Universiteit Leiden
The Netherlands

Comparing novel shotgun DNA sequencing and state-of-the-art proteomics approaches for authentication of fish species in mixed samples

Varunjikar, M.S.; Moreno-Ibarguen, C.; Andrade-Martinez, J.S.; Tung, H.S.; Belghit, I.; Palmblad, M.; ... ; Lie, K.K.

Citation

Varunjikar, M. S., Moreno-Ibarguen, C., Andrade-Martinez, J. S., Tung, H. S., Belghit, I., Palmblad, M., … Lie, K. K. (2022). Comparing novel shotgun DNA sequencing and state-of-the-art proteomics approaches for authentication of fish species in mixed samples. *Food Control*, *131*. doi:10.1016/j.foodcont.2021.108417

Version:	Publisher's Version
License:	Creative Commons CC BY 4.0 license
Downloaded from:	https://hdl.handle.net/1887/3250582

Note: To cite this publication please use the final published version (if applicable).

Food Control 131 (2022) 108417

Contents lists available at ScienceDirect

Food Control

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

Comparing novel shotgun DNA sequencing and state-of-the-art proteomics approaches for authentication of fish species in mixed samples

Madhushri S. Varunjikar[a], Carlos Moreno-Ibarguen[b], Juan S. Andrade-Martinez[b], Hui-Shan Tung[a], Ikram Belghit[a], Magnus Palmblad[c], Pål A. Olsvik[a,d], Alejandro Reyes[b], Josef D. Rasinger[a,*], Kai K. Lie[a,**]

[a] Institute of Marine Research, P.O. Box 1870 Nordnes, 5817, Bergen, Norway
[b] Max Planck Tandem Group in Computational Biology, Departament of Biological Sciences, Universidad de Los Andes, Bogotá, Colombia
[c] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, Netherlands
[d] Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway

ARTICLE INFO

Keywords:
Seafood
Fish
Food fraud
Authentication
Mislabelling
DNA-Sequencing
Spectral library

ABSTRACT

Replacement of high-value fish species with cheaper varieties or mislabelling of food unfit for human consumption is a global problem violating both consumers' rights and safety. For distinguishing fish species in pure samples, DNA approaches are available; however, authentication and quantification of fish species in mixtures remains a challenge. In the present study, a novel high-throughput shotgun DNA sequencing approach applying masked reference libraries was developed and used for authentication and abundance calculations of fish species in mixed samples. Results demonstrate that the analytical protocol presented here can discriminate and predict relative abundances of different fish species in mixed samples with high accuracy. In addition to DNA analyses, shotgun proteomics tools based on direct spectra comparisons were employed on the same mixture. Similar to the DNA approach, the identification of individual fish species and the estimation of their respective relative abundances in a mixed sample also were feasible. Furthermore, the data obtained indicated that DNA sequencing using masked libraries predicted species-composition of the fish mixture with higher specificity, while at a taxonomic family level, relative abundances of the different species in the fish mixture were predicted with slightly higher accuracy using proteomics tools. Taken together, the results demonstrate that both DNA and protein-based approaches presented here can be used to efficiently tackle current challenges in feed and food authentication analyses.

1. Introduction

In recent years, a significant increase in food fraud and adulteration has been observed (Moyer et al., 2017). Meat and fish products account for 27% of all reported cases and a high occurrence of mislabelled fish products has been recorded (Bouzembrak et al., 2018; Khaksar et al., 2015). According to European Union (EU) (REGULATION (EU) No 1169/2011), consumers should be properly informed about the contents of the food they consume. In addition to the EU law, food labelling also is addressed by the European Committee for standardization through standard: CWA 17369:2019 – "Authenticity and fraud in the feed and food chain – Concepts, terms, and definitions". To ensure that regulation can be enforced, and standards can be followed, reliable analysis methods must be in place which can correctly detect any fraudulent

Abbreviations: (BSE), Bovine Spongiform Encephalopathy; (qPCR), quantitative Polymerase Chain Reaction; (FDA), Food and Drug Administration; (NGS), Next Generation Sequencing; (COI), Cytochrome *c* Oxidase subunit I; (MS), Mass Spectrometry; (UHPLC-MS/MS), Multi-target Ultra-High-Performance Liquid Chromatography coupled to tandem Mass Spectrometry; (SLM), Spectral Library Matching; (RPMM) Reads Per Million bp of reference genome per Million reads sequenced, (TPP); *Trans*-Proteomic Pipeline, (MGF); Mascot Generic Format, (mzXML) mass to charge ratio in eXtensible Markup Language.

* Corresponding author.
** Corresponding author.
E-mail addresses: madhushri.shrikant.varunjikar@hi.no (M.S. Varunjikar), c.moreno@uniandes.edu.co (C. Moreno-Ibarguen), js.andrade10@uniandes.edu.co (J.S. Andrade-Martinez), hui-shan.tung@hi.no (H.-S. Tung), ikram.belghit@hi.no (I. Belghit), n.m.palmblad@lumc.nl (M. Palmblad), pal.a.olsvik@nord.no (P.A. Olsvik), a.reyes@uniandes.edu.co (A. Reyes), josef.rasinger@hi.no (J.D. Rasinger), kaikristoffer.lie@hi.no (K.K. Lie).

https://doi.org/10.1016/j.foodcont.2021.108417
Received 19 April 2021; Received in revised form 11 June 2021; Accepted 9 July 2021
Available online 11 July 2021
0956-7135/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

labelling of food.

DNA based techniques are commonly used for authentication of food and feed materials and shown to discriminate between closely related taxa including fish (Ivanova et al., 2007; Sawyer et al., 2003; Shokralla et al., 2015; Ward et al., 2005). Targeted methods such as quantitative polymerase chain reaction (qPCR) have been adopted as a standard for identification of bovine material in feed and feed ingredients as part of the European effort to combat the spread of bovine spongiform encephalitis (BSE) (Olsvik et al., 2017) and commonly used for species authentication (Sajali et al., 2020). Although well-designed qPCR assays have been shown to quantify as little as 0.001% (w/w) inclusion of a specific species in a mixture (Kim et al., 2020; Sawyer et al., 2003), targeted multiplex qPCR assays are restricted to detecting a limited number of pre-determined species at each run. DNA barcoding approaches for identification and authentication of fish species of unknown origin have been developed by the Food and Drug Administration (FDA), among others, applying a combination of PCR amplification using degenerate primers and Sanger sequencing for final identification (Yancy et al., 2008). This technique enables the distinction between closely related species in a single product from any type of species (Ivanova et al., 2007; Ward et al., 2005; Yang et al., 2018) depending on the primer design. However, due to its inherent limitations, Sanger sequencing cannot be applied to distinguish different species in mixture samples or to quantify abundance.

Species identification using next-generation sequencing (NGS) has increased in popularity and surpassed the use of Sanger sequencing (Lo & Shaw, 2018). The continuously evolving sequencing technologies allow for massively parallel sequencing of individual amplicons, making authentication of multiple untargeted species within the same sample possible. This has led to the development of methods combining metabarcoding with NGS for accurate identification of species present in a mixture, still involving a PCR step (Hellberg et al., 2017; Lo & Shaw, 2018; Shokralla et al., 2015; Xing et al., 2019). The determination of the relative composition of species in mixture samples such as burger meat or fish cakes gives rise to additional challenges. The combination of metabarcoding and NGS has the potential to determine the presence of different species in a mixture (Bruno et al., 2019; Xing et al., 2019) but this approach often falls short to estimate the correct relative abundance of individual species in the mixture (Hellberg et al., 2017; Lo & Shaw, 2018; Ripp et al., 2014; Shokralla et al., 2015; Xing et al., 2019). The PCR step in the barcoding approach is prone to bias due to its dependency on degenerate primers which assumes equal amplification of target gene from all species. Furthermore, the common use of mitochondrial target genes, such as cytochrome *c* oxidase subunit I (COI), though increases the sensitivity, it also increases the possibility of bias due to fluctuating levels of mitochondrial DNA per cell, tissue, or age (Nagata, 2011; Preuten et al., 2010; Robin & Wong, 1988). Although larger barcoding amplicons would solve some of the issues concerning specificity and false discoveries, larger amplicons are also more sensitive to DNA degradation (Hird et al., 2006). Thus, avoiding the PCR step altogether would be beneficial for accurately quantifying the biological content of mixture food products. Recent approaches using shotgun metagenome sequencing have successfully quantified the content of mixture products demonstrating the potential for this technique in food and feed control (Haiminen et al., 2019; Kobus et al., 2020; Ripp et al., 2014). Due to the massive parallel sequencing of short reads, this approach also will be less prone to bias due to processing mediated DNA degradation.

For highly processed food materials (e.g. thermally and acid-treated samples), species identification using protein-based methods represent a suitable alternative to established DNA-based methods (Carrera et al., 2013a). Different proteomics approaches have been developed for accurate species identification from processed food and feed products and mixtures; currently, several laboratories are developing proteomics-based tools and analysis protocols for quality assessment and food safety analyses (Belghit et al., 2019; Carrera et al., 2013b; Lecrenier et al., 2021; Nessen et al., 2016; Ohana et al., 2016; Rasinger et al., 2016; Wulff et al., 2013). Standard bottom-up proteomics commonly involves gel-based or gel-free separation of proteins and identification of proteins with specific mass spectrometry profiles of marker peptides or proteins (Rasinger et al., 2016; Wulff et al., 2013). Current methods used for food and feed authentication rely on species-specific peptide markers for which sequence information is available (Carrera et al., 2013b; Lecrenier et al., 2016, 2021; Steinhilber et al., 2018). However, targeted mass-spectrometry (MS) methods are at times difficult to implement, as reference proteomes of non-model species are not readily available (Belghit et al., 2019; Rasinger et al., 2016). Therefore, alternative approaches based on proteome-wide tandem mass spectrometry and spectral library matching (SLM) for the identification of species have been developed and implemented by several laboratories for food and feed fraud detection in processed meat, seafood, and processed animal proteins (PAPs), respectively (Belghit et al., 2019; Carrera et al., 2013b; Ohana et al., 2016; Rasinger et al., 2016; Wulff et al., 2013).

Non-targeted database-agnostic proteomics approaches have been used previously for fish species authentication; a total of 47 fish samples were correctly identified in both fresh and processed samples derived from 22 different species of fish (Wulff et al., 2013). Applying the SLM proteomics method on closely related flatfish species were correctly identified species in both processed and fresh samples (Nessen et al., 2016), demonstrating that MS is a promising tool for species authentication.

In the present study, based on shotgun DNA sequencing data of seven teleost fish species (*Melanogrammus aeglefinus, Oreochromis niloticus, Gadus morhua, Salmo salar, Esox lucius, Pangasianodon hypophthalmus, Xiphophorus maculatus*), a bioinformatic pipeline and a condensed reference library for quantification of relative abundance of fish species in fish mixture samples were developed. In addition, high resolution (HR) MS data were generated, a spectral library collection was compiled and it was tested if previously developed proteomics-based methods (Nessen et al., 2016; Ohana et al., 2016; Wulff et al., 2013) also allow for differentiation of individual species and abundance estimates of a complex fish mixture. Based on the genomics and proteomics data obtained, the strengths and weaknesses of these two complementary approaches when screening for food fraud in fish mixtures were discussed and a combined strategy of analyses to tackle current seafood authentication challenges is introduced.

2. Materials and methods

2.1. Sampling and animals

A total of seven teleost species were analyzed; namely, Atlantic cod (*Gadus morhua*), Atlantic haddock (*Melanogrammus aeglefinus*), Nile tilapia (*Oreochromis niloticus*), Northern pike (*Esox lucius*), Atlantic salmon (*Salmo salar*), platyfish (*Xiphophorus maculatus*) and pangasius (*Pangasianodon hypophthalmus*) which will hereafter be referred to as cod, haddock, tilapia, pike, salmon, platyfish, and pangasius, respectively. Individual fish species were purchased from a commercial vendor except for a pike, which was donated by a local recreational fisherman. Species assignations of fish were validated through visual inspection by a trained ichthyologist in addition to genetic verification. Prior to DNA and protein extraction, fish were frozen and stored at −20 °C. For the fish mixture, muscle tissues from platyfish, tilapia and cod were weighed and mixed in the following ratios: platyfish 1/6, tilapia 2/6 and cod 3/6, forming a mixed tissue sample ("fish mixture"). The tissue samples were flash-frozen on dry ice and ground to a fine powder using a mortar and pestle. The mortar was kept on dry ice during the entire grinding and homogenization process.

2.2. DNA sample preparation

2.2.1. DNA extraction

DNA from individual fish were extracted from 40 to 50 mg of muscle tissues using the DNeasy Blood & Tissue kit (Qiagen) according to the manufacturer's instructions. The DNA concentration was determined at 260/280 nm (DNA-50) using a Nanodrop ND-1000 spectrometer and Qbit dsDNA BR assay kit (Invitrogen, ThermoFisher). For visual validation of DNA integrity, 500 ng of DNA were run on a 1% (w/v) agarose gel. DNA from 50 mg of grinded fish mixture tissue sample was extracted from three replicate samples using DNeasy Blood &Tissue Kit, Qiagen, according to the manufacturer's instructions.

2.2.2. DNA sequencing

The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no). TruSeq PCR free library kit (Illumina Inc., CA, USA) was used to construct DNA libraries from each of the fish mixture and individual DNA samples following the manufacturer's protocol. For DNA library prep, 1.5 μM of DNA was used for the construction of each individual library. All libraries were tested using qPCR for quantification prior to sequencing on Illumina HiSeq 2500 (Illumina Inc.) using V4 clustering and sequencing reagents according to the manufacturer's instruction. Library preparation and sequencing were done by the Norwegian Sequencing Centre, Oslo, Norway. Image analysis and base calling were both performed using Illumina's RTA software

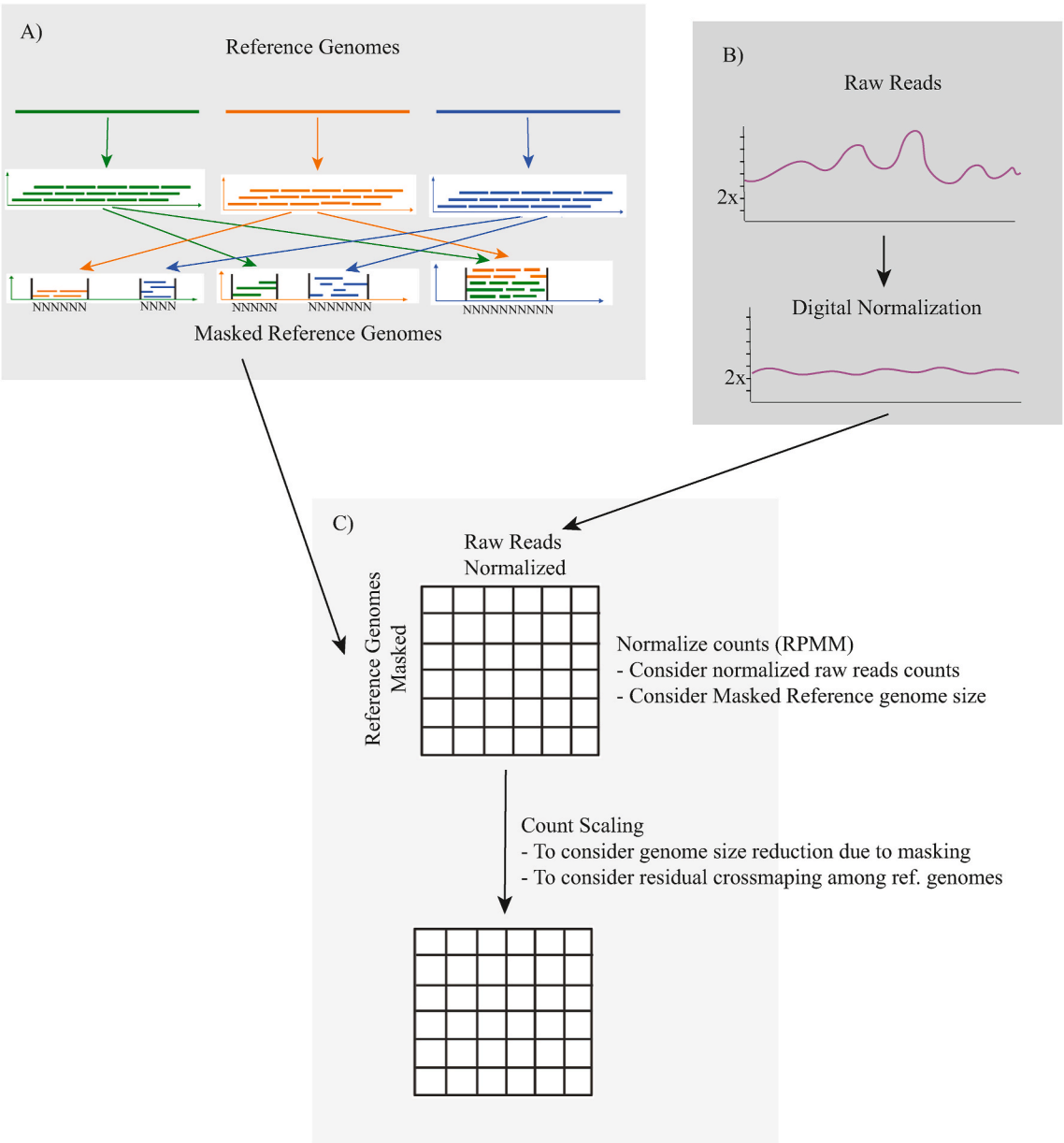

Fig. 1. Workflow of bioinformatics pipeline used for DNA sequencing analyses before calculating percentages. **(A)** Reference genome masking was conducted by generating a set of simulated reads from each genome followed by cross-mapping against all other reference genomes. Any identification of cross-mapping was masked (characters replaced by N's) to avoid cross-matching between species, leaving a masked genome with unique sequences for each species. This process was repeated three times in cases of the presence of duplicated regions and gene families. **(B)** Prior to the mapping of QC-controlled fish samples, a digital normalization was conducted to account for uneven sequence coverage throughout the length of the genome (as reflected by the peaks in the figure), which could be a major artefact considering extensive reference masking due to closely related species. **(C)** The last step is to calculate abundance estimation of read counts using RPMM (reads per million bp of reference genome per million reads sequenced) followed by a scaling process in order to account for the reduction in genome size after masking and the residual cross-mapping observed in the simulated genomes.

version 1.18.66.3. Low-quality reads were removed using Illumina's default chastity criteria. Compressed base call files (.bcl) were demultiplexed and converted to fastq files using the bcl2fastq software version 2.17.1.14. The quality of each library/fastq file was assessed using fastqc embedded in the bcl2fastq software. Between 8 and 11 M paired-end 125 bp reads were obtained from each sample (individual fish or fish mixture). Raw reads have been deposited to the SRA library (https://www.ncbi.nlm.nih.gov, BioProject accession number PRJNA716500).

2.2.3. Raw data cleanup and reference genomes retrieval

Paired-ends were cleaned for adapter contamination and low-quality bases (phred score below 20) using Trimmomatic (version 0.38 (Bolger et al., 2014),) with default parameters and a minimum length of 50. All reads with the presence of N's were removed to guarantee the quality of the sequences employed. In order to map and quantify the sequenced reads, the latest versions of the available fish reference genomes were retrieved (Supplementary Table 1).

For pangasius, no reference genome was available. Therefore, sequencing data generated in the present work were used to assemble a draft genome. SPADES (version 3.9.0 (Bankevich et al., 2012),) with default parameters was used for assembling. The resulting genome assembly together with the other retrieved reference genomes was checked for completeness and contamination using BUSCO (version 3.0.2 (Seppey et al., 2019, pp. 227–245),).

2.2.4. Reference genome de-replication and generation of simulated datasets

To avoid cross-mapping between reference genomes the conserved regions among related genomes were removed, generating a set of de-replicated reference genomes. For this, first, a set of simulated datasets was generated from the genome sequences by extracting sequence fragments of 100bp along the genome with a sliding window of 60bp (Fig. 1). Thus, every part of the genome had at least a 3x coverage. The set of simulated reads were mapped against all other reference genomes using bowtie as described below (2.2.5); any regions with positive mapping were masked using Bedtools (version 2.25.0 (Quinlan & Hall, 2010),). This procedure was repeated two additional times until the number of cross-mappings among the simulated genomes was minimal.

2.2.5. Mapping of raw reads and normalization

All simulated and generated sequencing reads were mapped using bowtie 2 (version 2–2.2.4 (Langmead & Salzberg, 2012),) with default parameters in the very fast and global alignment setup. The reads were mapped in paired-end mode keeping only the best match hit and only the number of matching pairs mapped were used for follow-up calculations in order to reduce potential false-positive mapping of single reads.

To account for potential variation when sequencing real samples, in which commonly not all regions of a genome are evenly sequenced, a digital normalization to an average coverage of 2x was implemented in the present study. To do so, BBNorm (version 37.57 (Bushnell et al., 2017),) was run with the prefilter option set to true and with target coverage set to 2. Fastq files for paired-end reads were normalized simultaneously using the paired-end functionality of BBNnorm. The resulting number of normalized reads were used as the library size for the RPMM normalization (see the following section).

2.2.6. Final mapping counts cleanup

After mapping the digitally normalized samples against the de-replicated genomes using the same procedure described above (see section 2.2.5), the RPMM (reads per million bp of reference genome per million reads sequenced; i.e. the abundance was normalized to the depth of sequencing and the variation in the length of the reference genomes) counts for each fish were calculated after subtracting the estimated residual cross-mapping among the reference genomes, thus considering potential false-positive mapping. Furthermore, the RPMM counts were adjusted for masked genomes and the genome size scaled to only the mappable nucleotide count, i.e., discarding the N's. Finally, a conversion factor was used to scale from the RPMMs obtained on the de-replicated genomes to the ones from the original reference genomes. This process was performed using a custom Perl script available upon request.

2.2.7. Skmer comparisons

To estimate genomic distances from the mappings and identify their closest match in the reference genomes, Skmer (version 3.0.2 (Sarmashghi, 2019),) was run using default parameters.

2.3. Proteomics analysis

2.3.1. Extraction, solubilization and quantification of proteins

Fish muscle tissue (100 mg) were weighed into test tubes of the PlusOne Sample Grinding kit (GE Healthcare Life Science, 80648337, Piscataway, NJ, USA) and solubilized with 1 mL lysis buffer (4% SDS, 0.1 M Tris-HCl, pH 7.6). Samples were kept on ice, homogenized and 1 M Dithiothreitol was added to obtain a final concentration of 0.1 M. Samples were centrifuged for 10 min at 15,000 g to remove resin and other debris. Supernatants were collected, heated at 95 °C for 5 min, centrifuged once again. The remaining supernatants were eventually collected into new tubes and stored at −20 °C until further processing. Protein concentrations of extracted samples were determined using a Pierce 660 assay (ThermoFisher Scientific) following the vendor's instructions. Fish mixture sample was prepared using extracted proteins in the following ratios: platyfish 1/6, tilapia 2/6 and cod 3/6.

2.4. In-solution digestion of proteins

Protein extracts were prepared for mass spectrometric analysis as described in Belghit et al. (2019). In short, following a Filter Aided Sample Preparation (FASP) digestion protocol (Wiśniewski, 2016), 40 μg of extracted proteins were diluted with 200 μL of 8 M urea solution prepared in Tris-HCl (100 mM, pH 8.5) and transferred to ultrafiltration spin column (Microcon 30, Millipore, Burlington, MA, USA). Proteins were alkylated with 50 mM of iodoacetamide (C2H4INO) for 20 min in the dark at room temperature. Subsequently, protein mixtures in the column were washed with 200 μL of 8 M urea solution along with 100 μL of 50 mM ammonium bicarbonate (NH4HCO3). Trypsin was added to the filters (1:50 enzyme to protein ratio), and tubes were incubated for 16 h at 37 °C. Filters were centrifuged and washed (40 μL of 50 mM ammonium bicarbonate solution followed by 0.5 M NaCl). Following a final centrifugation step, peptide concentration in the eluates was determined using a Nanodrop (Thermo Scientific). Subsequently, eluates were vacuum dried and stored at −20 °C.

2.5. Mass spectrometry

Digested peptide samples were analyzed at the Proteomics Unit at the University of Bergen, Norway (PROBE) as described in Bernhard et al. (2019). In short, dried peptides were dissolved in 2% acetonitrile (ACN) and 0.1% formic acid (FA). Samples were injected into an Ultimate 3000 RSLC system (Thermo Scientific, Sunnyvale, California, USA) connected to a linear quadrupole ion trap-orbitrap (LTQ-orbitrap Elite) mass spectrometer (Thermo Scientific, Bremen, Germany), equipped with a nanospray Flex ion source (Thermo Scientific). Raw data obtained in data-dependent-acquisition (DDA)-mode was analyzed as described below.

2.6. Proteomics bioinformatics

Using msConvert (version: 3.0., ProteoWizard (Kessner et al., 2008),) Thermo. raw files were converted to. mgf and. mzXML formats. Raw and

processed mass spectrometry data were deposited in an online repository (MSV000087017 (massive.ucsd.edu/ProteoSAFe)). For molecular phylogenetic analyses using compareMS2 (Palmblad & Deelder, 2012),.mgf files containing the top 500 most intense tandem mass spectra were created using msConvert (version: 3.0., ProteoWizard (Kessner et al., 2008)). The output of compareMS2 was used to create distance matrices and UPGMA trees in MEGA (version 10 (Palmblad & Deelder, 2012; Wulff et al., 2013)). For identification of peptides, tandem mass spectra were searched against UniProt *Danio rerio* reference proteomes (UP000000437 accessed on January 2021) using Comet (Eng et al., 2013) as implemented in the *Trans*-Proteomic Pipeline (TPP) (version 5.2.0 (Deutsch et al., 2015)) and shown in Fig. 2. In all searches, precursor mass tolerance was set to 20 ppm, trypsin was selected as a digestive enzyme (allowing for two non-enzymatic termini), and carbamidomethylation of carbon and oxidation of methionine were set as fixed and variable modification, respectively. Generated pepXML files were further analyzed using PeptideProphet (Keller et al., 2002). Based on mzXML and pepXML files, spectral libraries were created for each of the seven fish species using SpectraST (version 5.0 (Lam, 2011)). Subsequently, spectra from all fish species in the set were cross-matched against all spectral libraries created and dot products were calculated (Lam, 2011); a dot product of one indicates that spectra are identical whereas a dot product of zero indicates that spectra are mismatching (Belghit et al., 2021). Matching spectra with dot products above 0.8 were considered to be valid matches and the unique identifiers of these spectra were extracted and exported into a text file (spectra counts as given in Supplementary Table 6 A and Table 4). Using these text files, original mzXML files were filtered to remove contaminant-, common peptide- or non-peptide-spectra; filtered files were then searched against the UniProt *Danio rerio* reference proteome (UP000000437 accessed on January 2021) using Comet, as mentioned above. Based on these filtered data, the second set of masked spectral libraries were created using SpectraST (version 5.0 (Lam, 2011)). The fish mixture sample was matched against both raw and masked spectral library of each fish species for relative quantification of the percentage contribution of fish species to the mixture as shown in Fig. 2. Dot products above 0.9 or higher were considered valid matches and used for quantification. The percentage of fish in the mixture was calculated using R (version 3.6.1). Outputs were recoded using tidyverse functions (version 1.3.0 (Wickham et al., 2019)) and UpSetR (version 1.4.0). All R code is available on request from the authors.

3. Results

3.1. Genomic relatedness analysis of pure samples

In order to establish if the fast-genomic comparison could help identify relatedness between muscle tissue samples obtained from seven different fish species, Illumina sequencing reads of individual samples were compared in a pair-wise fashion among all possible comparisons using Skmer. Results show that very high similarity obtained between forward and reverse reads from the same samples; additionally, samples that were generated from closely related species such as cod and haddock appear closer together in the obtained dendrogram (Fig. 3A). Furthermore, comparisons of the samples against the reference genomes also show a high similarity and clustering with their corresponding reference genomes (Fig. 3B).

To identify and calculate the relative abundance from each of the different species present in individual and fish mixture samples, available reference genomes for each fish species of interest were retrieved (Supplementary Table 1). In the case of the pangasius, no reference genome was available; thus, an assembled draft version using SPADES was used. However, comparing the completeness of the different genomes (Supplementary Table 1) it was clear that the assembled pangasius genome resulted in very low completion, likely due to the very low sequencing coverage. The genomes of the remaining six species, even though not perfectly assembled, had metrics of high enough quality that allowed for comprehensive mapping. Of note, salmon showed a high level of duplicated genes (34.98%), which was found to be in agreement with a recent genomic duplication that occurred in the ancestor of this species 80 million years ago (Lien et al., 2016). In addition, it was noted that the haddock genome was the most

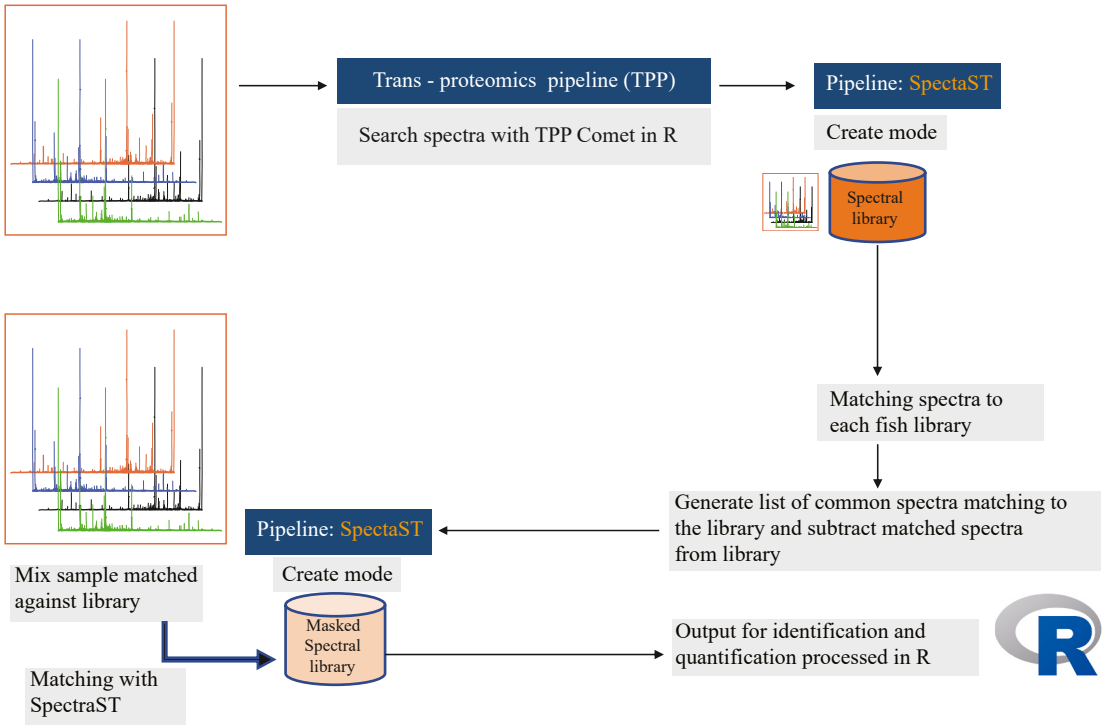

Fig. 2. Representation of proteomics bioinformatics methods used for calculation of percentages in the fish samples using spectral library workflow, where *Trans*-Proteomic Pipeline (TPP) was used for searching spectra and creating libraries as well as searching against the libraries.

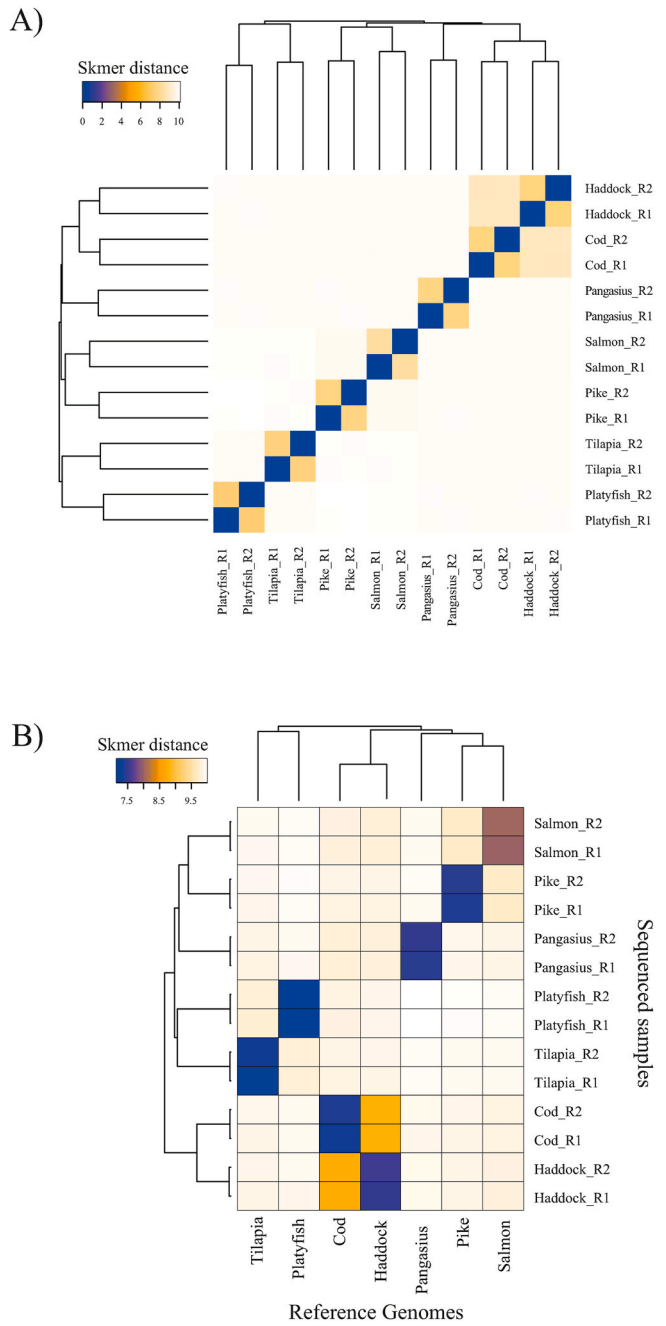

Fig. 3. Sample relatedness based on DNA sequencing using Skmer. (A) Analysis of pure samples with Skmer; note that forward (R1) and reverse (R2) reads from each sample are used, strong relatedness is identified from the corresponding paired sample. (B) Comparison of pure tissue samples against all the reference genomes shows the clustering of the samples with their reference genome. Note that scale bars are different for A and B, in both cases, they represent distances as determined by Skmer.

fragmented one of all genomes obtained online; it comprised only ~60% complete and ~32% fragmented genes.

Using the simulated reads generated from each reference genome, the cross-mapping was evaluated among reference genomes, which provided an estimate of the closeness and redundancy presented among the targeted fish genomes. As it can be seen in Table 1, the genomes of cod and haddock displayed the highest rate of cross-mapping, with close to 50% of the reads from one genome mapping to the other genome. By comparison, the cross-mapping among other genomes was relatively low; in general values of less than 1% were observed. To increase the accuracy of quantification, three rounds of masking for each reference genome present in the set were performed.

When mapping simulated reads against masked reference genomes (Mask-3), a significant reduction in the cross-mapping can be observed (Supplementary Table 2 C). A normalization step was performed in order to consider potential coverage variation on the real datasets, where certain regions of the genome could have more coverage than others, likely affecting the quantitation of the mapping. A digital normalization using *k*-mers to a 2x expected normalization was performed. Using the simulated dataset, it was possible to observe that the digital normalization had no effect on the raw datasets (Supplementary Table 2 A and B) but had minor variation in the masked genomes (Supplementary Table 2 C and D). Thus, it was needed to apply a final scaling factor to take into consideration the subtraction for the estimated cross-mapping and the re-scaling to the unmasked genome size; with this scaling, it was possible to obtain minimal cross-mapping counts while retaining the un-masked original mapping counts (Supplementary Table 2 E and F). Eventually, a final mapping and counting strategy was developed, which could be applied to all samples investigated in the present study. For this final strategy the reported numbers were normalized to the sequencing effort and the genome size, thus, are reported in RPMM (Reads Per Million bp of reference genome per million sequenced reads), see Supplementary Table 3 for equivalent results to Supplementary Table 2 but in RPMMs.

Following quality filtering and digital normalization, the reads of muscle tissues of seven individual fish species were mapped to the masked reference genomes and quantified according to the strategy described above. As can be seen in Table 2, despite several rounds of masking, a small degree of residual cross-mapping between closely related fish species was observed; in particular between cod and haddock. This observation is likely due to either (i) intra-species variation between the reference genome and the samples used or (ii) incompleteness of the reference genomes, as observed by the fact that the haddock reference genomes had a high amount of fragmented single-copy orthologs identified. Some low negative values were obtained due to the normalization effect; however, those counts were always very close to zero (Table 2).

3.2. Quantitation of fish mixture -DNA method

In addition to the fish mixture samples created by mixing muscle tissues of three fish; (N = 4), an additional set of samples was generated by mixing defined proportions of DNA post-extraction (N = 3). The quantitation of such fish mixtures revealed that mixing the DNA was able to recover the expected mixture ratio with minor divergence from expected values (Table 3, for RPMM counts see Supplementary Tables 4 and 5), demonstrating the accuracy of the method. Despite taking great care in homogenizing the samples, the observed variation within the tissue mixture group could be result of incomplete homogenization. This highlights the importance of the sample preparation step for obtaining reliable data.

3.3. Proteomic relatedness analysis of individual fish muscle samples with compareMS2

Using compareMS2, a phylogenetic tree was constructed based on the top n = 500 tandem mass spectra obtained from muscle samples of the seven fish species. All fish species were separated and branched according to their respective phylogeny (Fig. 4). In accordance with DNA data, a strong relatedness of cod and haddock was observed, which were placed on the same branch, while pangasius was placed on a different branch of the obtained tree.

3.4. Quantitation of fish mixture -proteomics method

Using SpectraST, tandem mass spectra of a representative fish

Table 1
Percent cross-mapping between species.

Reference genome	Mapping before the first masking						
	Sim-Cod	Sim- Haddock	Sim- Pike	Sim- Platyfish	Sim- Salmon	Sim-Tilapia	Sim- Pangasius
RG_ Cod	NA	46.15	1.41	0.18	1.39	0.21	0.04
RG_ Haddock	49.68	NA	2.00	0.19	2.83	0.26	0.04
RG_ Pike	0.94	2.22	NA	0.14	4.35	0.25	0.09
RG_ Platyfish	1.30	2.74	0.74	NA	1.42	0.74	0.03
RG_ Salmon	1.36	3.14	3.94	0.17	NA	0.27	0.09
RG_ Tilapia	0.84	1.77	0.80	0.88	1.60	NA	0.05
RG_ Pangasius	0.49	0.84	1.21	0.03	1.68	0.85	NA

[a]RG: Reference Genome; Sim: Simulated reads from the reference genome. Mapping simulated reads against individual whole-genome sequences; before any masking was performed. NA indicates perfect matching between library which is invalid as the sample inside the library is the same as the matching samples.

Table 2
Cross-mapping between species following genome masking.

RPKM	Cod	Haddock	Pangasius	Salmon	Pike	Tilapia	Platyfish
Cod	0.64	0.04	0.00	0.00	0.00	0.00	0.00
Haddock	0.02	0.66	0.00	0.00	0.00	0.00	0.00
Pangasius	0.00	0.00	2.41	0.00	0.00	0.00	0.00
Salmon	0.00	0.00	0.00	0.38	0.00	0.00	0.00
Pike	0.00	0.00	0.00	0.00	1.05	0.00	0.00
Tilapia	0.00	0.00	0.00	0.00	0.00	0.96	0.00
Platyfish	0.00	0.00	0.00	0.00	0.00	−0.01	1.48

[a]Values are stated as reads per kilobase million (RPKM).

Table 3
Quantitation of fish mixture (N = 4) and DNA mixture in percentage (N = 3), data are presented as means ± SD.

		Cod	Tilapia	Platyfish	Haddock	Pangasius	Pike	Salmon
	Expected (%)	50	33	17	–	–	–	–
Fish fillet mixture	**Match (%)**	53 ± 17	27 ± 13	16 ± 4	3 ± 1	0 ± 0	0 ± 0	0 ± 0
	Divergence	3	−3	1	−3	0	0	0
Fish DNA mixture	**Match (%)**	45 ± 1	39 ± 1	13 ± 0	4 ± 0	0 ± 0	0 ± 0	0 ± 0
	Divergence	−5	6	−4	0	0	0	0

[a]Fish fillet mixture - muscle tissues from platyfish, tilapia and cod were weighed and mixed; Fish fillet mixture - fillet from platyfish, tilapia and cod were mixed; Fish DNA mixture - DNA from platyfish, tilapia and cod were mixed; Expected (%) - platyfish 1/6, tilapia 2/6 and cod 3/6, forming a mixed tissue sample or DNA samples; Divergence - represents divergence from the expected percentages values in the mixture (% expected - % match), values were calculated for both fish fillet and DNA mixtures.

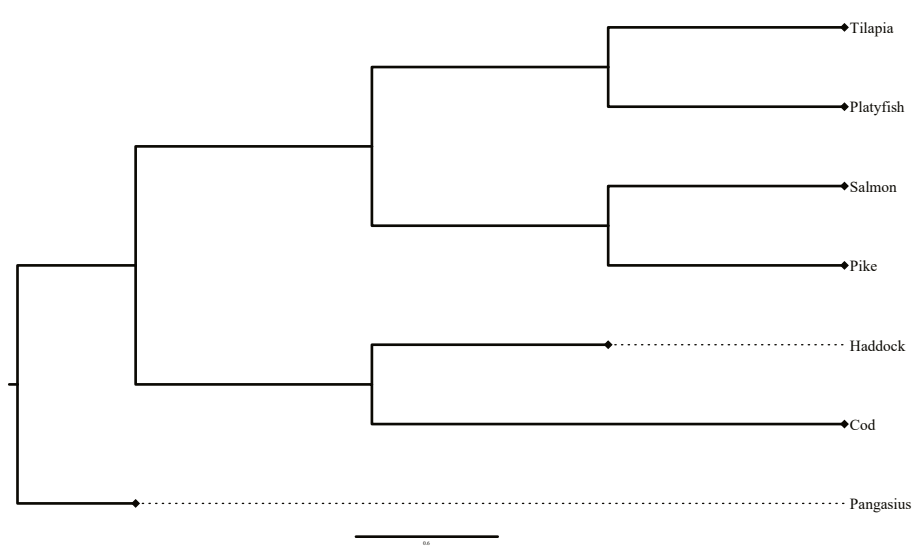

Fig. 4. Phylogenetic tree built from compareMS2 with top 500 spectra, which agrees with the phylogeny of the selected fish species. Scientific names of the species are as given here: Haddock (*Melanogrammus aeglefinus),* Tilapia (*Oreochromis niloticus),* Cod *(Gadus morhua),* Salmon *(Salmo salar),* Pike (*Esox lucius),* Pangasius (*Pangasianodon hypophthalmus),* Platyfish (*Xiphophorus maculatus*).

mixture sample comprising three different fish species (platyfish 1/6, tilapia 2/6 and cod 3/6) were matched against spectral library reference collection built from the seven fish species analyzed in the present study. Using a dot product cut off of 0.9, the percentage of each fish species in the mixture was determined (example matches reported in Supplementary Table 7). The results suggest that the fish mixture sample contained 23% (w/w) cod, which is lower than the nominal relative amount added to the fish mixture. The fish mixture sample also was found to contain 24% (w/w) tilapia and 18% (w/w) platyfish, which, when compared to the relative nominal concentrations of these fish in the fish mixture samples, represent an under- and overestimation, respectively of fish muscle tissues in the mixed sample (Table 5). On a taxonomic scale, cod and haddock belong to the same family, the gadoids. When quantifying protein data on the taxonomic family level, the data predicts a 47% inclusion level of gadoids (cod + haddock) in the sample, very close to the expected 50% of a gadoid fish added to the fish mixture. An example output of is given in Supplementary Table 7.

4. Discussion

Predicting the relative species composition of complex food and feed mixtures remains a major challenge for regulatory scientists and food authorities. The present study shows that, for single-species analysis, both the novel shotgun DNA sequencing approach based on masked reference libraries and recently introduced MS-based proteomics approaches can distinguish between closely related fish species within the same taxonomic infraclass (*Teleostei*), clade (*acanthomorphata*) and within the same family (*gadidae*), respectively.

DNA has traditionally been used for the taxonomic classification of animal species, either by whole-genome sequencing or relying on mitochondrial genomes (Kahlke & Ralph, 2019). MS-based proteomics approaches based on collection and analysis of tandem mass spectra were applied successfully for species- and tissue-specific classification of both raw and heavily processed samples (; Belghit et al., 2019, 2021; Nessen et al., 2016; Ohana et al., 2016; Rasinger et al., 2016; Steinhilber et al., 2018; Wulff et al., 2013). While authentication of pure fish muscle samples using either DNA or MS-based proteomics already has been reported on in literature (Nessen et al., 2016; Ward et al., 2005; Wulff et al., 2013; Yancy et al., 2008), in the present study, for the first time, both approaches are applied on the same sample set. In addition to individual pure fish muscle tissue, mixtures of fish samples were analyzed to test the applicability of both approaches in the context of authenticity testing of fish mixtures such as fish cakes and other seafood products commonly sold in Norwegian markets.

Shotgun DNA sequencing and mapping towards a masked reference library gave an approximate estimate of the percent inclusion of each species in mixed fish tissue samples and samples of fish-DNA mixed in the same ratio as the tissues (Table 3). Although some deviation from the expected ratio was observed, DNA shotgun sequencing in combination with masked reference libraries demonstrated its usefulness for disclosing species substitution and adulteration in a mixed seafood product. This implies that the DNA-based workflow presented here also could be applied to identify species in other mixed food products.

Commonly, for authentication and relative abundance estimation of species in mixtures metabarcoding in combination with NGS has previously been applied (Bruno et al., 2019; Leonard et al., 2015; Voorhuijzen-Harink et al., 2019). While metabarcoding approaches, in general, has been shown to predict species combinations with relatively high accuracy, it tends to fall short in predicting relative abundances (Xing et al., 2019). This shortcoming is mainly due to PCR bias and other method-intrinsic challenges as listed in the introduction. One advantage of the metabarcoding approach when compared to both methods presented here (i.e. shotgun DNA sequencing based on masked reference libraries and untargeted MS-based proteomics), is the availability of reference material sequences in public databases. At the time of writing, 321 k species were listed in the BOLD database (https://www.boldsystems.org/), a cloud-based analysis platform developed to support the generation and application of DNA barcode data. However, the number of publicly accessible whole-genome assemblies has been increasing exponentially in the recent past, paving the way for analytical approaches utilizing whole-genome data; currently, 599 fish genomes are available for download (https://www.ncbi.nlm.nih.gov/, 2021).

Previously reported shotgun sequencing approaches such as the all-food-seq (AFS) and the FASTER pipelines have shown great potential for estimation of species abundances in mixtures of land animals, showing high accuracy and low false discovery rates (Hellmann et al., 2020; Kobus et al., 2020; Ripp et al., 2014). However, the AFS was restricted to comparisons of only 10 complex genomes. Whereas, another k-mers based approach showed accuracy comparable to the workflow presented in the present study and also has the potential to be applied to an unlimited number of genomes (Kobus et al., 2020).

In short, all of the studies listed above, highlight the usefulness of DNA-based tools for the identification and quantification of species from a variety of taxonomic kingdoms and phyla, including animals, plant and bacteria, in one single mixture sample (Hellmann et al., 2020; Kobus et al., 2020; Ripp et al., 2014). Combining DNA sequencing with masked reference libraries offers the possibility to analyse mixed samples with high accuracy using limited computational resources and small reference libraries. This, in combination with the nano-sequencing approach e.g. miniaturized DNA sequencing devices such as MinION developed by Oxford Nanopore Technologies or Sequel II by PacBio (Huo et al., 2021), in the near future, open the possibility for rapid on-site analyses of a fixed set of targets (Voorhuijzen-Harink et al., 2019).

The MS-based proteomics spectral library matching (SLM) approach also yielded promising results (Table 5) when estimating the relative abundance of fish species in a mixture; especially, on the family level. Since protein and water constitute the bulk of muscle tissue in terms of mass, one would expect a higher accuracy in predicting species contribution of mixed tissue samples using SLM approach compared to DNA. In terms of accuracy, the calculated relative abundance of platyfish was in accordance with the relative amount added to the protein fish mixture while the concentration of tilapia was underestimated. When summarizing the results on the taxonomic (family) level SLM predicted a 47% inclusion of gadoids (cod and haddock in this case), which is very close to the expected 50% cod protein added to the fish mixture. As cod and haddock belong to the same family, highly conserved proteins and

Table 4
Matching of fish species against each spectral library.

Library	Cod[a]	Haddock[a]	Pangasius[a]	Pike[a]	Platyfish[a]	Salmon[a]	Tilapia[a]
Cod	NA	27.2	11.2	12.0	0.126	12.2	10.6
Haddock	25.9	NA	12.6	13.3	15.0	13.2	12.3
Pangasius	9.9	12.1	NA	16.3	15.2	12.4	12.9
Pike	10.7	12.4	13.9	NA	13.7	22.5	10.7
Platyfish	12.3	14.8	15.5	14.4	NA	12.3	16.1
Salmon	12.3	13.5	12.0	21.4	12.2	NA	9.9
Tilapia	10.8	14.0	14.9	12.5	18.1	11.3	NA

[a] Percent sequencing reads mapped against the libraries. Represents library and each species matched against this library; NA indicates perfect matching between library which is invalid as the sample inside the library is same as the matching samples.

Table 5
Quantitation of protein mixture in percentage.

	Cod	Tilapia	Platyfish	Haddock	Salmon	Pike	Pangasius	Total unique spectra
Expected (%)	50	33	17	0	0	0	0	–
Total spectra	23,748	24,632	19,698	21,358	24,722	23,972	25,604	-
Total matches	3328	3051	2503	3516	1521	1587	1735	–
Unique matches	1191	1305	823	1199	152	184	235	5089
Unique match (%)	23	26	16	24	3	4	5	–
Divergence	−27	−7	−1	0	0	0	0	–

[a]Fish mixture spectra hits and percentage re-calculated using unique spectra from SpectraST output. Expected (%) - platyfish 1/6, tilapia 2/6 and cod 3/6, forming a mixed tissue sample or DNA samples Total spectra - represents total spectra in the spectral library; Total matches - matches against the library; Unique matches-unique spectra matches from each fish species; Unique match (%) - percent values calculated based on SML matching; Divergence-represents divergence from the expected percentages values in the mixture (% expected - % unique match).

peptides are present in the muscles, i.e., similar tandem mass spectra will be recorded, which will affect spectral library matching. Possibly due to the conserved nature of proteins decreasing species specificity, the accuracy of DNA approach was higher for the closely related species. Thus, the results indicate that the SLM approach displayed higher accuracy than the DNA approach for 1 out of 3 cases at species level and 2 out of 3 cases at taxonomic family level.

The SLM approached used in the present study is independent of annotated genomes and simple to implement. It has been used successfully in earlier studies for accurate identifications of fish species in both raw and processed samples (Nessen et al., 2016; Wulff et al., 2013). Even battered and deep-fried fish were correctly identified using SLM and spectral hits were proportional to the amount of cod (10%) added to the sample (Nessen et al., 2016). SLM also has been applied for quantification of horse in cow meat mixture with reasonable accuracy; it was highlighted that method precision can be improved by removing non-peptide spectra from the spectral reference libraries (Ohana et al., 2016). The method was also applied recently to detect presence of bovine haemoglobin (1–10%) in the black solider fly (BSF) larvae fed on contaminated substrates with accuracy (Belghit et al., 2021). In the present study, it is shown for the first time that SLM also can be applied to more complex mixtures. Moreover, it was found that no masking of MS data is necessary, since masked and raw MS data yielded comparable quantification predictions, both very close to nominal values.

In terms of specificity and false-positive signals, SLM had a clear disadvantage compared to the DNA approach predicting 23% cod and 24% haddock in the fish mixture (Table 5). In addition, 3–5% spectra were matched to other species absent in the fish mixture. By comparison, mapping shotgun DNA sequence reads against masked and normalized reference libraries resulted in less than 3% hits against haddock (Table 2), and negligible hits against other species that were not included in the fish mixture. Similar results have reported by shotgun sequencing approaches demonstrating the discriminating power of shotgun DNA sequencing (Haiminen et al., 2019; Hellmann et al., 2020; Kobus et al., 2020; Ripp et al., 2014; Voorhuijzen-Harink et al., 2019).

Results from the present study highlighted the challenges arisen when analysing closely related species within the same family. The DNA analysis shows almost 50% overlap between the cod and haddock DNA read libraries. Similar results were obtained for the proteomic analysis with a spectral overlap between cod and haddock of ~27% (Table 4). Much less overlap was observed between the other species such as platy and tilapia as these species are distantly related and belong to different superorder i.e., *Protacanthopterygii*, *Ostariophysi*, and *Osteichthyes*.

The results confirm that distantly related species can be easily separated and quantified from the fish mixtures using SLM (Nessen et al., 2016; Ohana et al., 2016). It also was found that it is challenging to quantify the percentage inclusion of very closely related species in fish-mixtures, most probably due to the large degree of similarity in amino acid sequences of the respective peptides. If well-annotated reference proteomes were available for fish, further work could be done to target the analysis of very closely related species using a set of highly distinctive mass spectra representative of species-specific peptides. However, at the time of writing, only scaffold reference proteomes are available for download from online repositories (Supplementary Table 8). Once comprehensively annotated reference proteomes from more species become available, spectra identification using specific peptides will be attempted for accurate separation and abundance estimates as was recently proposed for PAP (Marbaix et al., 2016; Rasinger et al., 2016).

Only non-processed frozen material was used in the present study. Future studies applying the present analytical pipelines should investigate the effect of processing on the analytical outcome. However previous studies predicting the content of processed materials using shotgun DNA sequencing and proteomics indicate that processing, such as cooking (heat treatment), does not affect the predictive result (Haiminen et al., 2019; Kobus et al., 2020; Nessen et al., 2016; Ohana et al., 2016; Ripp et al., 2014). In comparison, metabarcoding approaches using large amplicon can be sensitive to DNA degradation following heat treatment (Hird et al., 2006). Therefore, the presented approaches should also be suitable for cooked fish samples even if they contain other ingredients such as flour or oil.

5. Conclusions

Food and feed scandals are breaching food safety legislation and violating consumer rights which have economic impacts (Moyer et al., 2017). Thus, efficient tools for fraud detection are needed. In the present study, for the first time, shotgun DNA sequencing and mass spectrometry-based proteomics were applied in parallel on the same samples to estimate the relative abundance of fish species in mixed samples. Both approaches show promise for use in future food control applications for species identification and authentication of mixed samples. While the untargeted SLM-based proteomics workflow showed some limitations in differentiating closely related species in comparison to shotgun DNA sequencing in combination with masked reference libraries, the data indicate that at the taxonomic family level, SLM displays a higher accuracy in predicting relative abundances of fish in mixtures. In practice, possibly a tiered approach taking advantage of the specificity of DNA sequencing and the abundance accuracy of proteomics would be best suited for tackling current food authentication challenges.

CRediT authorship contribution statement

Madhushri S. Varunjikar: Formal analysis, Data curation, Sample preparation, Investigation, Bioinformatics pipeline, Writing – original draft, Writing – review & original draft. **Carlos Moreno-Ibarguen:** Formal analysis, Data curation, Investigation, Bioinformatics pipeline, Software, Writing – original draft, Writing – review & original draft. **Juan S. Andrade-Martinez:** Formal analysis, Data curation, Investigation, Bioinformatics pipeline, Software, Writing – review & original draft. **Hui-Shan Tung:** Sample preparation, Writing – review & original draft. **Ikram Belghit:** Sample preparation, Writing – review & original draft. **Magnus Palmblad:**: Bioinformatics pipeline, Software, Writing –

review & original draft. **Pål A. Olsvik:** Conceptualization, Writing – review & original draft, Project administration, Writing – original draft. **Alejandro Reyes:** Formal analysis, Data curation, Bioinformatics pipeline, Software, Writing – original draft, Writing – review & original draft. **Josef D. Rasinger:** Investigation, Project administration, Writing – original draft, Writing – review & original draft. **Kai K. Lie:** Conceptualization, Data curation, Investigation, Project administration, Writing – original draft, Writing – review & original draft.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

This study was supported by the Nærings-og fiskeridepartementet, IMR.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2021.108417.

References

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology, 19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Belghit, I., Lock, E. J., Fumière, O., Lecrenier, M. C., Renard, P., Dieu, M., Berntssen, M. H. G., Palmblad, M., & Rasinger, J. D. (2019). Species-specific discrimination of insect meals for aquafeeds by direct comparison of tandem mass spectra. *Animals, 9*(5). https://doi.org/10.3390/ani9050222

Belghit, I., Varunjikar, M., Lecrenier, M.-C., Steinhilber, A. E., Niedzwiecka, A., Wang, Y. V., Dieu, M., Azzollini, D., Lie, K., Lock, E.-J., Berntssen, M. H. G., Renard, P., Zagon, J., Fumière, O., van Loon, J. J. A., Larsen, T., Poetz, O., Braeuning, A., Palmblad, M., & Rasinger, J. D. (2021). Future feed control – tracing banned bovine material in insect meal. *Food Control,* 108183. https://doi.org/10.1016/j.foodcont.2021.108183

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*(Issue 1). https://doi.org/10.1371/journal.pone.0211128

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bouzembrak, Y., Steen, B., Neslo, R., Linge, J., Mojtahed, V., & Marvin, H. J. P. (2018). Development of food fraud media monitoring system based on text mining. *Food Control*. https://doi.org/10.1016/j.foodcont.2018.06.003

Bruno, A., Sandionigi, A., Agostinetto, G., Bernabovi, L., Frigerio, J., Casiraghi, M., & Labra, M. (2019). Food tracking perspective: Dna metabarcoding to identify plant composition in complex and processed food products. *Genes, 10*(3). https://doi.org/10.3390/genes10030248

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – accurate paired shotgun read merging via overlap. *PloS One, 12*(10), Article e0185056. https://doi.org/10.1371/journal.pone.0185056

Carrera, M., Cañas, B., & Gallardo, J. M. (2013a). Fish authentication. *Proteomics in Foods*, (November), 205–222. https://doi.org/10.1007/978-1-4614-5626-1_12. Springer US.

Carrera, M., Cañas, B., & Gallardo, J. M. (2013b). Proteomics for the assessment of quality and safety of fishery products. *Food Research International, 54*(1), 972–979. https://doi.org/10.1016/j.foodres.2012.10.027

Deutsch, E. W., Mendoza, L., Shteynberg, D., Slagel, J., Sun, Z., & Moritz, R. L. (2015). Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics - Clinical Applications, 9*(7–8), 745–754. https://doi.org/10.1002/prca.201400164

Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: An open-source MS/MS sequence database search tool. *Proteomics, 13*(1), 22–24. https://doi.org/10.1002/pmic.201200439

Haiminen, N., Edlund, S., Chambliss, D., Kunitomi, M., Weimer, B. C., Ganesan, B., Baker, R., Markwell, P., Davis, M., Huang, B. C., Kong, N., Prill, R. J., Marlowe, C. H., Quintanar, A., Pierre, S., Dubois, G., Kaufman, J. H., Parida, L., & Beck, K. L. (2019). Food authentication from shotgun sequencing reads with an application on high protein powders. *Npj Science of Food, 3*(1), 1–11. https://doi.org/10.1038/s41538-019-0056-6

Hellberg, R. S., Hernandez, B. C., & Hernandez, E. L. (2017). Identification of meat and poultry species in food products using DNA barcoding. *Food Control, 80*, 23–28. https://doi.org/10.1016/j.foodcont.2017.04.025

Hellmann, S. L., Ripp, F., Bikar, S. E., Schmidt, B., Köppel, R., & Hankeln, T. (2020). Identification and quantification of meat product ingredients by whole-genome metagenomics (All-Food-Seq). *European Food Research and Technology, 246*(1), 193–200. https://doi.org/10.1007/s00217-019-03404-y

Hird, H., Chisholm, J., Sanchez, A., Hernandez, M., Goodier, R., Schneede, K., Boltz, C., & Popping, B. (2006). Effect of heat and pressure processing on DNA fragmentation and implications for the detection of meat using a real-time polymerase chain reaction. *Food Additives & Contaminants, 23*(7), 645–650. https://doi.org/10.1080/02652030600603041

Huo, W., Ling, W., Wang, Z., Li, Y., Zhou, M., Ren, M., Li, X., Li, J., Xia, Z., Liu, X., & Huang, X. (2021). Miniaturized DNA sequencers for personal use: Unreachable dreams or achievable goals. *Frontiers in Nanotechnology, 3*(February), 1–17. https://doi.org/10.3389/fnano.2021.628861

Ivanova, N. V., Zemlak, T. S., Hanner, R. H., & Hebert, P. D. N. (2007). Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes, 7*(4), 544–548. https://doi.org/10.1111/j.1471-8286.2007.01748.x

Kahlke, T., & Ralph, P. J. (2019). Basta – taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods in Ecology and Evolution, 10*(1), 100–103. https://doi.org/10.1111/2041-210X.13095

Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry, 74*(20), 5383–5392. https://doi.org/10.1021/ac025747h

Kessner, D., Chambers, M., Burke, R., Agus, D., & Mallick, P. (2008). ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics, 24*(21), 2534–2536. https://doi.org/10.1093/bioinformatics/btn323

Khaksar, R., Carlson, T., Schaffner, D. W., Ghorashi, M., Best, D., Jandhyala, S., Traverso, J., & Amini, S. (2015). Unmasking seafood mislabeling in U.S. Markets: DNA barcoding as a unique technology for food authentication and quality control. *Food Control*. https://doi.org/10.1016/j.foodcont.2015.03.007

Kim, M. J., Suh, S. M., Kim, S. Y., Qin, P., Kim, H. R., & Kim, H. Y. (2020). Development of a real-time PCR assay for the detection of donkey (Equus asinus) meat in meat mixtures treated under different processing conditions. *Foods, 9*(2). https://doi.org/10.3390/foods9020130

Kobus, R., Abuín, J. M., Müller, A., Hellmann, S. L., Pichel, J. C., Pena, T. F., Hildebrandt, A., Hankeln, T., & Schmidt, B. (2020). A big data approach to metagenomics for all-food-sequencing. *BMC Bioinformatics, 21*(1), 1–15. https://doi.org/10.1186/s12859-020-3429-6

Lam, H. (2011). Building and searching tandem mass spectral libraries for peptide identification. *Molecular & Cellular Proteomics, 10*(12), 1–10. https://doi.org/10.1074/mcp.R111.008565

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lecrenier, Caroline, M., Marien, A., Veys, P., Belghit, I., Dieu, M., Gillard, N., Henrottin, J., Herfurth, U. M., Marchis, D., Morello, S., Oveland, E., Poetz, O., Rasinger, J. D., Steinhilber, A., Baeten, V., Berben, G., & Fumière, O. (2021). Interlaboratory study on the detection of bovine processed animal protein in feed by LC-MS/MS-based proteomics. *Food Control, 125*, 1–7. https://doi.org/10.1016/j.foodcont.2021.107944. . (Accessed November 2020)

Lecrenier, M. C., Marbaix, H., Dieu, M., Veys, P., Saegerman, C., Raes, M., & Baeten, V. (2016). Identification of specific bovine blood biomarkers with a non-targeted approach using HPLC ESI tandem mass spectrometry. *Food Chemistry, 213*(1774), 417–424. https://doi.org/10.1016/j.foodchem.2016.06.113

Leonard, S. R., Mammel, M. K., Lacher, D. W., & Elkins, C. A. (2015). Application of metagenomic sequencing to food safety: Detection of shiga toxin-producing Escherichia coli on fresh bagged spinach. *Applied and Environmental Microbiology, 81*(23), 8183–8191. https://doi.org/10.1128/AEM.02601-15

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., & Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature, 533*(7602), 200–205. https://doi.org/10.1038/nature17164

Lo, Y. T., & Shaw, P. C. (2018). DNA-based techniques for authentication of processed food and food supplements. *Food Chemistry, 240*(August 2017), 767–774. https://doi.org/10.1016/j.foodchem.2017.08.022

Marbaix, H., Budinger, D., Dieu, M., Fumière, O., Gillard, N., Delahaut, P., Mauro, S., & Raes, M. (2016). Identification of proteins and peptide biomarkers for detecting banned processed animal proteins (PAPs) in meat and bone meal by mass spectrometry. *Journal of Agricultural and Food Chemistry, 64*(11), 2405–2414 (n.d.). MassIVE., ftp://MSV000087017@massive.ucsd.edu https://doi.org/10.1021/acs.jafc.6b00064massive.ucsd.edu/ProteoSAFe/.

Moyer, D. C., DeVries, J. W., & Spink, J. (2017). The economics of a food fraud incident – case studies and examples including Melamine in Wheat Gluten. *Food Control, 71*, 358–364. https://doi.org/10.1016/j.foodcont.2016.07.015

Nagata. (2011). Electron microscopic radioautographic study on the protein synthesis in the pancreas of aging mice with special reference to mitochondria. *Gastroenterology Research, 4*(3), 114–121. https://doi.org/10.4021/gr310e

Nessen, M. A., van der Zwaan, D. J., Grevers, S., Dalebout, H., Staats, M., Kok, E., & Palmblad, M. (2016). Authentication of closely related fish and derived fish products using tandem mass spectrometry and spectral library matching. *Journal of Agricultural and Food Chemistry, 64*(18), 3669–3677. https://doi.org/10.1021/acs.jafc.5b05322

Ohana, D., Dalebout, H., Marissen, R. J., Wulff, T., Bergquist, J., Deelder, A. M., & Palmblad, M. (2016). Identification of meat products by shotgun spectral matching. *Food Chemistry, 203*, 28–34. https://doi.org/10.1016/j.foodchem.2016.01.138

Olsvik, P. A., Fumière, O., Margry, R. J. C. F., Berben, G., Larsen, N., Alm, M., & Berntssen, M. H. G. (2017). Multi-laboratory evaluation of a PCR method for detection of ruminant DNA in commercial processed animal proteins. *Food Control, 73*, 140–146. https://doi.org/10.1016/j.foodcont.2016.07.041

Palmblad, M., & Deelder, A. M. (2012). Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Communications in Mass Spectrometry, 26*(7), 728–732. https://doi.org/10.1002/rcm.6162

Preuten, T., Cincu, E., Fuchs, J., Zoschke, R., Liere, K., & Börner, T. (2010). Fewer genes than organelles: Extremely low and variable gene copy numbers in mitochondria of somatic plant cells. *The Plant Journal, 64*(6), 948–959. https://doi.org/10.1111/j.1365-313X.2010.04389.x

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

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. *Journal of Proteomics, 147*, 125–131. https://doi.org/10.1016/j.jprot.2016.05.036

EGULATION (EU) No 1169/2011, (testimony of REGULATION (EU) No 1169/2011). https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:304:0018:0063:en:PDF.

Ripp, F., Krombholz, C. F., Liu, Y., Weber, M., Schäfer, A., Schmidt, B., Köppel, R., & Hankeln, T. (2014). All-food-seq (AFS): A quantifiable screen for species in biological samples by deep DNA sequencing. *BMC Genomics, 15*(1). https://doi.org/10.1186/1471-2164-15-639

Robin, E. D., & Wong, R. (1988). Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. *Journal of Cellular Physiology, 136*(3), 507–513. https://doi.org/10.1002/jcp.1041360316

Sajali, N., Wong, S. C., Abu Bakar, S., Khairil Mokhtar, N. F., Manaf, Y. N., Yuswan, M. H., & Mohd Desa, M. N. (2020). Analytical approaches of meat authentication in food. *International Journal of Food Science and Technology*, 1–9. https://doi.org/10.1111/ijfs.14797

Sarmashghi, S. (2019). *Skmer : Assembly-free and alignment-free sample identification using genome skims* (pp. 1–20).

Sawyer, J., Wood, C., Shanahan, D., Gout, S., & McDowell, D. (2003). Real-time PCR for quantitative meat species testing. *Food Control, 14*(8), 579–583. https://doi.org/10.1016/S0956-7135(02)00148-2

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). *BUSCO: Assessing genome assembly and annotation completeness*. https://doi.org/10.1007/978-1-4939-9173-0_14

Shokralla, S., Hellberg, R. S., Handy, S. M., King, I., & Hajibabaei, M. (2015). A DNA mini-barcoding system for authentication of processed fish products. *Scientific Reports, 5*, 1–11. https://doi.org/10.1038/srep15894

Steinhilber, A. E., Schmidt, F. F., Naboulsi, W., Planatscher, H., Niedzwiecka, A., Zagon, J., Braeuning, A., Lampen, A., Joos, T. O., & Poetz, O. (2018). Species differentiation and quantification of processed animal proteins and blood products in fish feed using an 8-plex mass spectrometry-based immunoassay. *Journal of Agricultural and Food Chemistry, 66*(39), 10327–10335. https://doi.org/10.1021/acs.jafc.8b03934

Voorhuijzen-Harink, M. M., Hagelaar, R., van Dijk, J. P., Prins, T. W., Kok, E. J., & Staats, M. (2019). Toward on-site food authentication using nanopore sequencing. *Food Chemistry, X*(June), 100035. https://doi.org/10.1016/j.fochx.2019.100035. *2*.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1462), 1847–1857. https://doi.org/10.1098/rstb.2005.1716

Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin, T., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., & Woo, K. (2019). *Welcome to the Tidyverse, 4*, 1–6. https://doi.org/10.21105/joss.01686

Wiśniewski, J. R. (2016). Quantitative evaluation of filter aided sample preparation (FASP) and multienzyme digestion FASP protocols. *Analytical Chemistry, 88*(10), 5438–5443. https://doi.org/10.1021/acs.analchem.6b00859

Wulff, T., Nielsen, M. E., Deelder, A. M., Jessen, F., & Palmblad, M. (2013). Authentication of fish products by large-scale comparison of tandem mass spectra. *Journal of Proteome Research, 12*(11), 5253–5259. https://doi.org/10.1021/pr4006525

Xing, R. R., Wang, N., Hu, R. R., Zhang, J. K., Han, J. X., & Chen, Y. (2019). Application of next generation sequencing for species identification in meat and poultry products: A DNA metabarcoding approach. *Food Control, 101*(February), 173–179. https://doi.org/10.1016/j.foodcont.2019.02.034

Yancy, H. F., Zemlak, T. S., Mason, J. A., Washington, J. D., Tenge, B. J., Nguyen, N. L. T., Barnett, J. D., Savary, W. E., Hill, W. E., Moore, M. M., Fry, F. S., Randolph, S. C., Rogers, P. L., & Hebert, P. D. N. (2008). Potential use of DNA barcodes in regulatory science: Applications of the regulatory fish encyclopedia. *Journal of Food Protection, 71*(1), 210–217. https://doi.org/10.4315/0362-028X-71.1.210

Yang, F., Ding, F., Chen, H., He, M., Zhu, S., Ma, X., Jiang, L., & Li, H. (2018). DNA barcoding for the identification and authentication of animal species in traditional medicine. *Evidence-based Complementary and Alternative Medicine*. https://doi.org/10.1155/2018/5160254. *2018*.