for this analysis was generated utilizing the union of all peptides identified by our proteogenomic pipeline. Protein expression levels for this matrix were estimated using a novel normalized spectral count approach developed for this work which corrects for protein length and the total number of spectra acquired across all samples in the set. This, in part, tackles some of the challenges associated with proteogenomics analyses, namely, the integration and visualization of data sets comprising quantitative measurements across multiple data modalities (Ruggles et al., 2017).

The integrated proteogenomics protocol and automated protein expression matrix generation presented here (Fig. 1), also pave the way for applying mechanistic research when working with newly or only scarcely annotated non-model organisms (and their respective *in vitro* models) central to marine, agricultural, veterinary or ecosystem

research. Here, as a preliminary proof of concept for the applicability of our proteogenomic annotation and analysis pipeline for mechanistic toxicology research in non-model species, we compared the Atlantic salmon liver tissue proteome with the proteome of cultured 3D primary hepatocyte using a single expression matrix (Suppl. Table 5). Except for the expression of blood-related proteins, which were missing in the cultured cells (probably as a result of the cell extraction method and culturing conditions), we found a high degree of consistency in expression profiles with a strong correlation between *in vitro* and *in vivo* data (Fig. 6; Suppl. Table 6). This is in line with previous studies on human samples, in which a high degree of similarity was observed when comparing proteomic expression data of human liver tissue and isolated hepatocytes (Vildhede et al., 2015). Similar findings also were obtained when genomic RNA-Seq data of liver models were compared with *in vivo* human liver data (Gupta et al., 2020). In addition to the global comparison of protein expression between liver tissue and hepatocytes, we specifically mined the proteogenomics augmented expression data for proteins involved in xenobiotic detoxification, such as cytochromes P450. Applying a search strategy for finding chemical defensome proteins in teleost fish (Eide et al., 2021), in the enhanced Atlantic salmon protein reference dataset, three cyp proteins (Q8QGP2, B5DG17, B5X2R4) were identified both *in vivo* and *in vitro*. Following the conversion of the salmon-specific protein annotations into their respective zebrafish orthologs, we found a total of 14 cyp paralogs in the unified *in vivo* and *in vitro* expression matrix (Suppl. Table 10). We also found that averaged PSK normalized abundance values for these proteins were strongly correlated between *in vitro* and *in vivo* systems (Fig. 6). These data corroborate that primary hepatocytes can be a suitable model system for mechanistic toxicity studies in Atlantic salmon liver.

The conversion of Atlantic salmon protein annotations into zebrafish-specific annotations increased the number of cyp proteins detected, and facilitated the detection of the xenobiotic biotransformation enzyme epx1 described above. Our proteogenomic evidence correctly identified salmon epx1 and suggested a correction for its reference genome annotation (Fig. 4). However, only after the conversion of salmon epx1 to its respective zebrafish ortholog was this enzyme detected when we programmatically screened for chemical defense proteins in the proteogenomics augmented expression matrix (Suppl. Table 5). This shows that despite improving the number and quality of protein identifications using proteogenomics reference databases, for a biological interpretation, non-model species protein identifiers still require conversion to their respective model-species orthologs using basic local alignment search tools (Blast). Blast conversions commonly are implemented in non-model species research to facilitate comparison to other studies and to facilitate downstream analyses as, depending on the origin of the non-model organism sequence database, identifiers may simply be a generic locus or transcript ID unsuitable for biological interpretation (Heck and Neely, 2020).

Future work to advance the understanding of the chemical defensome of the Atlantic salmon will focus on the analyses of *in vivo* and *in vitro* omics-level responses of xenobiotic defense proteins in reaction to the exposure to novel salmon feed contaminants including, pesticides, and mycotoxins (Berntssen et al., 2021; Olsvik and Søfteland, 2020; Søderstrøm et al., 2022). We anticipate that, following the generation of *in vivo* and *in vitro* proteogenomics expression data, and the subsequent conversion of Atlantic salmon protein identifiers to their respective model species counterparts, extensive gene-set enrichment and pathway analyses can be performed. These can then be linked to experimentally validated findings to posit novel mechanistic hypotheses concerning the toxicological effects xenobiotics exposure in Atlantic salmon. In addition to the envisaged mechanistic toxicology analyses, future *in vitro* work also will encompass specific case studies for chemicals relevant for food and feed safety assessments to allow for the calibration and validation of QIVIVE model approaches for the Atlantic salmon.

Our results suggest that a proteogenomic characterization of the Atlantic salmon provides a precious foothold to thoroughly assess the function of genes from a mechanistic and structural standpoint. In particular, recent advances in 3D modeling have shown that the more a proteome is soundly characterized, the easier and more fruitful is its investigation (Pedroni et al., 2023). Theoretically speaking, the whole coding part of a genome, which was validated using proteogenomics approaches, may be pipelined into molecular modeling workflows to systematically study the biomechanics of encoded proteins *in silico*. This is particularly relevant for the hazard characterization of xenobiotics of interest to food, feed, and ecosystems research where for rodents, fish, and several farm animals, quantitative activity structure–activity relationship (QSAR) models and physiologically-based kinetic models are being developed (Benfenati et al., 2019; Dorne et al., 2021; Gadaleta et al., 2021; Grech et al., 2019; Lautz et al., 2020c, 2020b, 2020a; Toropov et al., 2017; Toropova et al., 2017).

Benefitting from a OneHealth approach (Gao, 2021), these models can be implemented in risk assessment as part of the NAM batteries using batteries using a weight-of-evidence approach to predict both toxicokinetic and toxicodynamic properties in the absence of toxicity data. In addition, such models are both relevant to the (eco)toxicological assessment of food and feed chemicals in Atlantic salmon but can also serve as exposure inputs for human health risk assessment of chemicals from salmon consumption. This will ultimately contribute to the further advancement of the strongly desired implementation of the 3R strategy in toxicology and risk assessment (Benfenati et al., 2019; EFSA et al., 2017).

5. Conclusion

Our results suggest that the proteogenomic characterization of the Atlantic salmon genome improves upon current annotations and provides a large database on tissue-specific protein expression. The present study also provides evidence that for mechanistic toxicology studies on non-model species, extrapolations of *in vitro* omics data to whole tissue or organism responses are feasible. Our work shows that the proteogenomics workflow presented here yields multi-level omic data which can aid the development of NAMs and further support their use in regulatory science and (eco)toxicology studies. Here, the Atlantic salmon was used as an example, but in light to the rapid advancements in genome sequencing, assembly, and annotation of many non-model organisms, a wider range of application domains of this proteogenomics-based approach can be anticipated.

CRediT authorship contribution statement

M.S. Lin: Methodology, Formal analysis, Software, Writing - original draft. **M.S. Varunjikar:** Writing – original draft. **K.K. Lie:** Writing – original draft. **L. Dellafiora:** Writing – original draft. **R. Ørnsrud:** Writing – original draft. **M. Sanden:** Writing – original draft. **J.L.C.M. Dorne:** Writing – original draft. **V. Bafna:** Formal analysis, Methodology, Writing - original draft. **J.D. Rasinger:** Writing – original draft, Conceptualization, Supervision, Data curation, Project administration.

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

Data will be made available on request.

Acknowledgements

The present work was financed by the Norwegian Research Council (NFR) projects "*AuqaSafe*" (254807), "*AquaMyc*" (281032),

"*Proteogenomics*" (268088), "*ClimSeaFood*" (15804), and the "*MultiOmicsTools*" project (Ministry of Trade, Industry and 574 Fisheries, 15470, IMR). The views expressed here are the authors only and do not reflect the views of the European Food Safety Authority.

Appendix A. Supplementary material

Supplementary File 1. Stepwise protocol of the proteogenomics pipeline developed for the integrated analysis of omics data for mechanistic toxicology studies in non-model species.

Supplementary Figure 1. Peptide-spectrum matches with b-ion and y-ion matches (+1, +2) for proteogenomic events (L)LTVLFQNYA GTDA:ADDSK and (K)(K)LESDTNQLQTEVEEAVQEC+57.021R (related to **Fig. 3**).

Supplementary Figure 2. Peptide-spectrum matches with b-ion and y-ion matches (+1, +2) for proteogenomic events LDNNTGVYVPTGLA AFPN+0.984ELMHVPQSWAR, ETGYLHIQGTKPDSAGC+57.021GVND SPVGLAAYMLEK, and FSTWTDFNNR (related to **Fig. 4**)

Supplementary Figure 3. A. Hierarchical clustering of the PSK normalized expression matrix (23 spectrum files, top 600 proteins) in multi-tissue dataset (related to **Fig. 5**; **Suppl. Table 3**). **B.** Spectra identified in each of twelve tissues. Identification rates, shown as blue text, are defined as the number of identified spectra divided by the total number of spectra for each tissue type.

Supplementary Figure 4. G.O. terms for proteins in multi-tissue dataset (related to **Fig. 5**; **Suppl. Table 4**). Clusters are highlighted in **Suppl. Fig. 3A**.

Supplementary Figure 5. G.O. terms for proteins in liver tissue and primary hepatocyte cultures (related to **Fig. 6**; **Suppl. Table 6**). Clusters are highlighted in **Suppl. Fig. 6B**.

Supplementary Figure 6. A. Spectra identified in primary hepatocyte and liver tissue. Identification rates are shown in blue text. **B.** Hierarchical clustering of the PSK normalized expression matrix (104 spectrum files, top 1,000 proteins) reveals a separation of Salmon primary hepatocyte (green) from Salmon liver tissue (red) samples. Proteins highly expressed in liver tissue, but not primary hepatocyte samples include hemoglobin alpha and beta subunit proteins (orange rectangles; **Suppl. Table 6**).

Supplementary Figure 7. Hierarchical clustering of the PSK normalized expression matrix of Atlantic salmon chemical defensome proteins in multi-tissue dataset (related to Suppl. Table 5).

Supplementary Table 1. Augustus predicted genes and transcripts.

Supplementary Table 2. Augustus predicted transcript CDS overlap with Ensembl or RefSeq CDS, 5'UTR, and/or 3'UTR.

Supplementary Table 3. PSK normalized proteogenomics expression matrix of multi-tissue Atlantic salmon proteins.

Supplementary Table 4. Salmon protein clusters in multi-tissue dataset.

Supplementary Table 5. PSK normalized proteogenomics expression matrix of Atlantic salmon liver and primary hepatocyte proteins.

Supplementary Table 6. Salmon protein clusters in liver tissue and primary hepatocyte cultures.

Supplementary Table 7. Chemical defensome gene patterns from (Eide et al., 2021).

Supplementary Table 8. PAW_BLAST output of Atlantic salmon multi-tissue proteogenomic expression matrix against zebrafish reference proteome.

Supplementary Table 9. PAW_BLAST output Atlantic salmon liver and primary hepatocyte proteogenomic expression matrix against zebrafish reference proteome.

Supplementary Table 10. **Cytochrome P450 (Cyp)**. List of Cyp proteins related to xenobiotic metabolism detected in a PSK normalized proteogenomics expression matrix of Atlantic Salmon liver and primary hepatocytes. ZF, AS, L, and H denote zebrafish, Atlantic salmon, liver, and hepatocytes; M, LCI and UCI describe mean, lower, and upper bound 95% confidence interval of protein expression values. GN and DFSP stand for gene names and chemical defensome pattern, respectively.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2023.108309.

References

Ansong, C., Purvine, S.O., Adkins, J.N., Lipton, M.S., Smith, R.D., 2008. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief. Funct. Genomic. Proteomic. 7, 50–62.

Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., Hartmann, E.M., 2014. Non-model organisms, a species endangered by proteogenomics. J. Proteomics 105, 5–18.

Benfenati, E., Chaudhry, Q., Gini, G., Dorne, J.L., 2019. Integrating in silico models and read-across methods for predicting toxicity of chemicals: a step-wise strategy. Environ. Int. 131, 105060.

Bernhard, A., Rasinger, J.D., Wisløff, H., Kolbjørnsen, Ø., Secher Myrmel, L., Berntssen, M.H.G., Lundebye, A.-K., Ørnsrud, R., Madsen, L., 2018. Subchronic dietary exposure to ethoxyquin dimer induces microvesicular steatosis in male BALB/c mice. Food Chem. Toxicol. 118, 608–625.

Bernhard, A., Rasinger, J.D., Betancor, M.B., Caballero, M.J., Berntssen, M.H.G., Lundebye, A.-K., Ørnsrud, R., 2019. Tolerance and dose-response assessment of subchronic dietary ethoxyquin exposure in Atlantic salmon (*Salmo salar* L.). PLoS One 14, e0211128.

Berntssen, M.H.G., Rosenlund, G., Garlito, B., Amlund, H., Sissener, N.H., Bernhard, A., Sanden, M., 2021. Sensitivity of Atlantic salmon to the pesticide pirimiphos-methyl, present in plant-based feeds. Aquaculture 531, 735825.

Castellana, N., Bafna, V., 2010. Proteogenomics to discover the full coding content of genomes: a computational perspective. J. Proteomics 73, 2124–2135.

Davidson, W.S., Koop, B.F., Jones, S.J.M., Iturra, P., Vidal, R., Maass, A., Jonassen, I., Lien, S., Omholt, S.W., 2010. Sequencing the genome of the Atlantic salmon (*Salmo salar*). Genome Biol. 11, 403.

Dellafiora, L., Dall'Asta, C., 2017. Forthcoming challenges in mycotoxins toxicology research for safer food-a need for multi-omics approach. Toxins 9. https://doi.org/10.3390/toxins9010018.

Dorne, J.L.C.M., Richardson, J., Livaniou, A., Carnesecchi, E., Ceriani, L., Baldin, R., Kovarich, S., Pavan, M., Saouter, E., Biganzoli, F., Pasinato, L., Zare Jeddi, M., Robinson, T.P., Kass, G.E.N., Liem, A.K.D., Toropov, A.A., Toropova, A.P., Yang, C., Tarkhov, A., Georgiadis, N., Di Nicola, M.R., Mostrag, A., Verhagen, H., Roncaglioni, A., Benfenati, E., Bassan, A., 2021. EFSA's OpenFoodTox: an open source toxicological database on chemicals in food and feed and its future developments. Environ. Int. 146, 106293.

Dumas, T., Courant, F., Almunia, C., Boccard, J., Rosain, D., Duporté, G., Armengaud, J., Fenet, H., Gomez, E., 2022. An integrated metabolomics and proteogenomics approach reveals molecular alterations following carbamazepine exposure in the male mussel *Mytilus galloprovincialis*. Chemosphere 286, 131793.

Eide, M., Zhang, X., Karlsen, O.A., Goldstone, J.V., Stegeman, J., Jonassen, I., Goksøyr, A., 2021. The chemical defensome of five model teleost fish. Sci. Rep. 11, 1–13.

Fedoroff, N., Benfey, T., Giddings, L.V., Jackson, J., Lichatowich, J., Lovejoy, T., Stanford, J., Thurow, R.F., Williams, R.N., 2022. Biotechnology can help us save the genetic heritage of salmon and other aquatic species. Proc. Natl. Acad. Sci. U. S. A. 119, e2202184119.

Frankish, A., Uszczynska, B., Ritchie, G.R.S., Gonzalez, J.M., Pervouchine, D., Petryszak, R., Mudge, J.M., Fonseca, N., Brazma, A., Guigo, R., Harrow, J., 2015. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics 16 (Suppl 8), S2.

Gadaleta, D., Marzo, M., Toropov, A., Toropova, A., Lavado, G.J., Escher, S.E., Dorne, J. L.C.M., Benfenati, E., 2021. Integrated in silico models for the prediction of no-observed-(adverse)-effect levels and lowest-observed-(adverse)-effect levels in rats for sub-chronic repeated-dose toxicity. Chem. Res. Toxicol. 34, 247–257.

Gao, P., 2021. The exposome in the era of one health. Environ. Sci. Tech. 55, 2790–2799.

Gillson, J.P., Bašić, T., Davison, P.I., Riley, W.D., Talks, L., Walker, A.M., Russell, I.C., 2022. A review of marine stressors impacting Atlantic salmon *Salmo salar*, with an assessment of the major threats to English stocks. Rev. Fish Biol. Fish. 32, 879–919.

Glover, K.A., Wennevik, V., Hindar, K., Skaala, Ø., Fiske, P., Solberg, M.F., Diserud, O.H., Svåsand, T., Karlsson, S., Andersen, L.B., Grefsrud, E.S., 2020. The future looks like the past: introgression of domesticated Atlantic salmon escapees in a risk assessment framework. Fish Fish. 21, 1077–1091.

Goldstone, J.V., Hamdoun, A., Cole, B.J., Howard-Ashby, M., Nebert, D.W., Scally, M., Dean, M., Epel, D., Hahn, M.E., Stegeman, J.J., 2006. The chemical defensome: environmental sensing and response genes in the *Strongylocentrotus purpuratus* genome. Dev. Biol. 300, 366–384.

Grech, A., Tebby, C., Brochot, C., Bois, F.Y., Bado-Nilles, A., Dorne, J.-L., Quignot, N., Beaudouin, R., 2019. Generic physiologically-based toxicokinetic modelling for fish: integration of environmental factors and species variability. Sci. Total Environ. 651, 516–531.

Gupta, R., Schrooders, Y., Hauser, D., van Herwijnen, M., Albrecht, W., Ter Braak, B., Brecklinghaus, T., Castell, J.V., Elenschneider, L., Escher, S., Guye, P., Hengstler, J. G., Ghallab, A., Hansen, T., Leist, M., Maclennan, R., Moritz, W., Tolosa, L., Tricot, T., Verfaillie, C., Walker, P., van de Water, B., Kleinjans, J., Caiment, F., 2020. Comparing in vitro human liver models to in vivo human liver using RNA-Seq. Arch. Toxicol. https://doi.org/10.1007/s00204-020-02937-6.

Hammer, H., Schmidt, F., Marx-Stoelting, P., Pötz, O., Braeuning, A., 2021. Cross-species analysis of hepatic cytochrome P450 and transport protein expression. Arch. Toxicol. 95, 117–133.

Hampel, M., Alonso, E., Aparicio, I., Santos, J.L., Leaver, M., 2015. Hepatic proteome analysis of Atlantic salmon (*Salmo salar*) after exposure to environmental concentrations of human pharmaceuticals. Mol. Cell. Proteomics 14, 371–381.

EFSA, Hardy, A., Benford, D., Halldorsson, T., Jeger, M.J., Knutsen, H.K., More, S., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Schlatter, J.R., Silano, V., Solecki, R., Turck, D., Benfenati, E., Chaudhry, Q.M., Craig, P., Frampton, G., Greiner, M., Hart, A., Hogstrand, C., Lambre, C., Luttik, R., Makowski, D., Siani, A., Wahlstroem, H., Aguilera, J., Dorne, J.-L., Fernandez Dumont, A., Hempen, M., Valtueña Martínez, S., Martino, L., Smeraldi, C., Terron, A., Georgiadis, N., Younes, M., 2017. Guidance on the use of the weight of evidence approach in scientific assessments. EFSA J. 15, e04971.

Heck, M., Neely, B.A., 2020. Proteomics in Non-model organisms: a new analytical frontier. J. Proteome Res. 19, 3595–3606.

Kelkar, D.S., Provost, E., Chaerkady, R., Muthusamy, B., Manda, S.S., Subbannayya, T., Selvan, L.D.N., Wang, C.-H., Datta, K.K., Woo, S., Dwivedi, S.B., Renuse, S., Getnet, D., Huang, T.-C., Kim, M.-S., Pinto, S.M., Mitchell, C.J., Madugundu, A.K., Kumar, P., Sharma, J., Advani, J., Dey, G., Balakrishnan, L., Syed, N., Nanjappa, V., Subbannayya, Y., Goel, R., Prasad, T.S.K., Bafna, V., Sirdeshmukh, R., Gowda, H., Wang, C., Leach, S.D., Pandey, A., 2014. Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis. Mol. Cell. Proteomics 13, 3184–3198.

Kim, S., Pevzner, P.A., 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 5, 5277.

Krewski, D., Andersen, M.E., Tyshenko, M.G., Krishnan, K., Hartung, T., Boekelheide, K., Wambaugh, J.F., Jones, D., Whelan, M., Thomas, R., Yauk, C., Barton-Maclaren, T., Cote, I., 2020. Toxicity testing in the 21st century: progress in the past decade and future perspectives. Arch. Toxicol. 94, 1–58.

Lautz, L.S., Dorne, J.L.C.M., Oldenkamp, R., Hendriks, A.J., Ragas, A.M.J., 2020a. Generic physiologically based kinetic modelling for farm animals: Part I. Data collection of physiological parameters in swine, cattle and sheep. Toxicol. Lett. 319, 95–101.

Lautz, L.S., Hoeks, S., Oldenkamp, R., Hendriks, A.J., Dorne, J.L.C.M., Ragas, A.M.J., 2020b. Generic physiologically based kinetic modelling for farm animals: Part II. Predicting tissue concentrations of chemicals in swine, cattle, and sheep. Toxicol. Lett. 318, 50–56.

Lautz, L.S., Nebbia, C., Hoeks, S., Oldenkamp, R., Hendriks, A.J., Ragas, A.M.J., Dorne, J. L.C.M., 2020c. An open source physiologically based kinetic model for the chicken (*Gallus gallus domesticus*): calibration and validation for the prediction residues in tissues and eggs. Environ. Int. 136, 105488.

Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930.

Lien, S., Koop, B.F., Sandve, S.R., Miller, J.R., Kent, M.P., Nome, T., Hvidsten, T.R., Leong, J.S., Minkley, D.R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R.A., von Schalburg, K., Rondeau, E.B., Di Genova, A., Samy, J.K.A., Olav Vik, J., Vigeland, M.D., Caler, L., Grimholt, U., Jentoft, S., Våge, D.I., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D.R., Yorke, J.A., Nederbragt, A.J., Tooming-Klunderud, A., Jakobsen, K.S., Jiang, X., Fan, D., Hu, Y., Liberles, D.A., Vidal, R., Iturra, P., Jones, S.J.M., Jonassen, I., Maass, A., Omholt, S. W., Davidson, W.S., 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200–205.

Lundebye, A.-K., Lock, E.-J., Rasinger, J.D., Nøstbakken, O.J., Hannisdal, R., Karlsbakk, E., Wennevik, V., Madhun, A.S., Madsen, L., Graff, I.E., Ørnsrud, R., 2017. Lower levels of Persistent Organic Pollutants, metals and the marine omega 3-fatty acid DHA in farmed compared to wild Atlantic salmon (*Salmo salar*). Environ. Res. 155, 49–59.

Malmstrøm, M., Matschiner, M., Tørresen, O.K., Jakobsen, K.S., Jentoft, S., 2017. Whole genome sequencing data and de novo draft assemblies for 66 teleost species. Sci. Data 4, 160132.

Marowsky, A., Meyer, I., Erismann-Ebner, K., Pellegrini, G., Mule, N., Arand, M., 2017. Beyond detoxification: a role for mouse mEH in the hepatic metabolism of endogenous lipids. Arch. Toxicol. 91, 3571–3585.

Martyniuk, C.J., Denslow, N.D., 2009. Towards functional genomics in fish using quantitative proteomics. Gen. Comp. Endocrinol. 164, 135–141.

Marx-Stoelting, P., Rivière, G., Luijten, M., Aiello-Holden, K., Bandow, N., Baken, K., Cañas, A., Castano, A., Denys, S., Fillol, C., Herzler, M., Iavicoli, I., Karakitsios, S., Klanova, J., Kolossa-Gehring, M., Koutsodimou, A., Vicente, J.L., Lynch, I., Namorado, S., Norager, S., Pittman, A., Rotter, S., Sarigiannis, D., Silva, M.J., Theunis, J., Tralau, T., Uhl, M., van Klaveren, J., Wendt-Rasch, L., Westerholm, E., Rousselle, C., Sanders, P., 2023. A walk in the PARC: developing and implementing 21st century chemical risk assessment in Europe. Arch. Toxicol. https://doi.org/10.1007/s00204-022-03435-7.

McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.-B., Donnelly, P., 2014. Choice of transcripts and software has a large effect on variant annotation. Genome Med. 6, 26.

Mellingen, R.M., Myrmel, L.S., Lie, K.K., Rasinger, J.D., Madsen, L., Nøstbakken, O.J., 2021. RNA sequencing and proteomic profiling reveal different alterations by dietary methylmercury in the hippocampal transcriptome and proteome in BALB/c mice. Metallomics 13. https://doi.org/10.1093/mtomcs/mfab022.

Mellingen, R.M., Myrmel, L.S., Rasinger, J.D., Lie, K.K., Bernhard, A., Madsen, L., Nøstbakken, O.J., 2022. Dietary selenomethionine reduce mercury tissue levels and modulate methylmercury induced proteomic and transcriptomic alterations in hippocampi of adolescent BALB/c mice. Int. J. Mol. Sci. 23, 12242.

Merel, S., Regueiro, J., Berntssen, M.H.G., Hannisdal, R., Ørnsrud, R., Negreira, N., 2019. Identification of ethoxyquin and its transformation products in salmon after controlled dietary exposure via fish feed. Food Chem. 289, 259–268.

Nesvizhskii, A.I., 2014. Proteogenomics: concepts, applications and computational strategies. Nat. Methods 11, 1114–1125.

Nøstbakken, O.J., Rasinger, J.D., Hannisdal, R., Sanden, M., Frøyland, L., Duinker, A., Frantzen, S., Dahl, L.M., Lundebye, A.-K., Madsen, L., 2021. Levels of omega 3 fatty acids, vitamin D, dioxins and dioxin-like PCBs in oily fish; a new perspective on the reporting of nutrient and contaminant data for risk-benefit assessments of oily seafood. Environ. Int. 147, 106322.

Olsvik, P.A., Søfteland, L., 2020. Mixture toxicity of chlorpyrifos-methyl, pirimiphos-methyl, and nonylphenol in Atlantic salmon (*Salmo salar*) hepatocytes. Toxicol. Rep. 7, 547–558.

Pedroni, L., Louisse, J., Punt, A., Dorne, J.L.C.M., Dall'Asta, C., Dellafiora, L., 2023. A computational inter-species study on safrole phase I metabolism-dependent bioactivation: a mechanistic insight into the study of possible differences among species. Toxins 15, 94.

Pineda-Pampliega, J., Bernhard, A., Hannisdal, R., Ørnsrud, R., Mathisen, G.H., Solstad, G., Rasinger, J.D., 2022. Developing a framework for open and FAIR data management practices for next generation risk- and benefit assessment of fish and seafood. EFSA J. 20 https://doi.org/10.2903/j.efsa.2022.e200917.

Prasad, T.S.K., Mohanty, A.K., Kumar, M., Sreenivasamurthy, S.K., Dey, G., Nirujogi, R. S., Pinto, S.M., Madugundu, A.K., Patil, A.H., Advani, J., Manda, S.S., Gupta, M.K., Dwivedi, S.B., Kelkar, D.S., Hall, B., Jiang, X., Peery, A., Rajagopalan, P., Yelamanchi, S.D., Solanki, H.S., Raja, R., Sathe, G.J., Chavan, S., Verma, R., Patel, K. M., Jain, A.P., Syed, N., Datta, K.K., Khan, A.A., Dammalli, M., Jayaram, S., Radhakrishnan, A., Mitchell, C.J., Na, C.-H., Kumar, N., Sinnis, P., Sharakhov, I.V., Wang, C., Gowda, H., Tu, Z., Kumar, A., Pandey, A., 2017. Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. Genome Res. 27, 133–144.

Rasinger, J.D., Carroll, T.S., Lundebye, A.K., Hogstrand, C., 2014. Cross-omics gene and protein expression profiling in juvenile female mice highlights disruption of calcium and zinc signalling in the brain following dietary exposure to CB-153, BDE-47, HBCD or TCDD. Toxicology 321, 1–12.

Rasinger, J.D., Marbaix, H., Dieu, M., Fumière, O., Mauro, S., Palmblad, M., Raes, M., Berntssen, M.H.G., 2016. Species and tissues specific differentiation of processed animal proteins in aquafeeds using proteomics tools. J. Proteomics 147, 125–131.

Rasinger, J.D., Lundebye, A.-K., Penglase, S.J., Ellingsen, S., Amlund, H., 2017. Methylmercury induced neurotoxicity and the influence of selenium in the brains of adult zebrafish (*Danio rerio*). Int. J. Mol. Sci. 18, 725.

Rasinger, J.D., Carroll, T.S., Maranghi, F., Tassinari, R., Moracci, G., Altieri, I., Mantovani, A., Lundebye, A.-K., Hogstrand, C., 2018. Low dose exposure to HBCD, CB-153 or TCDD induces histopathological and hormonal effects and changes in brain protein and gene expression in juvenile female BALB/c mice. Reprod. Toxicol. 80, 105–116.

Rasinger, J.D., Frenzel, F., Braeuning, A., Bernhard, A., Ørnsrud, R., Merel, S., Berntssen, M.H.G., 2022. Use of (Q)SAR genotoxicity predictions and fuzzy multicriteria decision-making for priority ranking of ethoxyquin transformation products. Environ. Int. 158, 106875.

Regueiro, J., Negreira, N., Hannisdal, R., Berntssen, M.H.G., 2017. Targeted approach for qualitative screening of pesticides in salmon feed by liquid chromatography coupled to traveling-wave ion mobility/quadrupole time-of-flight mass spectrometry. Food Control 78, 116–125.

Ruggles, K.V., Krug, K., Wang, X., Clauser, K.R., Wang, J., Payne, S.H., Fenyö, D., Zhang, B., Mani, D.R., 2017. Methods, tools and current perspectives in proteogenomics. Mol. Cell. Proteomics 16, 959–981.

Søderstrøm, S., Lie, K.K., Lundebye, A.-K., Søfteland, L., 2022. Beauvericin (BEA) and enniatin B (ENNB)-induced impairment of mitochondria and lysosomes - potential sources of intracellular reactive iron triggering ferroptosis in Atlantic salmon primary hepatocytes. Food Chem. Toxicol. 161, 112819.

Søfteland, L., Olsvik, P.A., 2022. In vitro toxicity of glyphosate in Atlantic salmon evaluated with a 3D hepatocyte-kidney co-culture model. Food Chem. Toxicol. 164, 113012.

Sprenger, H., Rasinger, J.D., Hammer, H., Naboulsi, W., Zabinsky, E., Planatscher, H., Schwarz, M., Poetz, O., Braeuning, A., 2022. Proteomic analysis of hepatic effects of phenobarbital in mice with humanized liver. Arch. Toxicol. 96, 2739–2754.

Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B., 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32, W309–W312.

Toropov, A.A., Toropova, A.P., Marzo, M., Dorne, J.L., Georgiadis, N., Benfenati, E., 2017. QSAR models for predicting acute toxicity of pesticides in rainbow trout using the CORAL software and EFSA's OpenFoodTox database. Environ. Toxicol. Pharmacol. 53, 158–163.

Toropova, A.P., Toropov, A.A., Marzo, M., Escher, S.E., Dorne, J.L., Georgiadis, N., Benfenati, E., 2017. The application of new HARD-descriptor available from the CORAL software to building up NOAEL models. Food Chem. Toxicol. https://doi.org/10.1016/j.fct.2017.03.060.

Vildhede, A., Wiśniewski, J.R., Norén, A., Karlgren, M., Artursson, P., 2015. Comparative proteomic analysis of human liver tissue and isolated hepatocytes with a focus on proteins determining drug exposure. J. Proteome Res. 14, 3305–3314.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. Journal of Open Source Software. https://doi.org/10.21105/joss.01686.

Wisniewski, J.R., Zougman, A., Nagaraj, N., Mann, M., 2009. Universal sample preparation method for proteome analysis. Nat. Methods 6, 359.

Woo, S., Cha, S.W., Merrihew, G., He, Y., Castellana, N., Guest, C., MacCoss, M., Bafna, V., 2014a. Proteogenomic database construction driven from large scale RNA-seq data. J. Proteome Res. 13, 21–28.

Woo, S., Cha, S.W., Na, S., Guest, C., Liu, T., Smith, R.D., Rodland, K.D., Payne, S., Bafna, V., 2014b. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. Proteomics 14, 2719–2730.

Woo, S., Cha, S.W., Bonissone, S., Na, S., Tabb, D.L., Pevzner, P.A., Bafna, V., 2015. Advanced proteogenomic analysis reveals multiple peptide mutations and complex immunoglobulin peptides in colon cancer. J. Proteome Res. 14, 3555–3567.

Wu, P.-Y., Phan, J.H., Wang, M.D., 2013. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinf. 14 (Suppl 11), S8.

Zhao, S., Zhang, B., 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16, 97.A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16, 97. https://doi.org/10.1186/s12864-015-1308-8.