Proteomic Tools for Food and Feed Authentication

Madhushri Shrikant Varunjikar

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2023

UNIVERSITY OF BERGEN

Proteomic Tools for Food and Feed Authentication

Madhushri Shrikant Varunjikar

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 10.03.2023

Year: 2023

Title: Proteomic Tools for Food and Feed Authentication

Name: Madhushri Shrikant Varunjikar

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

This PhD was completed at the Institute of the Marine Research (IMR) in Bergen, Norway. During the PhD, I have been part of Marine toxicology research group and was employed to work on the Multi-Omics Tool project funded by the Institute of Marine Research.

This PhD was supervised by Dr. Josef D. Rasinger (principal supervisor), co-supervised by Dr. Kai Kristoffer Lie and Prof. Anne-Katrine Lundebye at the Institute of Marine Research. Dr. Josef D. Rasinger managed the Multi-OmicsTool project in close collaboration with the project group including Dr. Kai K. Lie, Dr. Ikram Belghit, and Dr. Eystein Oveland. The work was performed in the quality-assured proteomics and molecular biology laboratory facilities of the Institute of Marine Research. During this PhD, bioinformatics and technical support were provided by Prof. Magnus Palmblad from the Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands.

Acknowledgements

As per the saying of the great Buddha, *"It is better to travel well than to arrive"* and I realized this while doing this PhD. I feel like I have enjoyed this scenic journey of my PhD due to the excellent work environment at the Institute of Marine Research. Though the Covid-19 pandemic hit as soon as I started my PhD, I was able to cope with new changes due to the support of my supervisors and colleagues. 2020 was very challenging due to limited traveling opportunities and most activities being online, but everything worked out in the end.

I am very much thankful to the Institute of Marine Research, for funding the Multi-Omics Tool project and this PhD. I want to acknowledge the support from the Institute of Marine Research and the University of Bergen throughout this PhD.

I am grateful to the people who guided me throughout this PhD. Firstly, I would like to thank my principal supervisor, Dr. Josef D. Rasinger, for supporting me throughout this PhD. Thanks to his patience, guidance, teaching, and excellent supervision skills, I finished this PhD promptly. Also, I am very much grateful for all the collaboration opportunities that I got through him. I would like to thank Prof. Anne-Katrine, and Dr. Kai K. Lie for their supervision, guidance, fruitful discussions, and motivation. I am very much thankful to Prof. Robin Ørnsrud, Marine toxicology group's leader, for providing all the support I needed during this PhD.

Secondly, I would like to thank Dr. Belghit Ikram for training me in the proteomics laboratory and assisting in developing an in-house spectrometry method. She has been a dependable colleague and great collaborator throughout this PhD. I am also thankful to Dr. Eystein Oveland for his support while developing the proteomic method in-house as a part of this PhD. I am very much grateful to Prof. Magnus Palmblad from LUMC for his bioinformatics support and guidance during this PhD.

Being part of the Marine toxicology group at the IMR has been a pleasure. I was lucky to have co-workers and friends who were understanding and supported me. My officemate Charlotte has been part of this journey; thank you, Charlotte, for listening

to my frustrations and complaints. I am grateful to Sofie for all her valuable advice, tips, tricks, and support. My co-workers and friends, Angela, Annette, Gopika, Kjersti, Jojo, Cathrine, Marta, Sahar, Amalie, and Chavindi, always looked out for me. Thank you all, and it has been great to be with you.

I am thankful to my mother and father for always believing in me. My dear little sister Aboli have been always supporting and motivating me in hard times. My husband and best friend, Adwaith Nath, motivated me to take up this PhD during a difficult time when I was losing hope; he helped me to regain my motivation. Thank you, Adwaith and Aboli for your trust and support. I am also thankful to my mother and father-in-law for their love and encouragement.

-Madhushri S. Varunjikar-

Content

Abstract in English

Due to globally rising demands for food and feed, novel proteinaceous ingredients are introduced into our food systems on an increasing scale. The introduction of novel ingredients and circularity of the food system gives rise to novel challenges concerning the detection of feed and food fraud and the determination of feed and food authenticity, respectively. In this context, developing and increasing the implementation of rapid, sensitive, and robust molecular methods are essential. In the past, progress in applying such tools has been hampered by a general lack of well-annotated reference genomes of target species commonly used or newly introduced in feed or food preparations. This PhD focused on developing and implementing mass spectrometry-based approaches to identify, differentiate, and quantify proteinaceous ingredients of animal and plant origin in various food and feed mixes without using any genomic information.

The work presented in this PhD implemented bottom-up proteomic workflows using high-performance liquid chromatography (HPLC) tandem mass spectrometry (MS/MS). Data analyses were done using direct spectra comparison (compareMS2), spectra library matching (SLM), Trans-Proteomics Pipeline (TPP), and MaxQuant software. All data generated and published during this PhD have been made available on public repositories for proteomics data, such as the Mass Spectrometry Interactive Virtual Environment (MassIVE), following Findable, Accessible, Interoperable, and Reusable (FAIR) principles.

The untargeted proteomics SLM workflow implemented during this PhD successfully differentiated processed animal proteins such as bovine milk and bovine blood. The SLM was also used to identify and authenticate food and feed-grade insect species and to detect if black soldier fly (BSF) larvae were fed on the prohibited PAP. Using the SLM workflow, it was also possible to quantify and authenticate the different species in fish mixtures containing muscle tissues from three different fish species. It was also shown that untargeted proteomics could be used to identify common allergens in food-grade insect samples. Also, the proteomic approach was successfully implemented to

separate thirty-one ready-to-market soybean samples farmed organically, conventionally, and with genetic modifications (GM). Differential protein expression was detected between GM, conventionally, and organically farmed soybean samples. Additional bioinformatics analyses led to the detection of two novel peptide markers for the efficient tracing of GM crops in food and feed.

The proteomic tools implemented during this PhD were capable of species and tissues specific identification of proteinaceous food and feed ingredients, including processed animal proteins, plant, mammalian, and fish proteins. Future work should focus on the differentiation and detection of fraud in food and feed in the global food market. Web-based interphase will be developed for food and feed authentication using spectra libraries created during this PhD. Following proper quality testing, the web-based interphase will be released publicly to provide research and regulatory laboratories with an easily accessible platform for authenticating and identifying protein ingredients in feed and food samples.

Abstract in Norwegian

På grunn av globalt økende etterspørsel etter mat og fôr, introduseres nye proteinholdige ingredienser i matsystemene våre i økende skala. Innføring av nye ingredienser og introduksjon av sirkulære matsystemer gir nye utfordringer når det gjelder metoder for avsløring av henholdsvis fôr- og matsvindel. I denne sammenhengen er det viktig å utvikle raske, sensitive og robuste molekylære metoder som kan implementeres i kontroll og overvåkningsøyemed. Tidligere har fremskritt ved bruk av slike verktøy blitt hemmet av en generell mangel på annoterte referansegenomer for målarter som ofte brukes, eller nylig er introdusert, i fôr eller matpreparater. Fokuset for denne doktorgraden er å utvikle og implementere massespektrometriske metoder (LC-MS/MS) som er i stand til å identifisere, differensiere og kvantifisere proteinholdige ingredienser av animalsk og planteopprinnelse i ulike mat- og fôrblandinger ved bruk av massespektra fingeravtrykk.

Arbeidet som presenteres i denne doktorgraden omfatter «bottom-up» proteomiske arbeidsflyter ved bruk av høytrykksvæskekromatografi (HPLC) tandem massespektrometri (MS/MS). Databehandling ble utført ved å bruke direkte spektrasammenligning (compareMS2) og spektrabibliotekmatching (SLM) analyser ved bruk av verktøy fra Trans-Proteomics Pipeline (TPP) og annen åpen kildekode til bioinformatisk programvare. Alle data generert og publisert i løpet av denne doktorgraden har blitt gjort tilgjengelig på offentlige repositrium for MS-data, for eksempel Mass Spectrometry Interactive Virtual Environment (MassIVE), som følger FAIR-prinsippene.

Den SLM baserte arbeidsflyten brukt i denne doktorgraden klarte å differensiere ulike prosesserte animalske proteiner (PAP) som storfemelk og bovint blod. SLM ble også brukt til å differensiere ulike insektarter og for å detektere om larver av svart soldatflue (BSF) var fôret med PAP. SLM-metoden ble også brukt til å identifisere og kvantifisere innholdet i et blandingsprodukt av 3 ulike fiskearter. Det ble også funnet at SLM basert proteomikk kan brukes til å identifisere vanlige allergener i insektsprøver tiltenkt

humant konsum. Denne tilnærmingen ble også implementert med suksess for å differensiere mellom soyabønneprøver som var enten dyrket organisk, konvensjonelt eller inneholdt genetiske modifikasjoner (GM). I tillegg ble differensiell proteinekspresjon påvist mellom prøver av GM, konvensjonelt og økologisk dyrkede soyabønner. Dette førte til identifisering av to nye peptidmarkører for effektiv sporing av GM-avlinger i mat og fôr.

Denne doktorgraden har vist at den SLM baserte metoden er i stand til å identifisere både art og vevstype brukt i et proteinholdig matprodukt eller fôringredients det være seg PAP, plante-, pattedyr- eller fiskeproteiner. Fremtidig arbeid bør fokusere på differensiering og avsløring av svindel i sjømat, som nylig ble fremhevet som et fremvoksende tema i det globale matmarkedet. Alle arts- og vevsspesifikke MS-data samlet inn i det ovennevnte arbeidet vil gjøres tilgjengelig fra i dedikert nettbaserte tjenester. Sistnevnte utvikles for tiden internt, og etter skikkelig kvalitetstesting er det tenkt å bli utgitt offentlig for å gi forskningsmiljøer og myndigheter en lett tilgjengelig plattform for autentisering og identifisering av proteinholdige ingredienser i fôr- og mat.

List of Publications

Paper I

Belghit, I., **Varunjikar, M.**, Lecrenier, M.-C. C., Steinhilber, A. E., Niedzwiecka, A., Wang, Y. V. V., Dieu, M., Azzollini, D., Lie, K., Lock, E.-J. J., Berntssen, M. H. G. H. G., Renard, P., Zagon, J., Fumière, O., van Loon, J. J. A. J. A., Larsen, T., Poetz, O., Braeuning, A., Palmblad, M., & Rasinger, J. D. D. (2021). Future feed control – Tracing banned bovine material in insect meal. *Food Control*, *128*(April), 108183. https://doi.org/10.1016/j.foodcont.2021.108183

Paper II

Varunjikar, M.S., Moreno-Ibarguen, C., Andrade-Martinez, J.S., Tung, H.S., Belghit, I., Palmblad, M., Olsvik, P.A., Reyes, A., Rasinger, J.D. and 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*, p.108417. https://doi.org/10.1016/j.foodcont.2021.108417

Paper III:

Varunjikar, M. S., Belghit, I., Gjerde, J., Palmblad, M., Oveland, E., & Rasinger, J. D. (2022). Shotgun proteomics approaches for authentication, biological analyses, and allergen detection in feed and food-grade insect species. *Food Control*, 108888. https://doi.org/10.1016/j.foodcont.2022.108888

Paper IV:

Varunjikar M.S., Bøhn T., Sanden M., Belghit I., Palmblad M., Rasinger J.D. Analysis and differentiation of herbicide tolerant, conventional, and organic soybean seeds using proteomic approach (Submitted)

Paper V:

Marissen, R., **Varunjikar, M. S.**, Laros, J. F., Rasinger, J. D., Neely, B. A., & Palmblad, M. (2022). compareMS2 2.0: An Improved Software for Comparing Tandem Mass Spectrometry Datasets. *Journal of Proteome Research*. https://doi.org/10.1021/acs.jproteome.2c00457

Abbreviations

AF Analytical flow

AF-HPLC HR-MS Analytical flow High-Performance Liquid Chromatography High Resolution-Mass Spectrometry

AGC Automatic Gain Control

BvHb Bovine Hemoglobin powder

BSE Bovine Spongiform Encephalopathies

BSF Black Soldier Fly

CID Collision-Induced Dissociation

COI Cytochrome *c* Oxidase subunit I

DDA Data-Dependent Acquisition

EFSA European Food Safety Authority

EPSPS Enolpyruvylshikimate-3-Phosphate Synthase

EURL-AP European Union Reference Laboratory for Animal Protein

FAIR principles Findable, Accessible, Interoperable, and Reusable principles

FDR False Discovery Rate

GM Genetically Modified

GMO Genetically Modified Organism

HC House Cricket

HPLC-MS/MS High-Performance Liquid Chromatography Tandem Mass Spectrometry

HR-MS High Resolution-Mass Spectrometry

LW Lesser mealworm

MassIVE Mass Spectrometry Interactive Virtual Environment

MBM Meat Bone Meal

MS Mass Spectrometry

MGF Mascot Generic Format

MF Microflow

mzML open standard data format for mass spectrometry data

mzXML mass to charge ratio in eXtensible Markup Language

MF-HPLC QTOF Microflow High-Performance Liquid Chromatography Quadrupole time-of-flight

MRM multiple reaction monitoring

MW Morio worms

PCR Polymerase chain reaction

PSM Peptide-Spectrum Match

PAP Processed Animal Proteins

SL Spectra Library

SLM Spectra Library Matching

SRM Selected Reaction Monitoring

TPP Trans-Proteomic Pipeline

TSE Transmissible Spongiform Encephalopathies

qPCR quantitative Polymerase Chain Reaction

QTOF Quadrupole time-of-flight

YW Yellow meal Worm

List of Figures and Tables

Figures

Tables

1. Introduction

In the past decade, high throughput molecular technology or omics approaches have been extensively used as advanced analytical methods to address biological research questions (Capozzi & Bordoni, 2013; Ellis et al., 2016). Clinical assays, *in vivo*, and *in vitro* studies apply omics approaches frequently, and also, in food sciences and nutrition, omics and bioinformatics have been increasingly implemented (Ellis et al., 2016; Herrero et al., 2012). Genomics has been advancing swiftly; for many non-model species, data are available from sequencing projects such as Earth BioGenome Project, Vertebrate Genomes Project, DNA Zoo, Zoonomia Project, Bat1K Project, or Bird 10000 Genomes Project (Heck & Neely, 2020). With increased genome availability, the application of proteomics in non-model species has been advancing, which will benefit food safety research and various other areas, including molecular evolution, veterinary medicine, biomedical advancement, and agricultural research (Heck & Neely, 2020; Neely & Palmblad, 2021). Mass spectrometry-based proteomics has been widely used to analyze complex protein samples (Aebersold & Mann, 2003). In shotgun proteomics approaches, protein samples are digested using proteolytic enzymes such as trypsin. The resulting peptides are separated using liquid chromatography, and data are collected using tandem mass spectrometry. Tandem mass spectra are matched to a relevant proteomic database using bioinformatic search engines to interpret proteins (Duivesteijn, 2018). Proteomic methods have previously been used in food safety studies (Belghit et al., 2021; Lecrenier et al., 2016, 2018; Nessen et al., 2016; Rasinger et al., 2016). This PhD thesis aimed to implement shotgun proteomic and bioinformatic tools on non-model organisms for advancing the application of omics in food and feed safety research. The following sections briefly introduce shotgun proteomic methods and food and feed safety challenges. As this PhD thesis focuses on implementing proteomic methods, shotgun proteomics is introduced first, and food and feed safety challenges are discussed in later sections.

1.1 Shotgun Proteomics and Bioinformatics

Proteins are building blocks of cellular structures in an organism; studying proteins in an organism is called proteomics. Proteomics gives information regarding proteins present in a tissue of an organism at a given time point and provides a picture of the cellular, biological, and molecular processes. Protein profiles of tissues or cells can be identified and quantified with a shotgun proteomic approach (Duivesteijn, 2018). Proteins are extracted from any given sample and digested with enzymes such as trypsin which cleaves the protein backbone at the C-terminal to cite the end of arginine and lysine (Olsen et al., 2004). After the tryptic digestion of samples, separation is performed using reversed-phase high-performance liquid chromatography (HPLC) (Duivesteijn, 2018). The peptides eluting from the HPLC are analyzed using mass analyzers such as an Orbitrap using data-dependent acquisition (DDA) mode (Kalli et al., 2014). Mass analyzers measure the mass-to-charge (*m/z*) ratio of a molecule, which helps to detect amino acid sequences in the peptides (Figure 1). There are two scans in mass spectrometry (MS); the first MS1 records the *m/z* ratio and precursor ion intensities from the eluted peptides and selects the most intense ions for fragmentation and into smaller fractions using collision cells (Aebersold & Mann, 2003; Duivesteijn, 2018). The fragmentation technique used in mass spectrometry, such as an Orbitrap, collision-induced dissociation (CID), generates MS2 spectra. The MS2 spectra (tandem mass spectra) generated by the instruments are then used to identify peptides using proteomic database search or spectra library matching pipelines (Figure 1). In the proteomic database search, acquired tandem mass spectra are searched against theoretical spectra derived from a protein database using search engines (Duivesteijn, 2018). Various open-source software such as MaxQaunt, Trans-proteomic pipeline, and OpenMS can perform database searches using search engines such as Andromeda, Comet, X! Tandem, MS-GF plus, Mascot, and SEQUEST (Duivesteijn, 2018).

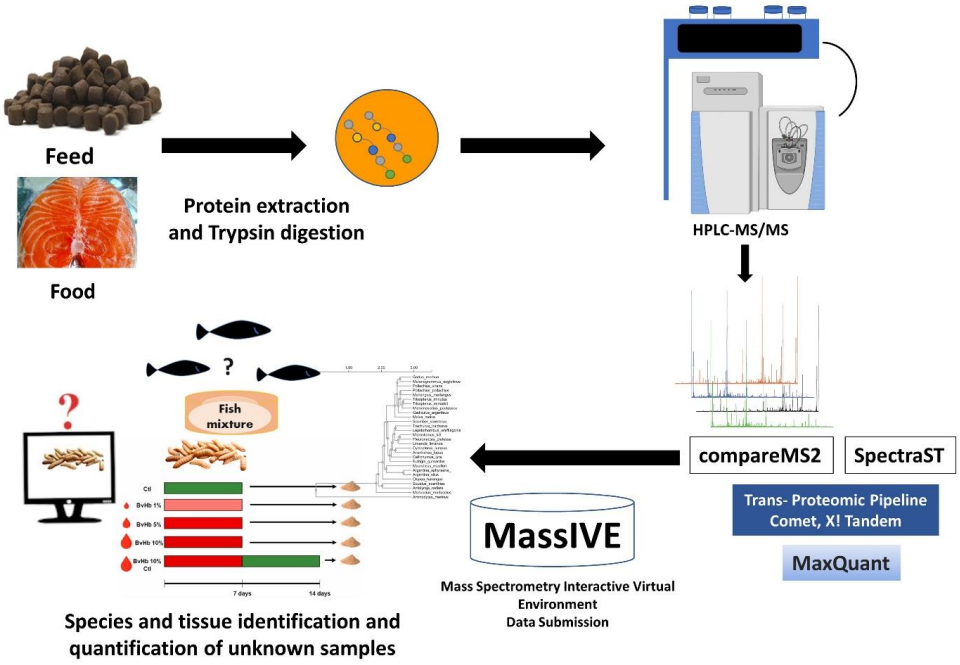


Figure 1: Shotgun proteomic workflow implemented during this PhD for proteomic experiments. Protein samples (food and feed) were processed and digested with trypsin enzyme. The digested samples were then separated using High-Performance Liquid Chromatography Tandem Mass Spectrometry (HPLC-MS/MS) to acquire proteomic data. Data were used to build spectra libraries, identify proteins, and submit to the Mass Spectrometry Interactive Virtual Environment (MassIVE) repository.

In the database search, the spectra matching search outcomes depend on the database size and search engine used. Protein sequences and proteomes are more available for model organisms (Figure 2). However, for non-model species, protein sequence databases are small, posing challenges to existing workflows (Heck & Neely, 2020). For example, in UniProtKB for non-model organisms, few reviewed and unreviewed protein sequences are present (Figure 2). In such a case, it is possible to identify proteins by homology searching in conventional database search using a proteome of closely related species (Heck & Neely, 2020).

Organisms (Fish)	Scientific name	Reviewed number of UniProt KB sequences	Unreviewed number of UniProt KB sequences
Zebrafish	*Danio rerio*	3601	114570
Atlantic cod	*Gadus morhua*	64	63698
Atlantic haddock	*Melanogrammus aeglefinus*	1	200
Nile tilapia	*Oreochromis niloticus*	22	79640
Northern pike	*Esox lucius*	113	80077
Atlantic salmon	*Salmo salar*	182	89282
Platyfish	*Xiphophorus maculatus*	8	35388
Pangasius	*Pangasianodon hypophthalmus*	0	21535

Organisms (Insects)	Scientific name	Reviewed number of UniProt KB sequences	Unreviewed number of UniProt KB sequences
Fruitfly	*Drosophila melanogaster*	4926	364541
Black soldier fly	*Hermetia illucens*	1	17698
Yellow mealworm	*Tenebrio molitor*	31	19356
Lesser mealworm	*Alphitobius diaperinus*	0	25
House cricket	*Acheta domesticus*	235	466
Morio worm	*Zophobas morio*	0	6

Figure 2: Comparison of reviewed and unreviewed sequences in model organisms and non-model organisms. The red box indicates that more reviewed sequences were present in model organisms from fish and insect compared to non-model species. Fish and insects were selected as an example due to their relevance to this PhD thesis.

With increasing proteomic applications in research, the generated proteomic data should be in line with FAIR principles, i.e., data must be Findable, Accessible, Interoperable, and Reusable (Caufield et al., 2021). A ProteomeXchange (PX) consortium was established to make proteomic data accessible to the scientific community (Keane et al., 2021). Members of the PX consortium are MassIVE, PRIDE, PeptideAtlas/PASSEL, Panorama Public, jPOST, and iProX (Jones et al., 2022). Submitting generated proteomic data to the PX consortium ensures the findability and accessibility of data. According to the proteomics standard initiative, open data formats are vital to ensure interoperability. However, achieving interoperability is difficult due to multiple vendor-designed software. During this PhD, interoperable search results were generated using open-source proteomic analyses software to ensure the reusability of the data. Submission of proteomic datasets to the members of the PX consortium will lead to the reusability of the data for implementing artificial intelligence tools, proteogenomic methods, and discoveries of post-translation modification (Jones et al., 2022). In this PhD, FAIR principles were applied for proteomic data using open-source bioinformatic tools described below and submitting data to the Mass Spectrometry Interactive Virtual Environment (MassIVE) repository.

1.1.1 compareMS2

Molecular phylogenetics is the study of the evolutionary relatedness of organisms by comparing DNA, RNA, or proteins. Due to cost-effectiveness, high-throughput DNA sequencing methods have been used for evolutionary studies. Before DNA sequencing became popular, a pioneering evolutionary study compared chromatographically separated protein patterns to establish phylogenetic relationships (Zuckerkandl et al., 1960). Recently, proteomic analyses of ancient samples have resolved the history of evolution in placental mammals highlighting the potential of tandem mass spectrometry for evolutionary studies (Welker et al., 2015). Using compareMS2, samples can be separated by their phylogenetic relationship by direct comparison of tandem mass spectra (Palmblad & Deelder, 2012). The method has successfully established the molecular phylogeny of mammals, fish, and several other species (Ohana et al., 2016; Rasinger et al., 2016; Wulff et al., 2013). The compareMS2 calculates the similarity between spectra using the compareMS2 cosine score, i.e., the

cosine of the angle from the vectorial representation of the tandem mass spectra after normalizing both spectra (Marissen et al., 2022). The software is openly available as compareMS2 Graphical User Interphase (GUI) for the scientific community (compareMS2 GUI, 2021). In this PhD, compareMS2 was used as a tool for quality checks before building spectra libraries from a dataset. The species studied as a part of this PhD were non-model species with limited genomic and proteomic information, and usage of compareMS2 generated rapid dendrograms for proteome comparisons.

1.1.2 SpectraST Spectra Library

Spectra searching against the sequence database is a preferred method in proteomics, and search results from these searches can be used to build spectra libraries. The spectra library search restricts search space to previously detected peptides, and fragmentation patterns are recorded and compiled into spectra libraries (Lam et al., 2008). Tools such as X! Hunter, Bibliospec, and SpectraST were developed for spectra library matching, and identified spectra were collected in the spectra library. The spectra library search is based on the fragmentation pattern of a molecule under fixed conditions; the unknown spectra acquired under the same conditions can be identified using spectra matching. However, the effectiveness of this matching depends on the quality of reference spectra and matching parameters accommodating imperfect matches.

In this PhD, the SpectraST (Spectra Search Tool) was used for building and matching spectra libraries. The SpectraST tool was developed at the Institute for Systems Biology (ISB) and released with the TPP software. Usually, high probability matches from spectra library search are included in the spectra library. However, with SpectraST 5.0, it is possible to build a spectra library with unidentified spectra, i.e., unidentified spectra archives (Frank et al., 2011). For food and feed-relevant species (non-model species) used in this PhD, a limited number of protein sequences were available in the proteomic database; therefore, all the spectra were not precisely identified using database searches. However, if unidentified spectra are added to the spectra library, the information will not be lost and could be used to identify the sample accurately. The advantage of this approach is search speed and precision, a shortcoming of protein sequence-based search engines. The unknown spectra are

matched against extensive collections of previously observed peptides by finding the best match to the existing spectra in the library, giving a high search speed (Lam, 2011).

After the matching, SpectraST output calculates a dot product of matching, which is the ratio of similarity between two spectra, meaning the higher the dot product higher the similarity. Therefore, the number of tandem mass spectra with a dot product above 0.7 were used to calculate percentages of species or tissue in the sample or identify the spectra (Wulff et al., 2013). A high dot product represents an accurate SpectraST match with a false discovery rate (FDR) of less than 1% (Wulff et al., 2013).

1.1.3 Trans-Proteomic Pipeline (TPP)

The Trans-Proteomic pipeline (TPP) is an open-source toolkit to analyze proteomics tandem mass spectra (Deutsch et al., 2015). The pipeline provides a complete suite for interpreting proteomic data with statistical validation (Deutsch et al., 2008). The TPP workflow aggregates information into peptides and protein tables with probabilities, spectra information, and sequence information. Using TPP, tandem mass spectra can be analyzed by searching against the database using search engines or matching against spectra libraries. The search engines such as X! Tandem, Comet, and Mascot can be employed for database searches by acquiring peptide spectrum matches (PSM) for each spectrum. Integrated software such as PeptideProphet and ProteinProphet can subsequently be used to interpret peptides and proteins from the results (Keller et al., 2002; Nesvizhskii et al., 2003). PeptideProphet uses PSM assignments to calculate the FDR. Using these FDR estimation, ProteinProphet calculates probabilities of the peptide to protein discoveries for protein identification (Nesvizhskii et al., 2003). After analyses, the output of the TPP is a list of proteins that can be used to plot heatmaps and identify the presence or absence of proteins (Duivesteijn, 2018). Similarly, SpectraST can be used via the TPP interphase with limited functionalities for spectra library matching.

1.1.4 MaxQuant

MaxQuant is an integrated proteomics software suite for analyzing quantitative shotgun proteomics data with the integrated search engine Andromeda (Cox & Mann,

2008; Tyanova et al., 2016). Both label-free and label-based quantifications are possible using this software. For peptide and protein identification the Andromeda search engine calculates probabilities and FDRs. MaxQuant eliminates systematic error via recalibration, where a mass error and correction factor are calculated for any given data (Cox & Mann, 2008; Tyanova et al., 2016). In this PhD, the label-free quantification workflow from MaxQuant was used in **Paper IV**. The MaxQuant output table was post-processed using the *proteus* R package, which is freely available for analyses (Gierlinski et al., 2018).

The following sections will introduce food and feed challenges and the need for implementing proteomic methods to ensure food safety.

1.2 Food and Feed Security

The world population will increase to 8.5 billion in 2030 (UN, 2022), ultimately increasing food demands. Increased food demands will pressure already over-exploited natural resources. According to the Food and Agricultural Organization (FAO) guidelines, people must have access to enough nutritious, safe, and affordable food to maintain a healthy and active life (FAO, 2022; McCarthy et al., 2018). The current food supply model depends on finite resources; there is a severe need for a more resource-efficient circular economy for food production.

One-half of the planet's vegetated land is used for food production, and possibly, with a growing population, more land will be used in the future (Mc Carthy et al., 2018; WRI, 2020). Also, poor management of the available agricultural soil is causing erosion, irrigation, and fertility problems (Maggio et al., 2015). Furthermore, water usage in agriculture accounts for around 70% percent of total freshwater consumption by humankind (Maggio et al., 2015). Water and land scarcity will eventually affect food production by reducing agricultural yield. Along with resource scarcity, climate change is yet another challenge for food production. The possible effects of climate change on the agricultural sector include metabolic and growth variation, plant productivity changes, and increased pest infestation (Singh Malhi et al., 2021).

Moreover, along with existing challenges, the COVID-19 pandemic directly or indirectly affected farming practices and food production globally (SDGs, 2022).

Due to the scarcity of land resources, sustainable utilization of marine resources is becoming important (FAO, 2018). Seafood is a part of traditional diets and provides excellent proteins and other nutrients (essential amino acids, minerals, vitamins, and omega-3 fatty acids) to consumers (FAO, 2022; VKM, 2022). Global fish production comprises captured fisheries from oceans, inland waters, and aquaculture. Aquaculture production has grown fast worldwide as a food industry (FAO, 2022). In 2016, aquaculture provided 50% of the production of finfish, mollusk, seaweed, and shrimp (FAO, 2020). Increased demand and production and the resulting exponential growth in the aquaculture sector put increased pressure on ocean and land-based resources as the growing aquaculture industry requires an ever-increasing amount of feed (Aas et al., 2019). For the sustainable expansion of aquaculture, sustainable feed ingredients need to be utilized with minimum carbon footprints and environmental impacts (van der Spiegel et al., 2013).

Previously, aquaculture relied primarily on marine capture fisheries for providing adequate nutrients to growing fish; therefore, majorly criticized for pressurizing already exhausted marine fish stocks (Aas et al., 2019) and directly competing with human consumption of these resources. The usage of fishmeal and fish oil has drastically reduced in the last decade due to the effective formulation and inclusion of plant-based ingredients like soybean proteins and oil (Aas et al., 2019). Therefore, commercial diets are considerably different now in aquaculture compared to a few decades ago (Burr et al., 2012; Aas et al., 2019). Sustainability concerns linked to industrialized farming for feed production include deforestation, water usage, pollution, and increased global greenhouse gas production, ultimately making aquaculture unsustainable (Woodgate et al., 2022; Zortea et al., 2018).

1.2.1 Circular Bio-based Economies and Marine Aquaculture

The transformation from a linear economy to a circular bio-based economy is necessary to produce sustainable aquafeed ingredients. A circular biobased economy is defined

as the sustainable utilization of biological resources and efficient utilization of food, feed, by-products, and bioenergy (Flynn et al., 2019). In circular bio-economies, products are generated by relocating the waste from one industry to supply raw materials to another (Kalmykova et al., 2018), and in this way, waste becomes valuable. Bio-economies focus on minimizing the impact of food production on the environment and producing sustainable, healthy, affordable food (Ghisellini et al., 2016; Mirabella et al., 2014).

In the European aquaculture industry, a shortage of protein-rich materials is considered a significant risk factor for food security (FEFAC Annual Report, 2018). Alternative proteins are introduced or reintroduced in the feed market to reduce the shortage of proteins. Alternative protein sources for feed production includes insect proteins, single-cell proteins, and discarded material from the food industry, such as by-products of farm animal production, known as processed animal proteins (PAP). Sustainable feed ingredients relevant to this PhD are discussed hereafter in detail.

1.2.2 Processed Animal Proteins (PAP)

Animal by-products (ABPs) are "animal products that are no longer intended for human consumption and safe to be used as animal feed" (Campos, 2019; van Raamsdonk et al., 2019). Due to the circularity aspect, such products should be used as feed products for terrestrial farm animal production and aquaculture. After 2002, ABPs were referred to as PAPs to reflect processing conditions. Under high pressure and temperature ABPs are transformed into valuable pet or livestock feed ingredients (Meeker & Hamilton, 2006). The use of PAPs, such as feather meal, poultry by-product meal, pork meat, and bone meal, poultry and pork blood meal, was common in the feed before 2002 (Woodgate et al., 2022). However, due to food safety concerns in 2002, the use of PAPs was prohibited in Europe after an outbreak of bovine spongiform encephalopathy (BSE). BSE is a neurodegenerative disorder in cattle which is a type of transmissible spongiform encephalopathies (TSE) (van Raamsdonk et al., 2019). Prions are causative agents of TSEs (van Raamsdonk et al., 2019). PAP usage in feed was the most probable cause of BSE spread observed in 2001, and hence, PAPs were prohibited from being used as animal feed in Europe (European Commission,

2001/999). A risk assessment performed by the European Food Safety Authority (EFSA) showed that the usage of ruminant proteins could increase the chances of a BSE outbreak. However, the risk of BSE epidemics was considered negligible if no ruminant proteins were present in animal feed (EFSA, 2005).

In 2013, after careful evaluation, non-ruminant PAPs were reauthorized to be used as aquaculture feed material. A quantitative polymerized chain reaction method (qPCR) was developed to detect the presence of ruminant material in feed samples (European Commission, 2013/51; European Commission, 2013/56). According to the modified regulation European Commission, 2013/56 and 2013/51, the non-ruminant PAP could reenter the food chain safely. An EFSA assessment in 2018 showed that the risk of BSE was reduced four times compared to in 2011 (EFSA, 2018a), and the reintroduction of non-ruminant PAPs, i.e., pig and poultry, were approved in the Europe (European Commission, 2021/1372). According to current regulations, pig PAPs are allowed in poultry feed, and poultry PAPs are allowed in pig feed (Woodgate et al., 2022).

The development of analytical techniques has helped legislative guidelines allow the safe reintroduction of non-ruminant PAPs in the food chain. However, as regulations are rapidly changing, new analytical toolkits are required to ensure the safety of PAP products (Belghit et al., 2021). For example, regulations changed from 2020 to 2022; ruminant collagen and gelatin were not authorized to be used in feed for pig, poultry, and aquaculture in 2020, but according to 2022 regulations, collagen and gelatin from ruminants were authorized in feed for pig, poultry, and aquaculture (EURL-AP Standard Operating Procedure, 2022). According to recent regulations, insect PAPs can be used in the aquaculture, pig, and poultry industry (Figure 3, EURL-AP Standard Operating Procedure, 2022).

1.2.3 Insect Proteins

Insects belong to the phylum Arthropoda, and it is estimated that one million species exist in this class (van Raamsdonk et al., 2019). Insects are a rich source of protein, can be produced sustainably, and hence are an ideal ingredient to be included in animal

feed (Lock et al., 2016; Makkar et al., 2014). The capability of insect species to utilize food waste and produce high-quality protein and fat is very appealing for circular bio-based economies (van Huis, 2020). Insect species can valorize food and farm waste to produce high-quality aquafeed (van Huis, 2020). Benefits of rearing insects include a wide range of substrate suitability, smaller space requirements when compared to conventional farm animals, high growth rates, and efficiency in converting low-grade waste into high-quality protein products (Belghit et al., 2019a; Lock et al., 2016; Makkar et al., 2014; van Huis, 2020).

	Feed for farmed Animals				
	Ruminants	Non ruminant herbivores (horses, rabbits...)	Pigs and poultry	Aquaculture	Pets and fur animals
Ruminant PAP (incl. ruminant blood meal)	U	U	U	U	A
Ruminant blood products	U	U	U	U	A
Non-ruminant blood products	U	A	A	A	A
Pig and poultry PAP (incl. pig and poultry blood meals)	U	U	A†	A	A
Non-ruminant PAP other than from pig and poultry	U	U	U	A	A
Non-ruminant blood meal other than from pig and poultry	U	U	U	A	A
Insect PAP	U	U	A	A	A
Fishmeal	U*	A	A	A	A
Ruminant collagen and gelatine	U	A	A	A	A
Non-ruminant collagen and gelatine	A	A	A	A	A
Hydrolysed proteins from ruminants other than those derived from hides and skins	U	U	U	U	A
Hydrolysed proteins from ruminants derived from hides and skins	A	A	A	A	A
Hydrolysed proteins from non-ruminants	A	A	A	A	A
Di and tricalcium phosphate of animal origin	U	A	A	A	A
Eggs and egg products, milk and milk products, colostrum and derivates	A	A	A	A	A
Animal proteins other than those mentioned ones	U	A	A	A	A

U: unauthorised
A: authorised
*: fishmeal is allowed for unweaned ruminants in milk replacers
†: intraspecies recycling is prohibited

Figure 3: Summary of Processed Animal Proteins and constituents of animal origin currently authorized in animal feed in the European Union. Source Table 1 *EURL-AP Standard Operating Procedure, 2022*.

Insects are part of the natural diet of farmed animals such as poultry and fish, therefore, regarded as excellent feed material (Belghit et al., 2019b; Makkar et al., 2014). Insect species were authorized in pig and poultry feed from 2021 and aquafeed from 2017 (European Commission, 2021/1372; European Commission, 2017/893; European

Commission 2021/1925). The eight insect species currently permitted in the European Union include (i) black soldier fly (BSF; *Hermetia illucens*), (ii) common housefly (*Musca domestica*), (iii) yellow mealworm (*Tenebrio molitor*), (iv) lesser mealworm (*Alphitobius diaperinus*), (v) house cricket (*Acheta domesticus*), (vi) banded cricket (*Gryllodes sigillatus*), (vii) field cricket (*Gryllus assimilis*), and (viii) silkworm (*Bombyx mori*) (European Commission, 2021/1372; European Commission, 2017/893; European Commission 2021/1925). Among these species, BSF has been considered the most promising for use in the aquaculture of commercially important fish species such as Atlantic salmon (Gillund & Myhr, 2010). Several feeding trials have been performed in which fish meal was successfully replaced with BSF meal (Belghit et al., 2019a; Lock et al., 2016). When used as feed ingredients in aquaculture (or any other animal feed), insects are considered PAPs (European Commission, 2021/1372). Insects are treated as farmed animals in the European Union (European Commission, 2017/893) and are thus subjected to the same rules and regulations (Figure 3), including the legislation on the prevention of TSE (European Commission, 2021/1372; European Commission, 2017/893).

Besides feed ingredients, insects can also be used as food ingredients due to their high protein, fat, and minerals. In the European food market, dried yellow mealworm (*Tenebrio molitor*) and frozen, dried, and powdered forms of house cricket (*Acheta domesticus*) are authorized as novel food (European Commission, 2021/882, European Commission, 2022/188).

1.3 Food and Feed Safety

Balancing security, safety, and sustainability in food production is crucial. Previous examples of safety failure, such as the mad cow disease epidemic outbreak due to bovine meat and bone meal usage, will help design future strategies (Vågsholm et al., 2020). Circular food systems, where slaughterhouse waste is processed and used as feed material, are excellent examples of recirculating quality nutrients from food chains (Woodgate et al., 2022). However, the outbreak of BSE indicated that circular economies could lead to severe hazards if regulations are not in place (Boqvist et al.,

2018). A cattle infected with BSE can infect 15-20 other cattle indicating that circular food production can also become a cycle of disease spread (Vågsholm et al., 2020). Therefore, it is essential to implement food safety regulations correctly in the circular food and feed chains.

Waste produced from food industries intended to be used as feed material must follow high safety standards to ensure consumer safety (Lavelli, 2021). The topics described below will provide an overview of hazards and precautions associated with circular food and feed chains.

1.3.1 PAP Ban and Anti-cannibalism Measures

Due to the risk of BSE and other TSEs, all PAPs were subsequently prohibited from animal feed according to the European Commission regulation (European Commission 2001/999). The objective of this ban was the eradication of TSEs, specifically BSE in cattle and scrapie in sheep (van Raamsdonk et al., 2019). After the complete ban on the use of PAP in 2001 to prevent the occurrence and spread of disease in the food chain, additional measures on anti-cannibalism were put into force in Europe (European Commission 2002/1774). In biobased economies implementing circularity, the avoidance of cannibalism of livestock is a crucial aspect. Regulations of anti-cannibalism have gained importance in preventing transmissible animal health-related issues such as viral and prion diseases (van Raamsdonk et al., 2019). The regulation extends to the use of by-products from the fishing and aquaculture industry, such as fishmeal; also, in this case, it is essential to identify species in the fish feed mixture to avoid the potential risk of disease due to cannibalism (van Raamsdonk et al., 2017).

The status of PAPs in Europe is “prohibited unless specifically exempted” as per the legislation that came into force in 2001 (European Commission 2001/999). Given the risks associated with BSE, ruminant PAP products such as blood, MBM, and other tissues are still prohibited from being used as feed products (van Raamsdonk et al., 2019). Whereas milk products from ruminants are legal feed ingredients; therefore, tissue-level differentiation of feed samples is crucial (Figure 3). Feed regulations

warrant the development of tools capable of detecting species and tissue-specific differences in feed material containing PAPs (Rasinger et al., 2016).

1.3.2 Allergenicity Risk Assessment of Insects as Food

Insects are regarded as novel food because they have not been widely consumed by humans in Europe prior to 1997 (Pali-Schöll et al., 2019a). In food safety, novel food materials should be safe for consumers concerning microbial, chemical composition, and allergenicity (Pali-Schöll et al., 2019b; van Huis, 2020). The EFSA evaluated the allergenic risk associated with novel insect ingredients as food (EFSA NDA panel 2021a; EFSA NDA panel 2021b) before several insect species were authorized by the European Commission for the European food market. Insect proteins can cause allergy by mediating adverse immune reactions due to exposure to insect and insect-derived products (Vågsholm et al., 2020). Methods for allergenicity assessment include immunological tests such as IgE reactivity (Pali-Schöll et al., 2019a). Mass spectrometry-based proteomic approaches were recently used to detect and quantify the allergen arginine kinase in house cricket samples (Bose et al., 2021). In other words, when collecting tandem mass spectrometry data for identifying insect PAP in feed, these data can also be screened for potential allergens, as recently shown by Varunjikar et al. (2022).

1.4 Food and Feed Fraud

Food fraud is increasing worldwide due to complex supply chains and the globalization of food markets (Visciano & Schirone, 2021). Acts such as adulteration, i.e., substitution with cheaper ingredients, dilution or adding impurities, and usage of unauthorized products, are referred to as food fraud by authorities worldwide. According to European Commission, “fraud is an act of omission relating to the use, or presentation of an incorrect, or incomplete statement or information or non-disclosure of information, violation of rules and any other illicit activity affecting the financial interests of the European Union” (Visciano & Schirone, 2021). Food fraud can affect consumer safety and rights; consumers must be appropriately informed about food contents. For example, if allergens in the food products are not well labeled or

undeclared, it might lead to severe consequences causing illness or even death of a consumer. The allergy-related labeling is thus crucial for sensitive consumers if fish, gluten, soybean, egg, or nuts are included in the food products (Visciano & Schirone, 2021). Also, for the consumers of soybean products, adulteration and mislabeling of genetically modified (GM) or non-GM can be a concern. For example, GM can present compositional differences and non-GM soybean seeds, or GM soybean seeds could be contaminated with glyphosate residues (Bøhn & Millstone, 2019). In Europe, if > 0.9% of an ingredient is derived from a GM product, then labeling the product as GM is mandatory (EFSA, 2010; European Commission 2003/1829). Also, according to European Union regulations, organic food should not include GM products (Lähteenmäki-Uutela et al., 2021). Therefore, it is necessary to differentiate GM soybean products from organic products in food and feed material.

In addition to missing information and mislabeling, the active adulteration of food might pose significant health hazards. For example, in 2008, melamine addition in milk in China was reported as a significant food safety incident (Bouzembrak & Marvin, 2016; Gossner et al., 2009). This incident affected about 300,000 Chinese infants and young children, causing severe health damage such as kidney and urinary tract effects and the death of six individuals (Gossner et al., 2009). Similarly, in 2013, in the European food market, the “horse meat scandal” was reported in seven countries where beef meat products were adulterated with horse meat (Madichie & Yamoah, 2017). The motivation for the action was purely economic gain; replacing beef with horse meat was an easy way to increase profits (McEvoy, 2016). Processed food products, such as burgers and sausages, were highly susceptible to adulteration. The horse meat scandal posed a safety hazard, as the adulterated horse meat was not initially destined for human consumption and was possibly contaminated with phenylbutazone, a veterinary steroidal drug (Visciano & Schirone, 2021). The adulteration raised food safety concerns due to this drug's toxicity and carcinogenic effects. Therefore, it is mandatory in the European Union to correctly label horse meat if present in food (Madichie & Yamoah, 2017).

Seafood is also susceptible to mislabeling and adulteration due to the immense diversity of seafood species being fished and cultured, the global nature of trade, and differential market values. According to the media monitoring system, 27% of frauds (mislabeling, artificial enhancement, substitution, or dilution) from 2000-2015 were seafood-related (Bouzembrak et al., 2018). A study on DNA metabarcoding of seafood in European mass catering showed that 26% of the samples were mislabeled (Pardo et al., 2018). For example, in Belgium, high-value fish such as sole (*Solea solea*) was replaced by cheaper species such as pangasius (*Pangasianodon hypophthalmus*) (Deconinck et al., 2020). Such illegal substitution and fraudulent activities also raise environmental concerns due to the potential inclusion of endangered species in the food chain. For example, an endangered blue shark (*Prionace glauca*) species was reported to be used as a substitute for the Atlantic cod (*Gadus morhua*) (Pardo et al., 2018). Such incidents can increase pressure on red-listed, overfished, and threatened species.

1.5 Food and Feed Forensics

Analytical tools for food and feed authenticity and traceability are becoming increasingly important in the global food chain (Saadat et al., 2022). Due to the complexity of food fraud, various tools are needed for food authentication (Silva, 2018). Food forensics is the capability to prove the authenticity of a food or feed product (Saadat et al., 2022). Several analytical techniques, including DNA-based methods, mass spectrometry, chromatography, immunoassay, and nuclear analytical techniques, can be applied to food and feed forensics (Saadat et al., 2022). Due to the wide variety of contaminants, adulterants, and hazards, a wide range of molecular tools are required for food forensics. For example, high-accuracy and cost-effective DNA methods have been used to develop rapid assessment tests for species identification in food samples (Beltramo et al., 2017; Toxqui Rodríguez et al., 2023). Due to the analytical flexibility offered by mass spectrometry, it has become a preferred tool for detecting infectious, allergenic, and toxic proteins in food products (Silva, 2018). The method can be used for species authentication and identifying adulterated substances in food or feed material (Saadat et al., 2022).

1.6 Existing Methods for Monitoring Feed

For the eradication of TSE and other prion diseases, three feed bans have been implemented in European legislation; (i) the ruminant ban, (ii) the extended feed ban, and (iii) the species-to-species ban (van Raamsdonk et al., 2019). The European Union Reference Laboratory for Animal Proteins (EURL-AP) is focused on developing analytical tools for PAP in the European Union (van Raamsdonk et al., 2019; Walloon Agricultural Research Centre, 2014). PAPs in feed products are highly processed, which can severely affect detection and identification methods. Although drastically heat-treated, PAP-containing feed products have been shown to contain amplifiable DNA for species detection (Fumière et al., 2006). The current operational protocol followed by the EURL-AP for feed analyses is shown in Figures 4A and B. The existing methods for detecting species and tissue are discussed below, indicating their strengths and limitations.

1.6.1 Microscopy Method

Optical light microscopy was the first official method to detect and characterize PAPs in feed materials after the implementation of the European ruminant ban (van Raamsdonk et al., 2019). The method can detect the presence of bone fragments in the given material and differentiate between fish, birds, and mammals (van Raamsdonk et al., 2017; van Raamsdonk et al., 2007). Specific staining techniques have been used to further enhance this method's sensitivity and tissue specificity to detect blood products and bone fragments (van Raamsdonk et al., 2011). Species-specific determination of PAP is not achievable with the classical microscopic method, and species-specific identification is becoming critical with the reintroduction of porcine PAP in the food chain (Fumière et al., 2006; Olsvik et al., 2017). A graphical representation in Figure 4A shows how testing is performed using this method.

1.6.2 DNA-based Method qPCR

The PAP ban is gradually being lifted in Europe, and the feed legislation demands more specific and sensitive methods for ruminant detection. Soon after the ban on PAP products, a DNA-based, qPCR method was developed (Fumière et al., 2006) and has

been used since 2013 to identify ruminant material in feed products. More recently, a standard operating procedure has been established by the EURL-AP (Figures 4A and B), which combines light microscopy and qPCR method to analyze feed containing animal products (EURL-AP, 2015). When a sample test positive using optical light microscopy, qPCR testing is performed, as shown in Figure 4B. For the qPCR method validation, multi-laboratory commercial PAP screening was performed, showing that the method can detect ruminant PAP in poultry, feather, and pork meals (Olsvik et al., 2017). The sensitivity and accuracy of the PCR-based method were among the key reasons that the EURL-AP chose it as the standard method for PAP detection (European Commission, 2013/51). However, DNA-based methods cannot be used to differentiate tissue origin, which is important given that ruminant milk is a legal ingredient and other tissue products from ruminants are prohibited in the feed material. Additional official methods are required that allow for species and tissue-specific PAP detection in feed.

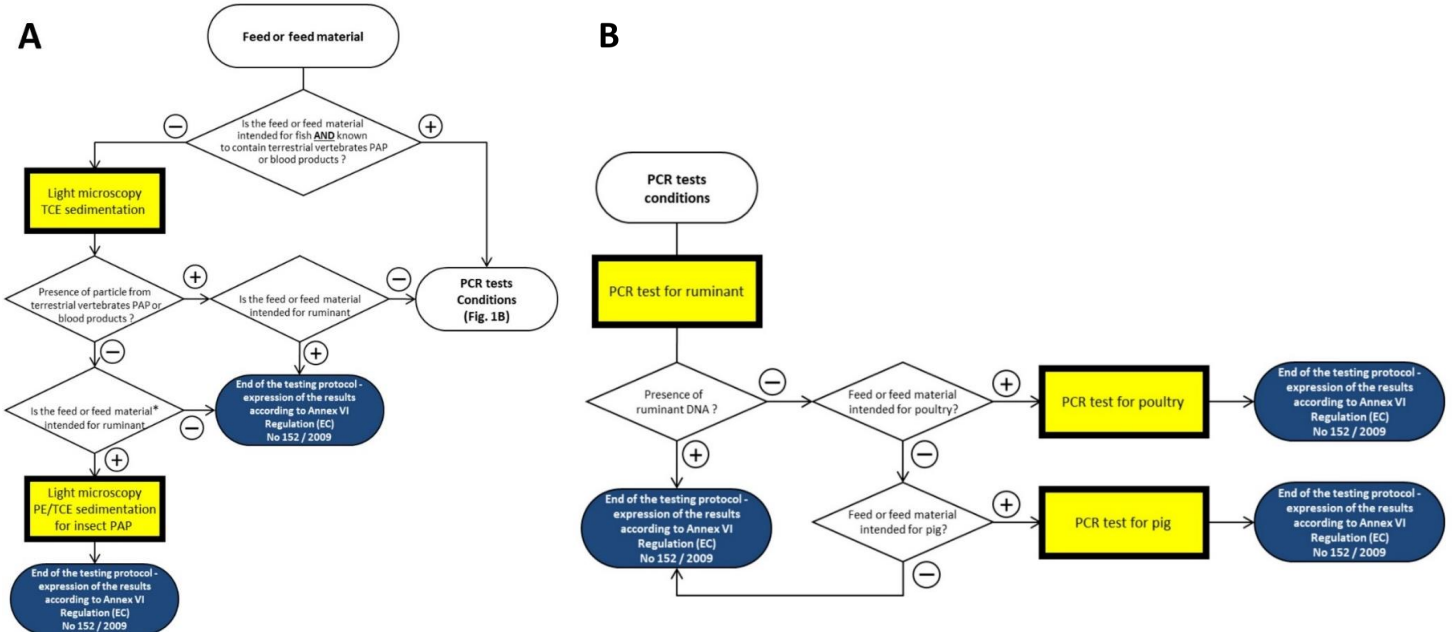


Figure 4: Standard operational protocol for the determination of animal source in feed material analyses followed by European Union Reference Laboratory for Animal Proteins (EURL-AP). Plus sign (+) indicates a positive result; minus sign a (−) negative result. Feed : compound feed; Feed material : products of vegetable or animal origin (A) light microscopy, (B) PCR test; Source: Figure 1A and B *EURL-AP Standard Operating Procedure, 2022*.

1.6.3 Other Methods

The near-infrared spectrometry (NIR) method is based on the profiling of samples using fundamental vibrations found in the region of the electromagnetic spectrum (van Raamsdonk et al., 2007). Light absorbed by the chemical bonds of any molecule can be measured using this method. This method was initially used following the Europe-wide PAP ban to detect the presence of PAP in the feed material. The limit of detection for PAP with NIR is generally higher than 1% and is above the current acceptable limit of detection of 0.1% (Fumière et al., 2006; van Raamsdonk et al., 2007; van Raamsdonk et al., 2019). The advantages of this method are cost-effectiveness and rapid analyses. However, the limit of detection is a drawback of this method, and further improvements are needed (van Raamsdonk et al., 2007).

As the established methods lacked tissue-specific identification of feed ingredients, antibody detection methods became important. Antibodies specific to tissue and species can be used as a method for the detection of PAP samples. However, heat treatment during processing can denature protein structures and affect detection (Huet et al., 2016). Heat-resistant proteins such as troponin, specific to muscle tissues, were considered ideal targets for developing antibody methods (van Raamsdonk et al., 2019). A commercial immunoassay kit developed for troponin-I protein detection (MELISA-TEK™ Ruminant) was found effective for PAP samples and successfully identified ruminant material in feed samples (Bremer et al., 2013).

1.6.4 Mass Spectrometry Methods

Mass spectrometry-based methods are used to perform proteomic analyses by separating and identifying peptides in the tryptic digest of PAP samples. For identifying tissue-specific proteins, untargeted or targeted proteomic methods can be used. Untargeted proteomic methods analyze the complete protein profile of samples without targeting a particular protein. Whereas, targeted methods analyzes one or multiple peptide markers using selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) (Yocum & Chinnaiyan, 2009). In general, SRM and MRM methods are suitable for detecting targeted proteins in food or feed matrix, given their high specificity and sensitivity of detection (Lecrenier et al., 2016; Marbaix et al.,

2016). Several peptide biomarkers have been applied to differentiate PAP proteins, such as hemoglobin, casein, beta-lactoglobulin, and collagen (Lecrenier et al., 2021, 2018). Using tissue-specific peptide markers, the separation of tissue from the same organism is feasible. Separating ruminant milk (which is a legal ingredient) from the illicit bone meal is possible with a MRM assay (Lecrenier et al., 2021). If required, it is possible to improve the sensitivity of targeted methods further when combined with immunoaffinity mass spectrometry (Marchis et al., 2017; Steinhilber et al., 2018b). For immunoaffinity mass spectrometry, protein digestion is followed by immunoprecipitation of the targeted peptides which are separated using mass spectrometry (Steinhilber et al., 2018a, 2018b; van Raamsdonk et al., 2019). However, significant hurdles associated with using this method for application in routine analyses are analytical time and high cost (van Raamsdonk et al., 2019).

Untargeted proteomic methods have also been used in differentiating species and tissue origin of PAPs. In tissue-specific-peptide biomarker detection, untargeted methods are used for screening tests before developing targeted assays (Marbaix et al., 2016). However, protein sequence information is required to analyze untargeted proteomic data, which is not readily available for food and feed-relevant species. Therefore, peptide biomarker detection is difficult with feed and food-relevant species. Moreover, with multiple novel feed and food ingredients being introduced into the food chain, developing and standardizing SRM and MRM assays for every species and tissue will be time-consuming. While dealing with feed and food adulterations, unknown protein sources can be included in the feed product, which is challenging to detect using targeted analyses. An emerging protein database-independent spectra library (SL)-based approach has been implemented for food analyses (Ohana et al., 2016; Wulff et al., 2013). The spectra library matching (SLM) approach has previously been used for tracing blood meal sources in ticks, identifying proteins from zebrafish embryos, fish and meat products, and PAP authentication (Ohana et al., 2016; Önder et al., 2013; Van Der Plas-Duivesteijn et al., 2014). However, generating SL for food and feed-relevant species is a prerequisite for using this method routinely in laboratories.

This PhD thesis implemented an untargeted shotgun proteomic method with a particular focus on the SLM for future food and feed safety regulations by identifying the tissue and species origin of samples.

2. Research Objectives

The goal of this PhD thesis was the development and implementation of untargeted proteomics tools for species and tissue-specific identification of protein sources in feed and food. To achieve this, a shotgun proteomic approach was implemented, and spectra libraries were created for feed and food samples.

The objectives of this work were as follows:

i. development and application of mass spectrometry-based shotgun proteomic workflows for regulatory science (**Papers I, III** and **V**),

ii. applying spectra library matching (SLM) for fish species identification in mixed samples (**Paper II**)

iii. development and testing of SLM-based proteomic approaches for tissue and species-specific PAP differentiation in feed (**Papers I** and **III**),

iv. implementing shotgun proteomics to detect and differentiate transgenic soy in feed (**Paper IV**).

3. Methodological Approach

3.1 Samples

During this PhD, food and feed samples were used to develop proteomic protocols and spectra libraries. Processed animal proteins, fish muscle tissues, and soybean sample were acquired from different sources described in respective papers.

3.1.1 Insect Samples (Paper I and III)

The control feeding medium (Ctl) for the black soldier fly larvae (BSF) was standard poultry feed, a reference medium for BSF larvae by the Laboratory of Entomology (Wageningen, The Netherlands). The control feed medium was mixed with bovine hemoglobin powder (BvHb) at three different concentrations, as described in **Paper I**.

Food and feed-grade insect species were selected for **Paper III**, where multiple insect species samples were collected from different orders. Eight samples of species from the Diptera order; black soldier fly larvae (BSF) (*Hermetia illucens*), nine samples of species from the Coleoptera order, including the yellow mealworm (YW) (*Tenebrio molitor*) and the lesser mealworm (LW) (*Alphitobius diaperinus*), and two samples from the Orthoptera order; house cricket (HC) (*Acheta domesticus*) were collected from different insect food and feed companies (Belghit et al., 2019b). Additionally, one morio worm (MW) (*Zophobas morio*) sample was included in the study (**Paper III,** Supplementary Table S1).

3.1.2 Fish Samples (Paper II)

A total of seven teleost species muscle tissues were selected for **Paper II** due to their commercial importance, 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*). For the fish mixture formulation, muscle tissues from platyfish, Nile tilapia, and Atlantic cod were weighed and mixed in the ratio: platyfish 1/6, tilapia 2/6, and cod 3/6, forming a mixed tissue sample ("fish mixture"). Details of the preparation were described in **Paper II.**

3.1.3 Soybean Samples (Paper IV)

Samples were obtained from fields in Iowa, USA, and information such as seed types, cultivation process, and pesticides usage was described by Bøhn et al., 2014. Detailed information is given in Supplementary Table S1 of **Paper IV**.

3.2 Sample Preparation

3.2.1 Protein Extraction

Samples were weighed into a test tube of the One Plus Grinding kit (GE Healthcare Life Science, 80648337, Piscataway, NJ, USA). Lysis buffer was added to the samples (4% SDS, 0.1M Tris-HCl, pH 7.6), and samples were homogenized in the tube containing resins with a pestle. Freshly prepared, 3 µL of 1M Dithiothreitol was added to this homogenate to obtain a final concentration of 0.1M; further, these tubes were centrifuged for 10 minutes at 15,000 g to remove resin and other debris. The supernatant was collected and heated at 95°C on a heat block for 5 min. After this, samples were centrifuged, and the supernatant was collected in new tubes and stored at -20°C until further processing. The protein concentration of extracted samples was determined by the Pierce 660 assay using BSA for the standard curve (Thermo Scientific, San Jose, CA).

3.2.2 Protein Digestion and Purification

Protein extracts from samples were digested with a filter-aided sample preparation method described in **Papers II** and **III**. Extracted protein (150 mg) was diluted with 200 µL of 8M urea solution prepared in Tris-HCl (100mM, pH 8.5). Disulfide bonds in the samples proteins were broken using 1,4-dithiothreitol (DTT). This solution was transferred to an ultrafiltration spin column (Microcon 30, Millipore, Burlington, MA, USA). Further, these proteins were alkylated with 50 mM of iodoacetamide (C_2H_4INO) for 20 min before incubation in darkness at room temperature. After incubation, the protein mixture in the column was washed with 200 µL of 8M urea solution along with 100 µL of 50 mM ammonium bicarbonate (NH_4HCO_3) solution. Trypsin was added to filters in a 1:50 enzyme-to-protein ratio, and tubes were incubated for 16 hours at 37 °C. Filters were centrifuged and washed with 40 µL of 50 mM ammonium bicarbonate

solution and later with NaCl (0.5M). Desalination and cleaning of the peptides were performed using Pierce™ C18 spin column (ThermoFisher, 89870) as described in **Paper III**.

3.3 High-performance Liquid Chromatography- Mass Spectrometry (HPLC-MS/MS)

The samples used during this PhD were analyzed on three different High-performance liquid chromatography – Tandem Mass spectrometry (HPLC-MS/MS) instrument setups. For **Paper I**, data was acquired from QTOF; for **Papers II** and **IV**, data were acquired from the proteomics facility at the University of Bergen (PROBE). Lastly, for **Paper III**, an in-house instrument, normal flow coupled with HPLC Q-Orbitrap, was used to acquire data using a method that was developed during this PhD.

3.3.1 HPLC-MS/MS UHR-TOF (Paper I and III)

The protein digest was analyzed by LC-ESI-MS/MS maXis Impact UHR-TOF (Bruker, Bremen, Germany) mass spectrometer (MS) coupled with a UPLC Dionex UltiMate 3000 (Thermo). The peptide samples were separated by reverse-phase liquid chromatography using a flow rate of 40 µL and Acclaim PepMap 100 C18 (1.0 mm × 15 cm) Thermo column in an Ultimate 3000 liquid chromatography system. Mobile phase A was 95% of water, 0.1% formic acid, and 2% acetonitrile. Mobile phase B was 20% water, 80% acetonitrile, and 0.1% formic acid. The digest (10 µl) was injected, and the organic content of the mobile phase was increased linearly from 5% to 40% in 75 min (**Paper I**), 4% to 40% B in 60 min (**Paper III**), and from 40% B to 95% B in 10 min. The column effluent was directly connected to the UHR-TOF instrument. In the survey scan, tandem mass spectra were acquired for 0.5 s in the *m/z* range between 50 and 2200. The 10 most intense peptide ions, 2+ to 4+, were fragmented. Mass spectrometry data were converted using DataAnalysis 4.2 (Bruker) and exported as mzXML files. In **Paper III**, the method was regarded as microflow-HPLC QTOF (MF-HPLC QTOF).

3.3.2 HPLC-MS/MS LTQ-Orbitrap Elite (Papers II and IV)

Peptide samples were dissolved in 2% acetonitrile and 0.1% formic acid, as described by Bernhard et al. (2018). Samples were injected into an Ultimate 3000 RSLC system (Thermo Scientific, CA, USA) coupled with a linear quadrupole ion trap-Orbitrap (LTQ-Orbitrap Elite) mass spectrometer (Thermo Scientific, Bremen, Germany). Samples were desalinated on a pre-column Acclaim PepMap 100 (2 cm×75 μm) nanoViper C18 column at a flow rate of 5 μl/min for 5 min with 0.1% trifluoroacetic acid. Peptides were separated using a biphasic acetonitrile gradient from two nanoflow UPLC pumps (flow rate of 270 nl/min) on a 50 cm analytical Acclaim PepMap 100 (50 cm×75 μm) nanoViper column. Solvents A and B were 0.1% trifluoroacetic acid (v/v) in water and 100% acetonitrile, respectively. The gradient composition was 5% for 5 min, followed by 5–7% B for 1 min, 7–21% B for 134 min, 21–34% B for 45 min, and 34–80% B for 10 min. The mass spectrometer was operated in the DDA mode to automatically switch between full-scan MS and MS/MS acquisition. Instrument control was through Tune 2.7.0 and Xcalibur 2.2. Survey full-scan MS spectra were acquired in the Orbitrap with an Automatic Gain Control (AGC) target value of 1×10^6. The 12 most intense eluting peptides were fragmented in the high-pressure linear ion trap by CID. Mass spectrometry data were collected in .Raw format.

3.3.3 HPLC-MS/MS HR-MS Orbitrap (Paper III)

For the optimization, HPLC analyses were performed using Vanquish Horizon binary HPLC (Thermo Scientific, San Jose, CA). Separations were performed using 2.2 μm Acclaim Vanquish C18, 2.1 x 250 mm (Thermo Scientific, San Jose, CA). A and B solvents were 0.1% (v/v) formic acid in high-purity water and 0.1% formic acid (v/v) in 100% acetonitrile, respectively. Gradient conditions were described in Supplementary Table S2, **Paper III,** with different gradient lengths varying from 60-80 min. The flow rate varied between 300 and 400 μL/min (**Paper III**, Supplementary Table S2). Different amounts of HeLa cells digest were loaded (**Paper III,** 0.5-40 μg, Supplementary Table S3).

Eluting peptides were analyzed on a High resolution - Mass spectrometry (HR-MS) Q Exactive Orbitrap (Thermo Scientific, San Jose, CA). MS instrumental tune parameters

were set as follows: ESI spray voltage was 3.5 kV, sheath gas flow rate was 40 AU, the auxiliary gas flow rate was 10 AU, the capillary temperature was 320°C, probe heater temperature was 400°C, and S-lens RF level was set to 50. In DDA mode, resolution settings of 17,500, 35,000, and 70,000 were tested (**Paper III**, Supplementary Table S2). The mass range was set at 200-2000 *m*/*z*, and an AGC target was 5.0×10^5 up to 3.0×10^6 with a maximum injection time of 50 ms. For MS2, the resolution settings were 17,500 and 35,000 at a fixed first mass of 140 *m/z* with an AGC target value of 5.0×10^5 and an isolation window of 1.2 *m*/*z*. The normalized collision energy set was 32, and the top 10 precursors were selected for fragmentation. The signal intensity threshold was 2.0×10^4 with dynamic exclusion of 10, 20, and 30 s (**Paper III,** Supplementary Table S2). This method was called analytical flow-HPLC HR-MS (AF-HPLC HR-MS) in **Paper III**.

After the optimization of the HPLC and MS parameters with the HeLa Digest, the developed AF-HPLC HR-MS workflow was implemented to analyze nineteen insect meal samples (**Paper III**). Gradient conditions were 2% B to 35% B in 62 min, hold at 95% B until 5 min, and 2% B from 67.1 until 80 min. The flow rate was 400 µL/min (**Paper III**, test number 19 in Supplementary Table S2). MS scans were obtained at a resolution of 70,000. The mass range was set at 350-2000 *m*/*z*, and the AGC target was 3.0×10^6 with a maximum injection time of 50 ms. For MS2, the resolution was 35,000 at a fixed first mass of 140 *m*/*z* with an AGC target value of 3.0×10^6 and an isolation window of 1.2 *m*/*z*. The normalized collision energy set was 32, and the top 10 precursors were selected for fragmentation. The signal intensity threshold was 2.0×10^4 with dynamic exclusion of 30 s.

3.4 Bioinformatics Analyses

Data analysis is a central aspect of proteomic research and the most time-consuming part of analyses. Various open-source software are available for analyses of proteomics data due to the growing interest in this field (Perez-Riverol et al., 2014). In this PhD, proteomic data were analyzed using open-source software to ensure the free usage of the developed method across all laboratories and institutions.

3.4.1 Direct Spectra Comparison Using compareMS2

For molecular phylogenetic analyses using compareMS2 (Palmblad & Deelder, 2012) version 0.0.4 and version 0.0.5 (compareMS2 GUI, 2021) .mgf files with 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 calculate distance matrices and UPGMA trees in MEGA (**Papers II**, **III**, and **IV**).

3.4.2 Spectra-database Matching Using Search Engine

Mass spectrometry data generated were converted from .Raw or .baf format and exported as mzML files using msConvert (version: 3.0., ProteoWizard (Kessner et al., 2008). Depending on the sample, dataset reference proteomes for the search were selected; for example, in **Paper I**, bovine hemoglobin and milk data spectra were searched against the bovine reference proteome obtained from UniProt (UP000009136; accessed on December 2020). Insect data was matched against *Hermetia illucens* proteins (UniProtKB; accessed on December 2020) using X! Tandem (Craig & Beavis, 2004). Generated pepXML files were further analyzed using PeptideProphet and ProteinProphet with 1% FDR (Keller et al., 2002). The list of Uniprot proteomes and Uniprot KB ids as given in Table 1, along with the search engine used for the spectra searching, and details of search settings were described in respective **Papers I**, **II**, **III**, and **IV**. For **Papers I**, **II** and **III**, TPP were used and for **Paper IV** MaxQuant software were used for database searches.

Table 1: Uniprot reference proteome or Uniprot KB id and search engines used in respective papers.

Organism	Scientific name	Uniprot ID	Paper	Search Engine
Bovine	*Bos taurus*	Proteome id UP000009136	I	X! Tandem
Black Soldier Fly	*Hermetia illucens*	UniprotKB "*Hermetia illucens"*	I	X! Tandem
Zebrafish	*Danio rerio*	Proteome id UP000000437	II	Comet
Arthropoda species	-	-	III	Comet
Soybean	*Glycine max*	UP000008827 + sequence of A0A140GBJ6	IV	Comet, Andromeda

3.4.3 Spectra Library Building

Spectra libraries (SL) were created using SpectraST (version 5.0), as described by Lam (2011). All sample spectra were searched against respective spectra libraries for relative quantification of samples using TPP (Deutsch et al., 2015). Dot products above 0.7 were considered valid matches and used for quantification. A graphical overview of the SLM workflow and example output of matching spectra are shown in Figures 3 and 4, respectively. Outputs of SLM were recorded using tidyverse functions (version 1.3.0 (Wickham et al., 2019)) and UpSetR (version 1.4.0). In **Paper II**, the SLM approach was slightly modified as described in the method section. Data used in published papers and SLs were made available on MassIVE.

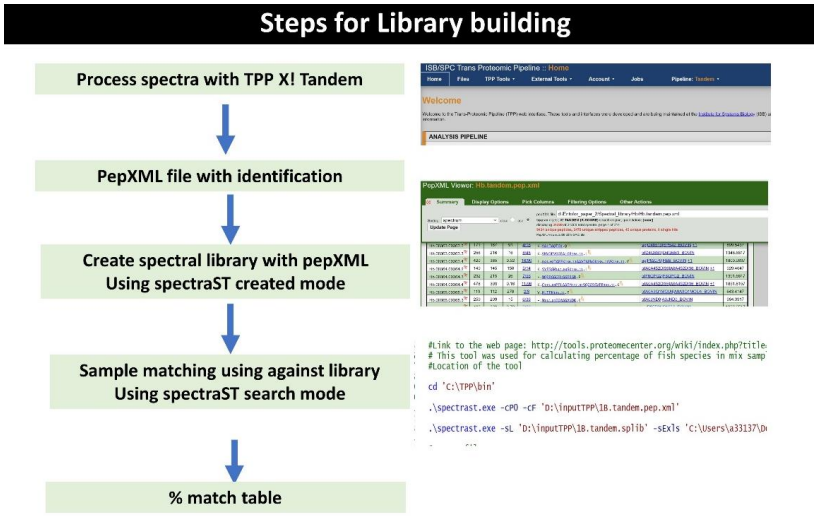
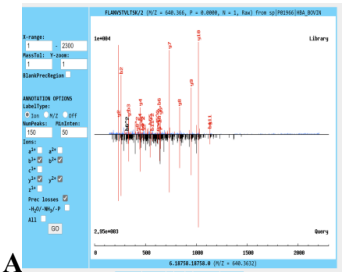


Figure 5: Workflow used for building spectra libraries for authentication of black soldier fly larvae reared on prohibited substrate. Source: Supplementary material of **Paper I**.



A

b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
148.0757	74.5415	1	F	12		
261.1598	131.0835	2	L	11	1132.6572	566.8322
332.1969	166.6021	3	A	10	1019.5732	510.2902
446.2398	223.6235	4	N	9	948.5360	474.7717
545.3082	273.1577	5	V	8	834.4931	417.7502
632.3402	316.6738	6	S	7	735.4247	368.2160
733.3879	367.1976	7	T	6	648.3927	324.7000
832.4563	416.7318	8	V	5	547.3450	274.1761
945.5404	473.2738	9	L	4	448.2766	224.6419
1046.5881	523.7977	10	T	3	335.1925	168.0999
1133.6201	567.3137	11	S	2	234.1448	117.5761
		12	K	1	147.1128	74.0600

B

Figure 6: Spectra library matching of samples to bovine hemoglobin library for dot product calculation (A) Spectra matching and (B) Table of ion annotation. Source: Supplementary material of **Paper I.**

3.4.4 Allergen Detection

In **Paper III**, due to the food relevance of the insect species, shotgun-proteomics was employed for allergen detection. A list of food allergens relevant for insect species was downloaded from www.allergen.org, along with allergen families and biochemical names. These allergen sequences were downloaded from UniProt to create a database. The list is given in Supplementary Table S6, **Paper III**. Spectra from the UHR-TOF and HR-MS instruments were searched against the database using TPP and Comet search engines to evaluate allergen detection ability. Data processing and statistical

comparison of detected allergenic proteins in samples were performed in Omics Explorer.

3.4.5 Pathway Analyses in AgriGO

Differentially expressed proteins with $p<0.01$ and a log fold change of higher than 0.5 or lower than -0.5 were subjected to pathway-level analyses using the tool AgriGO V2.0 (Du et al., 2010) (**Paper IV**, Supplementary Table S5). GO term reduction was performed as described in **Paper IV**.

4. General Discussion

In sustainable food production systems, using advanced analytical tools for next-generation risk assessment is critical. For regulatory purposes, omics tools are considered key to supporting and upholding current food and feed safety standards (EFSA, 2018b). By 2030, it is expected that EFSA will routinely apply omics approaches in food chain analytics to enhance food and food safety-related risk assessments (EFSA, 2022). In parallel with other omics approaches, the field of proteomics has been advancing rapidly over the past decade. Implementing proteomic tools for non-model organisms, including farm animals, aquaculture, and insect species, can benefit feed and food safety research (Heck & Neely, 2020; Neely & Palmblad, 2021). This PhD aimed to implement and develop proteomics-based approaches for regulatory science in food and feed analyses.

There are safety challenges associated with food and feed products. From the perspective of circularity, the reauthorization of PAP as feed ingredients in Europe raised authentication challenges, which required the development of analytical tools to differentiate species and tissue-level identities of samples. For the control of feedstuff across Europe, standard procedures were established by the EURL-AP, including optical light microscopy and qPCR for ruminant DNA detection (European Commission, 2013/51). The qPCR method is sensitive and valuable for species detection, but this method is not tissue-specific; for example, authorized milk products cannot be differentiated from blood or bone products from ruminants (Lecrenier et al., 2020). During this PhD, in **Papers I** and **III**, proteomic approaches were developed for detecting, differentiating, and tracing prohibited PAP in the feed chain as required by current European Union legislations (European Commission, 2013/51; European Commission, 2013/56; European Commission, 2021/1372; European Commission, 2017/893; European Commission, 2017/1017; European Commission 2021/1925). The untargeted proteomic approaches developed during this work were also implemented to solve food fraud challenges, such as substituting expensive fish species with cheaper ones, adding non-permitted species into food material, or mislabeling GM products (**Papers II** and **IV**). Recently, omics tools also were proposed as a strategy for an

efficient evaluation of the safety of GM products (Gould et al., 2022). The proteomic analyses of GM and non-GM soybean seed performed during this PhD can contribute to the omics-based evaluation of GM plants (**Paper IV**). Moreover, in light of the potential allergy risk of insects as novel foods (Ribeiro et al., 2021), the proteomics workflows developed in this PhD thesis can be implemented to assess the allergenicity of food-relevant insect samples (**Paper III**).

Database-independent and database-dependent approaches are available for untargeted proteomics data analyses. Both were implemented during this PhD thesis to address food and feed safety regulatory challenges (Figure 7). Due to a general lack of reference proteomes for food and feed-relevant species (Rasinger et al., 2016), proteomic database-independent approaches were considered most suitable for this work. Database-independent proteomic tools compareMS2 and SpectraST were implemented during this PhD (Figure 7). The compareMS2 was used as quality control software to build a molecular phylogenetic tree to evaluate tandem mass spectrometry data acquired from relevant species (**Paper V**). SLM was implemented using the SpectraST tool to identify and quantify species and tissue origin. SpectraST has been previously used to detect food fraud and to reveal the species origin of blood samples (Nessen et al., 2016; Ohana et al., 2016; Önder et al., 2013; Wulff et al., 2013). The analyses conducted during this PhD thesis (**Papers I**, **II**, and **III**) showed that, in addition to earlier reports in the literature, the SLM-based proteomic approach was also suitable for identifying species and tissue origin of food and feed samples.

Proteomics data collected from food and feed-relevant samples were analyzed using database-dependent approaches by matching data against proteomic reference databases in **Papers III** and **IV** (Figure 7). Previously, results from database searches were used to discover species or tissue-specific peptide markers to develop targeted proteomic assays for PAP in feed products (Lecrenier et al., 2018; Marbaix et al., 2016; Niedzwiecka et al., 2019; Steinhilber et al., 2018a, 2018b, 2019). Species or tissue-specific peptides detected during untargeted proteomic analyses of PAP samples could also be used to rapidly develop MRM assays for novel food and feed products when new legislation is enforced. In **Papers III** and **IV**, species-specific protein markers for

insect species and GM soybean were detected, which can be potential targets for developing new MRM assays. Since MS data from **Papers III** and **IV** were made available publicly (MassIVE ids: MSV000088034, MSV000087026, MSV000087017, and MSV000089618), the development of such markers is not only restricted to the laboratories at the IMR but can be performed by any stakeholder in the food and feed sector.

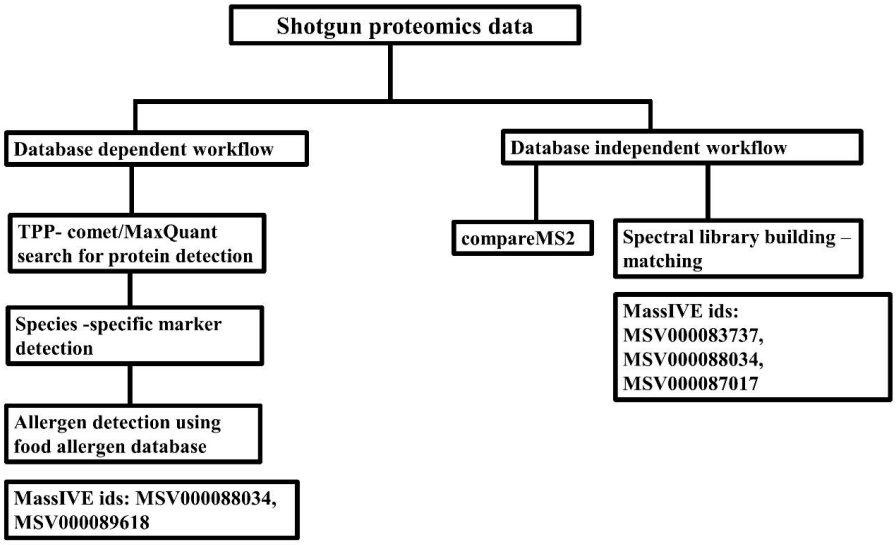


Figure 7: Proteomic bioinformatics workflow implemented in this PhD work for food and feed-relevant samples, modified from **Paper III**, Supplementary Figure 2.

In addition to peptide marker detection, proteomics also can be implemented to identify allergenic peptides from insect species. At the time of writing, PAPs from eight insect species were permitted to be used as feed material in Europe (IPIFF, 2021). The authentication of these permitted species is essential in the regulatory aspect (European Commission, 2021/1372; European Commission, 2017/893; European Commission 2021/1925). In addition, two insect species are permitted to be used as foods in the European Union (IPIFF, 2022). For insect species authorized to be used for human consumption, in addition to their detection and tracing in the food chain, allergenicity assessment of these products has become vital for food safety (Ribeiro et al., 2021). As a part of this PhD, in **Paper III**, allergenic proteins were detected using the database-

dependent approach (Figure 7). Similar to PAP peptide marker detection, allergenic peptides detected during this PhD can be used in developing MRM assays for food-allergen detection, as Bose et al. (2021) demonstrated.

The results obtained in this PhD thesis showed that proteomic methods were crucial to address challenges in sustainable and circular food systems. In the following sections, a more detailed account is presented of how the different proteomics approaches investigated in this PhD thesis can contribute to the detection and differentiation of food and feed-relevant species using shotgun proteomic methods.

4.1 Tissue and Species Differentiation of PAPs (Papers I and III)

Due to the risk of TSE spread, insects PAPs are subjected to strict regulations. Potential fraud with insect PAPs in the feed chain involves using prohibited insect species and insects reared on non-permitted substrates such as ruminant blood, bone, or other tissues (except milk products). Bovine milk is a permitted material used in feed products and can also be used to feed insects. In contrast, bovine blood is a prohibited material and is not to be used as a feeding ingredient according to current feed regulations (European Commission, 2013/56). Mass spectrometry-based proteomic methods can be implemented along with the official qPCR method to tackle the challenge of tissue-specific identification (Lecrenier et al., 2021; Rasinger et al., 2016).

During this PhD, a SLM was developed and implemented for tissue-level differentiation of PAP samples. Most of the mass spectrometry-based proteomics approaches for authentication and identification of feed samples quantify or identify one or more target-specific peptides in the samples (Lecrenier et al., 2018, 2021, Steinhilber et al., 2018b, 2019). SLM is an alternative proteomic approach that identifies and quantifies samples using previously collected reference libraries (Lam et al., 2008; Nessen et al., 2016; Ohana et al., 2016; Wulff et al., 2013). Spectra libraries used for SLM can be created using either identified or unidentified spectra; this allows the identification of non-model organisms when genomic or proteomic information is

unavailable (Nessen et al., 2016; Önder et al., 2013). Previously, spectra libraries with unidentified tandem mass spectra were effectively used for tracking the source of blood meal in parasitic arthropods (Önder et al., 2013), showing the effectiveness of the SLM in identifying blood remnants from mammalian origin. Therefore, this method was implemented to detect illicit ingredients such as bovine blood (PAP) in the substrate and BSF insect larvae (**Paper I**) feed and insect species.

In **Paper I**, BSF larvae were reared on feeding media spiked with PAP, bovine hemoglobin (BvHb) powder at 1%, 5%, and 10% (w/w) for seven days. An additional dietary group was fed 10% BvHb initially and then a control diet for seven more days. Contents of BvHb in spiked feeding media and remnants of BvHb from BSF larvae were detected when samples of insect diet and insects were matched against the BvHb reference spectra library. The SLM successfully separated bovine milk protein from bovine blood (Table 2) in feeding media spiked with BvHb. When BSF larvae fed on contaminated media were analyzed using SLM, BvHb was detected only at 5% and 10% w/w (Table 3). In BSF larvae fed 1% BvHb, the ruminant blood protein was not detected using SLM (Table 3).

Table 2: Detection of ruminant material in the feeding media used for the black soldier fly larvae growth trial (modified from Table 2 **Paper I**).

	qPCR (labs A, B)		Targeted MS (labs A, B, C)								SLM (lab D)	
			LC-MS/MS		IA-LC-MS/MS (protein IP)	IA-LC-MS/MS (peptide IP)						
	Ruminant DNA		Hb	MP[1]	Hb	Hb	PP	MP[2]	MY	CP	Hb	MP
Ctl	+	+	+	+	-	+	-	-	-	-	+	+
BvHb 1%	+	+	+	+	+	+	-	-	-	-	+	+
BvHb 5%	+	+	+	+	+	+	+			-	+	+
BvHb 10%-Ctl	+	+	+	+	+	+	+			-	+	+

Table 3: Detection of ruminant material in black soldier fly larvae grown on feeding media containing bovine hemoglobin powder (n = 2) (modified from Table 3 **Paper I**).

	qPCR (labs A, B)		Targeted MS (labs A, B, C)								SLM (lab D)	
			LC-MS/MS		IA-LC-MS/MS (protein IP)	IA-LC-MS/MS (peptide IP)						
	Ruminant DNA		Hb	MP[1]	Hb	Hb	PP	MP[2]	MY	CP	Hb	MP
Ctl	-	-	-	-	-	+	-	-	-	-	-	+
	-	-	-	-	-	+	-	-	-	-	-	+
BvHb 1%	+	-	-	-	-	+	-	-	-	-	-	+
	+	-	-	-	-	+	-	-	-	-	-	+
BvHb 5%	+	-	-	-	+	+	-	-	-	-	+	+
	+	+	+	-	+	+	-	-	-	-	+	+
BvHb 10%	+	+	-	-	+	+	-	-	-	-	+	+
	+	+	-	-	+	+	-	-	-	-	+	+
*BvHb 10%	+	-	-	-	-	+	-	-	-	-	-	+
	-	-	-	-	-	+	-	-	-	-	-	+

Footnote Tables 2 and 3: Plus sign (+) indicates a positive result; minus sign a (−) negative result. Unexpected results were marked in red. Workflows: LC-MS/MS (laboratory A, triple quadrupole); immunoaffinity-LC-MS/MS (IA-LC- MS/MS), IA on protein level (laboratory B, Q-TOF); IA-LC-MS/MS, IA on peptide level (laboratory C, triple quadrupole); SLM, spectra library matching (laboratory D, Q-TOF). Bovine proteins identified: Hb, hemoglobin; PP, plasma proteins: α2 macroglobulin and complement component 9; MP, milk protein: 1 Beta-lactoglobulin, casein, and 2 osteopontins; MY, muscle protein: myosin 7; CP, cartilage protein: matriline 1. Detailed analysis outputs were presented in Supplementary Tables 1-6 **Paper I**.

In addition to SLM, five other molecular methods were employed for BvHb detection (**Paper I**), including (i) real-time-PCR analysis, (ii) multi-target ultra-high performance liquid chromatography coupled to tandem mass spectrometry (UHPLC-MS/MS), (iii) protein-centric immunoaffinity-LC-MS/MS, (iv) peptide-centric immunoaffinity-LC-MS/MS, and (v) compound-specific amino acid analysis (CSIA) (Results are not shown in Table 3). Peptide-centric immunoaffinity LC-MS/MS (Tables 2 and 3) displayed lower limits of detection when compared to SLM, as it detected the presence of BvHb in all categories of BSF larvae (Table 3). The observed

differences in the limits of detection of BvHb might be due to the implementation of immunoaffinity binding in this method to enrich targeted peptides. However, SLM detected BvHb without targeting any specific peptide but using all tandem mass spectra collected from samples, arguably less laborious than peptide-centric immunoaffinity LC-MS/MS. The sensitivity of peptide-centric immunoaffinity LC-MS/MS involves target detection, development of affinity binding assay, and high cost per sample. Whereas the untargeted method SLM is robust and easy to implement. In addition, differences in the homogeneity of samples could also have interfered with the correct detection in complex feed matrices affecting the results of SLM (Marbaix et al., 2016).

The SLM method was applied for detection of insect PAPs from four authorized insect species and one unauthorized species. For these insect species, spectra libraries were created using two different instruments to analyze the robustness of SLM further when different HPLC gradients and MS instruments were used. Data generated from the micro flow-HPLC QTOF (MF-HPLC QTOF) and analytical flow- HPLC HR-MS (AF-HPLC HR-MS) was used to build spectra libraries (**Paper III**). The libraries created with MF-HPLC QTOF data had an average of 12,617 spectra, and the libraries created on AF-HPLC HR-MS comprised of 9,433 tandem mass spectra. Tandem mass spectra from samples of insect species were matched to these five insect spectra libraries. The identity of samples was confirmed by calculating the number of matches against reference spectra libraries of insect species (**Paper III**, Figure 2C).

In **Paper III**, libraries built on two different instruments with different collision-induced dissociation patterns were tested to evaluate the robustness of SLM. The Q-Orbitrap from Thermo-Scientific uses higher energy collision-induced dissociation (HCD) Cells (Kalli et al., 2014; Nessen et al., 2016), and collision energy influences spectra matching in SLM (Lam, 2011; Nessen et al., 2016). Therefore, query samples collected from one instrument were matched against the reference spectra library created on another instrument and vice versa (i.e., a cross-matching of datasets was performed). The output of the cross-matching is given in Table 4. When query spectra were collected from the same instrument used for spectra library creation, the number of matching spectra was much high, and species were correctly identified (Table 4).

Alternatively, when query spectra were collected from different instruments to the ones used to build reference libraries, the number of matching spectra was much low, yet species were still identified correctly (Table 4). The results were similar to a previous study on the differentiation of closely related flatfish species comparing a Q-Orbitrap and amaZon ion trap (Nessen et al., 2016). The outcomes of comparisons showed that with SLM, correct identification of species is possible even if spectra libraries and query samples were obtained on different instruments. Results show that spectra libraries created in this PhD thesis can be used across different laboratories for the identification of insect PAP samples. In future, regulatory laboratories such as the EURL-AP can provide standard materials to build universally usable spectra libraries for PAP detection. Using SLM, standardized procedures and protocols (SOPs) could be developed for the species and tissue-specific identification of feed material of animal origin.

In summary, **Papers I** and **II** show that it is possible to differentiate tissue and species origin of samples destined for use in feed using SLM. This method can complement official analysis methods (qPCR and light microscopy), and novel targeted MS-based methods currently developed by EURL-AP. As a part of this PhD, bovine blood, milk, and insect spectra libraries were built for in-house use. In addition, data was made publicly available to the scientific and regulatory community through the MassIVE data repository (MassIVE id: MSV000087026, MSV000083737, and MSV000088034).

Table 4: SpectraST output table indicating spectra matches of samples to four insect species spectra libraries built on different instruments (Directly taken from the supplementary material of **Paper III**).

QTOF vs QTOF				
Sample	**BSF library**	**HC library**	**LW library**	**YW library**
BSF-1	1210	68	82	75
BSF-2	1059	152	182	173
BSF-3	1253	125	107	126
BSF-4	971	130	135	149
BSF-5	1112	180	132	114
BSF-6	NA	122	127	151
BSF-7	1428	133	296	133
BSF-8	1041	96	89	97
HC-17	111	1458	423	566
HC-18	113	NA	429	431
LW-14	85	471	NA	661
LW-15	138	414	1072	449
LW-16	115	228	1573	481
MW-19	144	297	631	445
YW-10	93	328	487	2071
YW-11	101	243	404	NA
YW-12	101	155	227	2300
YW-13	114	312	365	2860
YW-9	121	141	305	1674

HR-MS vs HR-MS				
Sample	**BSF library**	**HC library**	**LW library**	**YW library**
BSF-1	1726	246	246	338
BSF-2	1562	315	280	284
BSF-3	2039	276	234	261
BSF-4	1995	345	259	321
BSF-5	1246	203	182	167
BSF-6	NA	381	326	334
BSF-7	2006	401	296	365
BSF-8	1013	95	83	96
HC-17	373	2570	580	504
HC-18	392	NA	596	520
LW-14	314	573	NA	810
LW-15	380	1008	1833	1009
LW-16	403	551	1947	1037
MW-19	452	605	1048	839
YW-10	438	662	953	3586
YW-11	331	501	815	NA
YW-12	220	223	412	1353
YW-13	271	420	628	2759
YW-9	279	366	611	2633

QTOF vs HR-MS				
Sample	**BSF library**	**HC library**	**LW library**	**YW library**
BSF-1	533	18	21	21
BSF-2	575	39	30	48
BSF-3	502	27	24	26
BSF-4	469	40	30	41
BSF-5	291	38	31	34
BSF-6	NA	26	27	29
BSF-7	461	33	26	41
BSF-8	430	25	21	31
HC-17	81	210	87	81
HC-18	77	NA	95	56
LW-14	51	114	NA	125
LW-15	54	68	163	91
LW-16	40	67	163	80
MW-19	51	69	72	68
YW-10	68	84	115	302
YW-11	145	147	192	NA
YW-12	124	76	109	407
YW-13	165	198	236	586
YW-9	36	45	63	261

HR-MS vs QTOF				
Sample	**BSF library**	**HC library**	**LW library**	**YW library**
BSF-1	45	22	23	34
BSF-2	53	22	21	22
BSF-3	70	31	22	19
BSF-4	57	27	21	21
BSF-5	32	17	15	13
BSF-6	NA	29	25	27
BSF-7	63	27	20	21
BSF-8	28	14	10	11
HC-17	12	108	43	42
HC-18	10	NA	40	39
LW-14	12	35	NA	53
LW-15	15	65	83	59
LW-16	14	37	87	53
MW-19	17	41	49	50
YW-10	15	39	66	173
YW-11	11	28	41	138
YW-12	14	18	25	69
YW-13	13	30	40	138
YW-9	11	30	38	121

Footnote Table 4: Number of matches to libraries created on both instruments were reported (data used for Figure 3C in **Paper III**); TOF vs. TOF: data collected from TOF and library created on TOF; HR-MS vs. HR-MS: data collected from HR-MS and library created on HR-MS; TOF vs. HR-MS: data collected from TOF and library created on HR-MS; HR-MS vs. TOF: data collected from HR-MS and library created on TOF; BSF – black soldier fly; YW – yellow mealworm; LW – lesser mealworm; HC– house cricket; MW – Morio Worms.

4.2 Species Authentication and Quantification of Fish Mixtures (Paper II)

Untargeted proteomics provides a helpful toolkit for food forensics due to its analytical flexibility and the capability to simultaneously detect and quantify several proteins (Silva, 2018; Saadat et al., 2022). When dealing with food fraud, such as substituting

species of high economic value with species of low value, untargeted proteomics can thus be a suitable analytical approach. Seafood is susceptible to mislabeling and adulteration (Bouzembrak et al., 2018), and proteomics approaches have been applied previously for the authentication of fish products (Nessen et al., 2016; Wulff et al., 2013). SLM was implemented as a part of this PhD in the authentication and quantification of fish species in mixed samples. For commercially important fish species, few protein sequences are available in the UniprotKB database (Figure 2). However, proteomic database-independent tools such as compareMS2 and SpectraST have been shown to be suitable for detecting fish species (Nessen et al., 2016; Wulff et al., 2013). Therefore, to quantify the fish mixture containing three commercially important fish species, the SpectraST tool was implemented (**Paper II**).

For this work, tandem mass spectra were collected from commercially important fish species. Using compareMS2, the direct comparison of tandem mass spectra from each species was performed for a quality check before building spectra libraries. Tandem mass spectra from the mixture were matched against spectra libraries from seven reference fish species using SpectraST. Based on the calculation, the mixture contained 23% (w/w) of cod, 24% (w/w) tilapia, and 18% (w/w) platyfish (Figure 8). The details of the percentage quantification are given in Table 5, **Paper II**. SLM yielded reliable results when quantifying the relative abundance of distantly related fish species in the mixture. As water and proteins are among the main components of muscle tissue, higher accuracies were observed when calculating the relative contents of species in mixtures using the SLM method (quantifies abundant peptides) compared to the DNA-based method (quantifies less abundant nucleic acids). For example, the calculated percentage of platyfish was accurate regarding the relative amount added to the mixture, while tilapia was slightly underestimated. This discrepancy can be attributed to mixture preparation or water quantities in the mixture (**Paper II**).

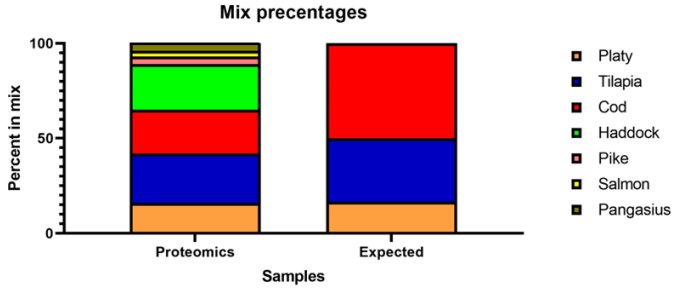


Figure 8: Spectra library matching was implemented to calculate percentages of fish species in the fish mixture. This method can authenticate processed fish products available in the market. The bar graph indicated that cod and haddock were not well separated using this approach.

The SLM method could not differentiate cod (quantified as 23%) from haddock (quantified as 24%), probably due to the close relationship between cod and haddock, belonging to the same family, i.e., gadoids. Due to the close relatedness, cod and haddock contain highly conserved peptides across muscle proteins, i.e., similar spectra, affecting spectra library matching (~27% common spectra). Previously, the conserved peptides from abundant proteins affected calculations in binary mixtures (Ohana et al., 2016). The conserved nature of proteins in closely related species reduces quantification accuracy with the SLM method. In contrast, the DNA-based approach is more accurate when quantifying closely related species (**Paper II**).

A comparison of shotgun DNA sequencing with SLM for the quantification of mixtures proved that both approaches were suitable for fraud detection in fish mixtures. SLM could be used successfully if the species are distantly related; however, SLM had a disadvantage in terms of specificity and false positive results for quantifying closely related fish species in mixtures. In this case, results from DNA-based methods were more accurate (**Paper II**). To resolve challenges concerning the accurate estimation of closely related species in mixed samples using SLM, precalculated conserved peptide overlap between proteomes of two species could be used as a correction factor, as described by Ohana et al. (2016) for mixtures containing horse and cow muscle tissues. However, the availability of this genomic information from species is a prerequisite for calculating the overlap between conserved peptides for two or three species.

Nevertheless, species-specific peptide markers can be identified by collecting high-quality tandem mass spectra from several individuals of the same fish species.

4.3 Authentication and Allergen Detection in Feed and Food-grade Insect Species (Paper III)

Insect species are important in circular economies due to their high nutrient content and feed conversion efficiency (Makkar et al., 2014). However, genomic and proteomic databases for these species contain limited information, which hampers the development of molecular analysis tools (Belghit et al., 2019b; Bose et al., 2021). During this PhD, samples previously described by Belghit et al. (2019b) were reanalyzed using a newly developed method on an instrument set up for proteomics analysis in-house (AF-HPLC HR-MS). The results of the earlier published dataset (Belghit et al., 2019b, MassIVE id: MSV000083737) were compared with newly collected data obtained using AF-HPLC HR-MS.

As mentioned before, compareMS2 compares tandem mass spectra and has been used previously as a quality control tool for proteomic datasets (**Paper V**). To compare the quality of data acquired with two different instruments, compareMS2 GUI was used in **Paper III**. The dendrogram output of compareMS2 (**Paper III**, Figure 1 A and B) showed that all insect spectra acquired using the newly developed AF-HPLC HR-MS method were of high quality and comparable with data obtained previously by Belghit et al. (2019b). A similar HPLC method using standard flow for analyses was developed recently called Standard Flow Multiplexed Proteomics (SFloMPro), generating comparable results to NanoLC (Orsburn et al., 2022). The SFloMPro method used 20 times higher quantities of samples to generate comparable results to NanoLC (Orsburn et al., 2022). Usually, it is possible to acquire high quantities of samples for food and feed analytics. Therefore, methods such as AF-HPLC HR-MS and SFloMPro can easily be implemented in regulatory laboratories for food authentication. Routine mass spectrometry equipment such as AF-HPLC and HR-MS can make the proteomic methods suitable for food authentication laboratories (Sentandreu & Sentandreu, 2011). Instruments such as NanoLC workflows are high-maintenance and increase the

cost per sample. In contrast, workflows such as AF-HPLC HR-MS and SFloMPro can reduce the cost per sample and aid the implementation of proteomics methods in routine and regulatory laboratory settings (Orsburn et al., 2022; Sentandreu & Sentandreu, 2011).

Further analyses of datasets using database-dependent Comet search engine tools identified ~4000 proteins. A comparison of the detected proteins using Venn diagrams revealed that 45% of identified proteins (2758 proteins) were consistently detected in both datasets Belghit et al. (2019b) and AF-HPLC HR-MS (**Paper III**, Figure 3A). Proteomic analyses revealed that for differentiating yellow mealworm and house crickets (edible insect species), larval cuticle protein A2B and cytochrome c oxidase proteins could be used as markers (**Paper III**, Supplementary Figure 4A and B, Supplementary Table S6).

In recent EFSA opinions, requirements for allergenicity risk assessment of insect proteins were laid out for safely introducing edible insects into the food market (EFSA NDA panel 2021a, EFSA NDA panel 2021b, EFSA NDA panel, 2022). Concerns were raised regarding allergens in house crickets and lesser mealworms in risk assessments performed by the EFSA (EFSA NDA panel 2021b, EFSA NDA panel, 2022). Risk assessments assessing potential hazards from allergens took into consideration the proteomic and bioinformatic analyses of allergens in both species. Proteomic data collected in the present study (**Paper III**), AF-HPLC HR-MS and previously by Belghit et al. (2019b), were screened for allergens by matching black soldier fly, yellow mealworm, lesser mealworm, house cricket, and one unauthorized species morio worm samples against a publicly available list of allergen families and sequences (**Paper III**, Supplementary Table S7). Detected allergens in these datasets include tropomyosin, tropomyosin-2, EF-hand proteins, troponin C, and arginine kinase (**Paper III**, Figures 5A and B). Based on these data, MRM assays could be developed to quantify the concentrations of allergens present in insect species investigated.

Allergen detection performed during the study (**Paper III**) showed that the proteomic data can be used for food safety assessment of novel food ingredients. More insect

species are expected to be authorized in the European and global food markets in future, hence identifying potential hazards due to allergens will become necessary. Future risk assessments on edible insect species can use such data to predict the allergenic risks of novel foods.

4.4 Untargeted Proteomics for Differentiation of Soybean Samples (Paper IV)

Omics tools are a promising technology for the risk assessment of GM plants (EFSA, 2022). The application of omics-based assessment strategies is believed to benefit the development of a product-focused safety assessment of crops (Gould et al., 2022). Soybean (*Glycine max*) is an essential component of food and feed products, and GM soybean is widely used in the food market (John et al., 2017). In a study conducted by Bøhn et al. (2014), "substantial non-equivalence" of GM, non-GM and organic soybean samples were evaluated, and residues of glyphosate and aminomethyl phosphonic acid (AMPA) were detected in GM samples. It was argued that herbicide residues in the plants might affect or disturb the plant metabolism in herbicide-tolerant GM varieties (Bøhn et al., 2014). During this PhD, the same 31 samples previously examined by Bøhn et al. (2014) were analyzed using proteomic methods followed by database-independent and database-dependent bioinformatic analyses (**Paper IV**).

A comparative proteomics analysis using compareMS2 indicated that the proteome of all samples was similar, contradicting the previous hypothesis of "substantial non-equivalence" (**Paper IV**, Figure 1). However, further analyses using database-dependent tools did identify a small set of differentially expressed proteins whose expression changes seemed to depend on the cultivation conditions. Pathway analyses of these 39 differentially expressed proteins pointed towards several biological processes which potentially could contribute to the variation in nutrient profiles of GM, non-GM conventionally farmed, and organic soybean samples reported by Bøhn et al. (2014). Previous studies on GM and non-GM soybean parent lines also reported altered biological processes. However, no link to any allergenic or toxic proteins could be

made, and it was concluded that GM soybeans are safe for human consumption (Benevenuto et al., 2021; Jin et al., 2022; Liu et al., 2018, 2020; Natarajan et al., 2020).

Further development of omics technologies for a product-based evaluation of genetically engineered crop varieties and crops developed using novel methods, such as gene editing, will be required (EFSA, 2022; Gould et al., 2022). In addition, for a comprehensive risk assessment and to effectively screen data for unintended molecular and metabolic changes in such novel crops, an increased harmonization of tools and access to open data will be crucial (Benevenuto et al., 2022; Gould et al., 2022).

4.5 compareMS2 2.0 for Food and Feed Safety (Paper V)

In the database-independent workflow of this PhD, compareMS2 (Palmblad & Deelder, 2012) was used to evaluate the quality of samples and data under investigation. The method was initially developed for molecular phylogenetic analyses. However, it has been "repurposed" previously as a quality analyses tool for proteomic datasets (Duivesteijn, 2018; Ohana et al., 2016; Van Der Plas-Duivesteijn et al., 2016; Wulff et al., 2013). For a quick overview of similarities and differences in the proteome of samples, compareMS2 is an excellent tool (Duivesteijn, 2018). In the present PhD, all tandem mass spectra acquired from non-model species were analyzed using compareMS2 for quality assessment before building spectra libraries (**Papers II** and **III**). As compareMS2 does not use any genomic or proteomic information for the classification, it is suitable to analyze food and feed samples. In **Paper III**, for comparison of datasets collected on two different instruments, compareMS2 GUI output was used. Dendrogram outputs of compareMS2 from two datasets helped to assess reproducibility across two platforms (**Paper III**, Figure 1A and B).

Similarly, compareMS2 GUI was also implemented to analyze proteomic data from soybean samples cultivated under different conditions (**Paper IV**, Figure 1). As the compareMS2 dendrogram indicated, the samples were clustered randomly. The result indicated that proteomes of samples in three categories, i.e., GM, conventional, and organically farmed, were similar (**Paper IV**). The compareMS2 output validated

results from database-dependent analyses, where only 39 differentially expressed proteins were detected across three categories of samples (**Paper IV**, Figure 3B). However, samples were always well separated using compareMS2 if considerable differences existed on the proteome level. For example, fish and insect species were separated as per the molecular phylogeny in **Papers II** and **III** and previously published studies (Ohana et al., 2016; Rasinger et al., 2016; Van Der Plas-Duivesteijn et al., 2014; Wulff et al., 2013).

Overall, compareMS2 was found to be a very useful tool for food and feed authentication, either as a quality control tool or for differentiating species and tissue origin of samples (**Papers II**, **III**, and **IV**). compareMS2 software is open-source, capable of assessing the information quickly, and available with additional functionality as recently introduced GUI (**Paper V**).

4.6 FAIR data Practices for Regulatory Science

The proteomic data generated during this PhD was intended to address food and feed safety challenges (**Papers I**, **II**, **III**, and **IV**). Besides developing SLM for PAP and food authentication, spectra libraries created during this PhD for non-model organisms such as insects and fish were made openly available in accordance with FAIR principles. It will be possible to reproduce these results using the same or newly generated data using these libraries (Summarized in Table 5). Furthermore, the proteomic data generated in the present PhD was interoperable (i.e., it was stored and processed in open data formats using open-source tools), which increases the accessibility, visibility, and reusability of these data. Access to FAIR omics data are believed to be vital for facilitating next-generation risk assessments of novel insect food ingredients and GM plants (EFSA, 2022).

In addition to food and feed safety-focused research and risk assessments, the data created during the PhD can also be used in the future to advance the understanding of the biology of the non-model species (Table 5), namely, cow (*Bos taurus*), black soldier fly larvae (*Hermetia illucens*), yellow mealworm (*Tenebrio molitor*), lesser mealworm

(*Alphitobius diaperinus*), house cricket (*Acheta domesticus*), morio worm (*Zophobas morio*), Atlantic cod (*Gadus morhua*), Atlantic haddock (*Melanogrammus aeglefinus*), Nile tilapia (*Oreochromis niloticus*), Northern pike (*Esox lucius*), Atlantic salmon (*Salmo salar*), platyfish (*Xiphophorus maculatus*), pangasius (*Pangasianodon hypophthalmus*), and soybean (*Glycine max*). The spectra libraries built for insect, plant, and fish species can help advance proteomic research in these species. The comparative proteome analyses performed using compareMS2 can may benefit the systematics of non-model organisms in evolutionary context.

Table 5: Summary of all the data generated from food and feed-relevant non-model organisms during this PhD, along with MassIVE ids

Organisms	Scientific name	MassIVE ids	Paper
Cow	*Bos taurus*	MSV000087026	I
Black soldier fly	*Hermetia illucens*	MSV000083737,MSV000087026, & MSV000088034	I & III
Yellow mealworm	*Tenebrio molitor*	MSV000087026, MSV000088034	III & Belghit et al., 2019b
Lesser mealworm	*Alphitobius diaperinus*	MSV000087026, MSV000088034	III & Belghit et al., 2019b
House cricket	*Acheta domesticus*	MSV000087026, MSV000088034	III & Belghit et al., 2019b
Morio worm	*Zophobas morio*	MSV000087026, MSV000088034	III & Belghit et al., 2019b
Atlantic cod	*Gadus morhua*	MSV000087017	II
Atlantic haddock	*Melanogrammus aeglefinus*	MSV000087017	II
Nile tilapia	*Oreochromis niloticus*	MSV000087017	II
Northern pike	*Esox lucius*	MSV000087017	II
Atlantic salmon	*Salmo salar*	MSV000087017	II
Platyfish	*Xiphophorus maculatus*	MSV000087017	II
Pangasius	*Pangasianodon hypophthalmus*	MSV000087017	II
Soybean	*Glycine max*	MSV000089618	IV

5. Conclusions

The present work developed and implemented proteomic approaches for food and feed safety using database-independent and database-dependent bioinformatic tools.

1. SLM successfully differentiated PAP samples at a tissue and species level (**Papers I** and **III**). The method was found to be robust and accurate in species identification when data from different types of mass-spectrometers was used in **Paper III**.
2. SLM was found suitable for identifying and quantifying fish species in mixed samples (**Paper II**). A comparison of shotgun DNA sequencing with SLM showed that for quantification of closely related species, SLM requires further improvement.
3. Database-dependent proteomics successfully identified protein markers for the differentiation of edible insect species (*Tenebrio molitor* and *Acheta domesticus*). In proteomic data from insects, known allergens were detected, demonstrating that omics data can aid allergenicity risk assessments in the future (**Paper III**).
4. GM and non-GM soybean samples cultivated under conventional and organic conditions were differentiated using proteomics data. The study showed that proteomic data could be used for tracing GM samples in feed and food and for product-based risk assessments of genetically modified plants (**Paper IV**).
5. CompareMS2 was used to compare proteomes for food and feed-relevant species as a quality control tool. Together with SLM, compareMS2 2.0 can be applied widely for sample quality control before species and tissue differentiation of food and feed samples (**Paper V**).
6. Proteomic data generated during this PhD can help to understand the biology of non-model species, including farmed animals, insects, and fish. Following FAIR principles, HR-MS data generated in this PhD (**Papers I**, **II**, **III**, and **IV**) were made available online in public repositories (MassIVE ids: MSV000088034, MSV000087026, MSV000087017, and MSV000089618) and will help advance proteomics research in non-model species.

6. Future Directions

Proteomics methods are increasingly recognized as promising tools to complement current standard techniques for food and feed safety analyses. Proteomic methods developed during this PhD will help advance proteomics in regulatory research and risk assessment; for example, the allergenicity assessment of novel foods, safety assessments of GM crops, and PAP detection (Belghit et al., 2021; Benevenuto et al., 2022; Gould et al., 2022). Data collected during this PhD can be used to develop a web-based platform for identifying and quantifying unknown food or feed samples using SLM. This online platform could help regulatory agencies and researchers monitor food and feed samples and report to authorities if any irregularities are detected. The platform will implement a tool to convert raw data to the required format and later visualize results by matching the "unknown data" to the spectra library. The cross-platform compatibility of the data will be prioritized while building this platform following the FAIR principle. SLM implemented in this PhD for PAP detection can be assessed further, similar to a recent inter-laboratory study for targeted methods (Lecrenier et al., 2021). This testing could confirm the potential of the SLM method to resolve the analytical gaps in the detection and differentiation of PAP in Europe. It also could be assessed if targeted regulatory methods could be combined with untargeted methods to assess if the accuracy and sensitivity of both approaches could be improved.

The data collected during this PhD provided a strong foundation for developing additional targeted assays for the safety assessment of food and feed samples. Proteins and peptides detected in insects and GM soybean samples during this PhD thesis can also be used to develop additional MRM assays for targeted proteomic analyses. Similarly, MRM assay also can be developed to detect allergenic proteins of insects in novel food products and aid the efficient risk assessment of novel food ingredients.

Developing user-friendly and open-source database-dependent and independent tools will be crucial for routinely implementing omics methods in laboratories. The newly developed compareMS2 2.0 used during the present work is an example of user-friendly open-source software. Such tools can advance omics in non-model species.

For example, after sequencing non-model species, protein-coding genes can be annotated using high-resolution proteomic data acquired from different tissues (Kelkar et al., 2014) or for discovering new biomarkers in response to environmental stressors (Eide et al., 2021).

7. References

Aas, T. S., Ytrestøyl, T., & Åsgård, T. (2019). Utilization of feed resources in the production of Atlantic salmon (Salmo salar) in Norway: An update for 2016. *Aquaculture Reports*, *15*(August), 100216. https://doi.org/10.1016/j.aqrep.2019.100216

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics: Abstract: Nature. *Nature*, *422*(6928), 198–207. https://www.nature.com/articles/nature01511

Belghit, I., Varunjikar, M., Lecrenier, M.-C. C., Steinhilber, A. E., Niedzwiecka, A., Wang, Y. V. V., Dieu, M., Azzollini, D., Lie, K., Lock, E.-J. J., Berntssen, M. H. G. H. G., Renard, P., Zagon, J., Fumière, O., van Loon, J. J. A. J. A., Larsen, T., Poetz, O., Braeuning, A., Palmblad, M., & Rasinger, J. D. D. (2021). Future feed control – Tracing banned bovine material in insect meal. *Food Control*, *128*(April), 108183. https://doi.org/10.1016/j.foodcont.2021.108183

Belghit, Ikram, Liland, N. S., Gjesdal, P., Biancarosa, I., Menchetti, E., Li, Y., Waagbø, R., Krogdahl, Å., & Lock, E. J. (2019a). Black soldier fly larvae meal can replace fish meal in diets of sea-water phase Atlantic salmon (Salmo salar). *Aquaculture*, *503*, 609–619. https://doi.org/10.1016/J.AQUACULTURE.2018.12.032

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

Beltramo, C., Riina, M. V., Colussi, S., Campia, V., Maniaci, M. G., Biolatti, C., Trisorio, S., Modesto, P., Peletto, S., & Acutis, P. L. (2017). Validation of a DNA biochip for species identification in food forensic science. *Food Control*, *78*, 366–373. https://doi.org/10.1016/J.FOODCONT.2017.03.006

Benevenuto, R. F., Venter, H. J., Zanatta, C. B., Nodari, R. O., & Agapito-Tenfen, S. Z. (2022). Alterations in genetically modified crops assessed by omics studies: Systematic review and meta-analysis. *Trends in Food Science & Technology*, *120*, 325–337. https://doi.org/10.1016/J.TIFS.2022.01.002

Benevenuto, R. F., Zanatta, C. B., Guerra, M. P., Nodari, R. O., & Agapito-Tenfen, S. Z. (2021). Proteomic Profile of Glyphosate-Resistant Soybean under Combined Herbicide and Drought Stress Conditions. *Plants*, *10*(11), 2381. https://doi.org/10.3390/plants10112381

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 and Chemical Toxicology*, *118*(March), 608–625. https://doi.org/10.1016/j.fct.2018.06.005

Bøhn, T., Cuhra, M., Traavik, T., Sanden, M., Fagan, J., & Primicerio, R. (2014). Compositional differences in soybeans on the market: Glyphosate accumulates in Roundup Ready GM soybeans. *Food Chemistry*, *153*, 207–215. https://doi.org/10.1016/j.foodchem.2013.12.054

Bøhn, Thomas, & Millstone, E. (2019). The Introduction of Thousands of Tonnes of

Glyphosate in the food Chain—An Evaluation of Glyphosate Tolerant Soybeans. *Foods*, *8*(12). https://doi.org/10.3390/foods8120669

Boqvist, S., Söderqvist, K., & Vågsholm, I. (2018). Food safety challenges and One Health within Europe. *Acta Vet Scand*, *60*, 1. https://doi.org/10.1186/s13028-017-0355-3

Bose, U., Broadbent, J. A., Juhász, A., Karnaneedi, S., Johnston, E. B., Stockwell, S., Byrne, K., Limviphuvadh, V., Maurer-Stroh, S., Lopata, A. L., & Colgrave, M. L. (2021). Protein extraction protocols for optimal proteome measurement and arginine kinase quantitation from cricket Acheta domesticus for food safety assessment. *Food Chemistry*, *348*, 129110. https://doi.org/10.1016/J.FOODCHEM.2021.129110

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

Bouzembrak, Yamine, & Marvin, H. J. P. (2016). Prediction of food fraud type using data from Rapid Alert System for Food and Feed (RASFF) and Bayesian network modelling. *Food Control*, *61*, 180–187. https://doi.org/10.1016/j.foodcont.2015.09.026

Bremer, M. G. E. G., Margry, J. C. F., Vaessen, J. C. H., Van Doremalen, M. H., Van Der Palen, J. G. P., Van Kaathoven, G. C., Kemmers-Vonken, E. M., & Van Raamsdonk, L. W. D. (2013). Evaluation of a Commercial ELISA for Detection of Ruminant Processed Animal Proteins in Non-Ruminant Processed Animal Proteins AGRICULTURAL MATERIALS. *Journal of AoaC InternatIonal*, *96*(3). https://doi.org/10.5740/jaoacint.11-556

Burr, G. S., Wolters, W. R., Barrows, F. T., & Hardy, R. W. (2012). Replacing fishmeal with blends of alternative proteins on growth performance of rainbow trout (Oncorhynchus mykiss), and early or late stage juvenile Atlantic salmon (Salmo salar). *Aquaculture*, *334–337*, 110–116. https://doi.org/10.1016/j.aquaculture.2011.12.044

Campos, G. (2019). *Processed animal by-products as sustainable ingredients in diets for European seabass (Dicentrarchus labrax)*. *72*, 2019. https://repositorio-aberto.up.pt/bitstream/10216/122044/2/348367.pdf

Capozzi, F., & Bordoni, A. (2013). Foodomics: A new comprehensive approach to food and nutrition. *Genes and Nutrition*, *8*(1), 1–4. https://doi.org/10.1007/s12263-012-0310-x

Caufield, J. H., Fu, J., Wang, D., Guevara-Gonzalez, V., Wang, W., & Ping, P. (2021). A Second Look at FAIR in Proteomic Investigations. *Journal of Proteome Research*, *20*(5), 2182–2186. https://doi.org/10.1021/acs.jproteome.1c00177

compareMS2 GUI. (2021). *https://github.com/524D/compareMS2*. https://github.com/524D/compareMS2

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, *26*(12), 1367–1372. https://doi.org/10.1038/nbt.1511

Craig, R., & Beavis, R. C. (2004). TANDEM: matching proteins with tandem mass

spectra. *Bioinformatics*, *20*(9), 1466–1467. https://doi.org/10.1093/bioinformatics/bth092

Deconinck, D., Volckaert, F. A. M., Hostens, K., Panicz, R., Eljasik, P., Faria, M., Monteiro, C. S., Robbens, J., & Derycke, S. (2020). A high-quality genetic reference database for European commercial fishes reveals substitution fraud of processed Atlantic cod (Gadus morhua) and common sole (Solea solea) at different steps in the Belgian supply chain. *Food and Chemical Toxicology*, *141*(May), 111417. https://doi.org/10.1016/j.fct.2020.111417

Deutsch, E. W., Lam, H., & Aebersold, R. (2008). Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiological Genomics*, *33*(1), 18–25. https://doi.org/10.1152/physiolgenomics.00298.2007

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

Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, *38*(suppl_2), W64–W70. https://doi.org/10.1093/nar/gkq310

Duivesteijn, S. J. (2018). *Advancing zebrafish models in proteomics*. https://hdl.handle.net/1887/66790

EFSA. (2005). Opinion of the Scientific Panel on biological hazards (BIOHAZ) on the "Quantitative risk assessment of the animal BSE risk posed by meat and bone meal with respect to the residual BSE risk." *EFSA Journal*, *3*(9). https://doi.org/10.2903/J.EFSA.2005.257

EFSA. (2010). Scientific Opinion on applications (EFSA-GMO-RX-40-3-2[8-1a/20-1a], EFSA-GMO-RX-40-3-2) for renewal of authorisation for the continued marketing of (1) food containing, consisting of, or produced from genetically modified soybean 40-3-2; (2) feed containi. *EFSA Journal*, *8*(12). https://doi.org/10.2903/j.efsa.2010.1908

EFSA. (2018a). Updated quantitative risk assessment (QRA) of the BSE risk posed by processed animal protein (PAP). *EFSA Journal*, *16*(7). https://doi.org/10.2903/J.EFSA.2018.5314

EFSA. (2018b). Aguilera, J., Aguilera-Gomez, M., Barrucci, F., Cocconcelli, P. S., Davies, H., Denslow, N., Lou Dorne, J., Grohmann, L., Herman, L., Hogstrand, C., Kass, G. E. N., Kille, P., Kleter, G., Nogué, F., Plant, N. J., Ramon, M., Schoonjans, R., Waigmann, E., Wright, M. C., &. EFSA Scientific Colloquium 24 – 'omics in risk assessment: state of the art and next steps. *EFSA Supporting Publications*, *15*(11). https://doi.org/10.2903/sp.efsa.2018.EN-1512

EFSA. (2022). Theme (Concept) paper - Application of OMICS and BIOINFORMATICS Approaches: Towards Next Generation Risk Assessment. *Iacono, Giovanni Guerra, Beatriz Kass, Georges Paraskevopoulos, Konstantinos Kleiner, Juliane Heppner, Claudia Hugas, MartaEFSA Supporting Publications*, *19*(5). https://doi.org/10.2903/sp.efsa.2022.e200506

EFSA NDA panel 2021a. Safety of dried yellow mealworm (Tenebrio molitor larva) as a novel food pursuant to Regulation (EU) 2015/2283. *EFSA Journal*, *19*(1), 1–29. https://doi.org/10.2903/j.efsa.2021.6343

EFSA NDA panel 2021b. Safety of frozen and dried formulations from whole house crickets (Acheta domesticus) as a Novel food pursuant to Regulation (EU) 2015/2283. *EFSA Journal*, *19*(8). https://doi.org/10.2903/j.efsa.2021.6779

EFSA NDA panel 2022. (2022). Safety of frozen and freeze-dried formulations of the lesser mealworm (Alphitobius diaperinus larva) as a Novel food pursuant to Regulation (EU) 2015/2283. *EFSA Journal, Turck, Dominique Bohn, Torsten Castenmiller, Jacqueline De Henauw, Stefaan Hirsch-Ernst, Karen Ildico Maciuk, Alexandre Mangelsdorf, Inge McArdle, Harry J. Naska, Androniki Pelaez, Carmen Pentieva, Kristina Siani, Alfonso Thies, Frank Tsabou*, *20*(7). https://doi.org/10.2903/j.efsa.2022.7325

Ellis, D. I., Muhamadali, H., Allen, D. P., Elliott, C. T., & Goodacre, R. (2016). A flavour of omics approaches for the detection of food fraud. *Current Opinion in Food Science*, *10*(2016), 7–15. https://doi.org/10.1016/j.cofs.2016.07.002

EURL-AP. (2015). EURL-AP Standard Operating Procedure Operational protocols for the combination of light microscopy and PCR. *BENEDETTO, Alessandro BERBEN, Gilbert BROLL, Hermann FUMIÈRE, Olivier FRICK, Geneviève HALDEMANN, Christoph HOUGS, Lotte MÅRTENSSON, Jette PARADIES-SEVERIN, Inge SCHOLTENS, Ingrid VRHOVNIK, Igor UJČIČ VEYS, Pascal*, *32*(0), 5–8. https://www.eurl.craw.eu/wp-content/uploads/2015/12/EURL-AP-SOP.pdf

EURL-AP Standard Operating Procedure. (2022). *EURL-AP Standard Operating Procedure Operational protocols for the combination of light microscopy and PCR*. *32*(0), 5–8. https://www.eurl.craw.eu/wp-content/uploads/2022/05/EURL-AP-SOP-operational-schemes-V5.1.pdf

European Comission 2001/999. (2001). Regulation (EC) No 999/2001 Of the European Parliment and of the Council of 22 May laying down rules for the prevention, control and eradication of certain transmissible spongiform encephalopathies (31/5/2001). *Official Journal of the European Communities*, *May*, 1–69. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32001R0999&from=EN

European Comission (2002). REGULATION (EC) No 1774/2002 of the European Parliament and of the Council of 3 October 2002 laying down health rules concerning animal by-products not intended for human consumption. *248*(1083), 1–89. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02002R1774-20100928&from=EN

European Comission (2003) Regulation (EC) No 2003/1829 of the European parliament and of the council on genetically modified food and feed. *EU regulation 1829*. 1–5. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:2006R1881:20100701:EN:PDF%0Ahttps://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A01985L0374-19990604&qid=1604918047856

European Comission (2013). Commission Regulation (EU) No 56/2013 of 16 January 2013 amending amending Annexes I and IV to Regulation (EC) No 999/2001 of the European Parliament and of the Council laying down rules for the prevention, control and eradication of certain transmissible. *Official Journal of*

the European Union, *L*(21), 3–16. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R0056&from=EN

European Commission. (2013). COMMISSION REGULATION (EU) No 51/2013of 16 January 2013amending Regulation (EC) No 152/2009 as regards the methods of analysis for the determination of constituents of animal origin for the official control of feed. *Official Journal of the European Union*, *2013*(L 20/33), 33–43. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R0051&from=EN

European Commission. (2017). Commission Regulation (EU) 2017/893. amending Annexes I and IV to Regulation (EC) No 999/2001 of the European Parliament and of the Council and Annexes X, XIV and XV to Commission Regulation (EU) No 142/2011 as regards the provisions on processed animal protein. *Official Journal of the European Union*, *2017*(May), 1–25. https://publications.europa.eu/en/publication-detail/-/publication/7d1c640d-62d8-11e7-b2f2-01aa75ed71a1/language-en%0Ahttps://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A32017R0893

European Commission Regulation (2017) Commission Regulation 2017/1017 of 15 June 2017 amending Regulation (EU) No 68/2013 on the Catalogue of feed materials https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R1017&from=EN

European Comission (2021). Commission Regulation (Eu) 2021/1925. amending certain Annexes to Regulation (EU) No 142/2011 as regards the requirements for placing on the market of certain insect products and the adaptation of a containment method. *Official Journal of the European Union*, *1925*(1069), 4–8. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R1925&from=EN

European Comission (2021) Commission Regulation 2021/1372 of 17 August 2021 amending Annex IV to Regulation (EC) No 999/2001 of the European Parliament and of the Council as regards the prohibition to feed non-ruminant farmed animals, other than fur animals, with derived from animals. *Official Journal of the European Union*, *64*, 1–21. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R1372&from=EN

European Comission. (2021). Commission Implementing Regulation (EU) 2021/882 of 1 June 2021 authorising the placing on the market of dried Tenebrio molitor larva as a novel food under Regulation (EU) 2015/2283 of the European Parliament and of the Council, and amending Commission Im. *Official Journal of the European Union*, *64*(L194), 16–20. http://data.europa.eu/eli/reg_impl/2021/882/oj

European Commission. (2022). Commission Regulation 2022/188 Authorising the placing on the market of frozen, dried and powder forms of Acheta domesticus as a novel food under Regulation (EU) 2015/2283 of the European Parliament and of the Council, and amending Commission Implementing Regulation (EU) 2017/2470. *Official Journal of the European Union*, *188*(30), 108–114.

FAO. (2018). *World Fisheries and Aquaculture 2018*. https://www.fao.org/3/i9540en/i9540en.pdf

FAO. (2022). Food security and nutrition in the world repurposing food and healthy

diets more affordable. https://doi.org/10.4060/cc0639en

FEFAC annual Report. (2018). The EU Protein Gap. https://www.gmoinfo.eu/uk/files/353-eu-protein-gap-wcover-06-08.pdf

Flynn, K., Villarreal, B. P., Barranco, A., Belc, N., Björnsdóttir, B., Fusco, V., Rainieri, S., Smaradóttir, S. E., Smeu, I., Teixeira, P., & Jörundsdóttir, H. Ó. (2019). An introduction to current food safety needs. *Trends in Food Science & Technology*, *84*, 1–3. https://doi.org/10.1016/J.TIFS.2018.09.012

Frank, A. M., Monroe, M. E., Shah, A. R., Carver, J. J., Bandeira, N., Moore, R. J., Anderson, G. A., Smith, R. D., & Pevzner, P. A. (2011). Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nature Methods*, *8*(7), 587–591. https://doi.org/10.1038/nmeth.1609

Fumière, O., Dubois, M., Baeten, V., Von Holst, C., & Berben, G. (2006). Effective PCR detection of animal species in highly processed animal byproducts and compound feeds. *Analytical and Bioanalytical Chemistry*, *385*(6), 1045–1054. https://doi.org/10.1007/s00216-006-0533-z

Ghisellini, P., Cialani, C., & Ulgiati, S. (2016). A review on circular economy: The expected transition to a balanced interplay of environmental and economic systems. *Journal of Cleaner Production*, *114*, 11–32. https://doi.org/10.1016/j.jclepro.2015.09.007

Gierlinski, M., Gastaldello, F., Cole, C., & Barton, G. (2018). Proteus : an R package for downstream analysis of MaxQuant output. *BioRxiv*, 416511. https://doi.org/10.1101/416511

Gillund, F., & Myhr, A. I. (2010). Perspectives on Salmon Feed: A Deliberative Assessment of Several Alternative Feed Resources. *Journal of Agricultural and Environmental Ethics*, *23*(6), 527–550. https://doi.org/10.1007/s10806-010-9237-7

Gossner, C. M. E., Schlundt, J., Embarek, P. Ben, Hird, S., Lo-Fo-Wong, D., Beltran, J. J. O., Teoh, K. N., & Tritscher, A. (2009). The melamine incident: Implications for international food and feed safety. *Environmental Health Perspectives*, *117*(12), 1803–1808. https://doi.org/10.1289/ehp.0900949

Gould, F., Amasino, R. M., Brossard, D., Buell, C. R., Dixon, R. A., Falck-Zepeda, J. B., Gallo, M. A., Giller, K. E., Glenna, L. L., Griffin, T., Magraw, D., Mallory-Smith, C., Pixley, K. V., Ransom, E. P., Stelly, D. M., & Stewart, C. N. (2022). Toward product-based regulation of crops. *Science*, *377*(6610), 1051–1053. https://doi.org/10.1126/science.abo3034

Heck, M., & Neely, B. A. (2020). Proteomics in Non-model Organisms: A New Analytical Frontier. *Journal of Proteome Research*, *19*(9), 3595–3606. https://doi.org/10.1021/acs.jproteome.0c00448

Herrero, M., Simõ, C., García-Cañas, V., Ibáñez, E., & Cifuentes, A. (2012). Foodomics: MS-based strategies in modern food science and nutrition. *Mass Spectrometry Reviews*, *31*(1), 49–69. https://doi.org/10.1002/mas.20335

Huet, A. C., Charlier, C., Deckers, E., Marbaix, H., Raes, M., Mauro, S., Delahaut, P., & Gillard, N. (2016). Peptidomic Approach to Developing ELISAs for the Determination of Bovine and Porcine Processed Animal Proteins in Feed for Farmed Animals. *Journal of Agricultural and Food Chemistry*, *64*(47), 9099–9106. https://doi.org/10.1021/acs.jafc.6b03441

IPIFF. (2021). *An overview of the European market of insects as feed*. 2020–2021. https://ipiff.org/

IPIFF. (2022). *Insects as novel foods – an overview*. ipiff.org/insects-novel-food-eu-legislation-2

Jin, L., Wang, D., Mu, Y., Guo, Y., Lin, Y., Qiu, L., & Pan, Y. (2022). Proteomics analysis reveals that foreign cp4-epsps gene regulates the levels of shikimate and branched pathways in genetically modified soybean line H06-698. *GM Crops & Food* https://doi.org/10.1080/21645698.2021.2000320

Jones, A. R., Deutsch, E. W., & Vizcaíno, J. A. (2022). Is DIA proteomics data FAIR? Current data sharing practices, available bioinformatics infrastructure and recommendations for the future. *PROTEOMICS*, 2200014. https://doi.org/10.1002/pmic.202200014

Kalli, A., Smith, G. T., Sweredoski, M. J., & Hess, S. (2014). Evaluation and optimization of mass spectrometric mode: Focus on LTQ-Orbitrap Mass analyzers. *Journal of Proteome Research*, *12*(7), 3071–3086. https://doi.org/10.1021/pr3011588

Kalmykova, Y., Sadagopan, M., & Rosado, L. (2018). Circular economy - From review of theories and practices to development of implementation tools. *Resources, Conservation and Recycling*, *135*(October 2017), 190–201. https://doi.org/10.1016/j.resconrec.2017.10.034

Keane, T. M., O'Donovan, C., & Vizcaíno, J. A. (2021). The growing need for controlled data access models in clinical proteomics and metabolomics. *Nature Communications*, *12*(1), 5787. https://doi.org/10.1038/s41467-021-26110-4

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

Lähteenmäki-Uutela, A., Rahikainen, M., Lonkila, A., & Yang, B. (2021). Alternative proteins and EU food law. *Food Control*, *130*, 108336. https://doi.org/10.1016/J.FOODCONT.2021.108336

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

Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E., & Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. *Nature Methods*, *5*(10), 873–875. https://doi.org/10.1038/nmeth.1254

Lavelli, V. (2021). Circular food supply chains – Impact on value addition and safety. *Trends in Food Science & Technology*, *114*, 323–332. https://doi.org/10.1016/J.TIFS.2021.06.008

Lecrenier, M.-C., 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). Inter-laboratory study on the detection of bovine processed animal protein in feed by LC-MS/MS-based proteomics. *Food Control*, *125*(November 2020), 107944. https://doi.org/10.1016/j.foodcont.2021.107944

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

Lecrenier, M. C., Planque, M., Dieu, M., Veys, P., Saegerman, C., Gillard, N., & Baeten, V. (2018). A mass spectrometry method for sensitive, specific and simultaneous detection of bovine blood meal, blood products and milk products in compound feed. *Food Chemistry*, *245*(September), 981–988. https://doi.org/10.1016/j.foodchem.2017.11.074

Liu, W., Xu, W., Li, L., Dong, M., Wan, Y., He, X., Huang, K., & Jin, W. (2018). iTRAQ-based quantitative tissue proteomic analysis of differentially expressed proteins (DEPs) in non-transgenic and transgenic soybean seeds. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-35996-y

Liu, W., Zhang, Z., Liu, X., & Jin, W. (2020). iTRAQ-based quantitative proteomic analysis of two transgenic soybean lines and the corresponding non-genetically modified isogenic variety. *The Journal of Biochemistry*, *167*(1), 67–78. https://doi.org/10.1093/jb/mvz081

Lock, E. R., Arsiwalla, T., & Waagbø, R. (2016). Insect larvae meal as an alternative source of nutrients in the diet of Atlantic salmon (Salmo salar) postsmolt. *Aquaculture Nutrition*, *22*(6), 1202–1213. https://doi.org/10.1111/anu.12343

Madichie, N. O., & Yamoah, F. A. (2017). Revisiting the European Horsemeat Scandal: The Role of Power Asymmetry in the Food Supply Chain Crisis. *Thunderbird International Business Review*, *59*(6), 663–675. https://doi.org/10.1002/tie.21841

Maggio, A., van Criekinge, T., & Malingreau, J. P. (2015). Global Food Security 2030 Assessing trends with a view to guiding. Publications Office, 2016 https://doi.org/10.2788/5992

Makkar, H. P. S., Tran, G., Heuzé, V., & Ankers, P. (2014). State-of-the-art on use of insects as animal feed. *Animal Feed Science and Technology*, *197*, 1–33. https://doi.org/10.1016/J.ANIFEEDSCI.2014.07.008

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. https://doi.org/10.1021/acs.jafc.6b00064

Marchis, D., Altomare, A., Gili, M., Ostorero, F., Khadjavi, A., Corona, C., Ru, G., Cappelletti, B., Gianelli, S., Amadeo, F., Rumio, C., Carini, M., Aldini, G., & Casalone, C. (2017). LC-MS/MS Identification of Species-Specific Muscle Peptides in Processed Animal Proteins. *Journal of Agricultural and Food Chemistry*, *65*(48), 10638–10650. https://doi.org/10.1021/acs.jafc.7b04639

Marissen, R., Varunjikar, M. S., Laros, J. F. J., Rasinger, J. D., Neely, B. A., & Palmblad, M. (2022). compareMS2 2.0: An Improved Software for Comparing

Tandem Mass Spectrometry Datasets. *Journal of Proteome Research*, 0–5. https://doi.org/10.1021/acs.jproteome.2c00457

Mc Carthy, U., Uysal, I., Badia-Melis, R., Mercier, S., O'Donnell, C., & Ktenioudaki, A. (2018). Global food security – Issues, challenges and technological solutions. *Trends in Food Science & Technology*, *77*, 11–20. https://doi.org/10.1016/J.TIFS.2018.05.002

McEvoy, J. D. G. (2016). Emerging food safety issues: An EU perspective. *Drug Testing and Analysis*, *8*(5–6), 511–520. https://doi.org/10.1002/dta.2015

Meeker, D. L., & Hamilton, C. R. (2006). An overview of the rendering industry. *Essential Rendering*, *September*, 1–16.

Mirabella, N., Castellani, V., & Sala, S. (2014). Current options for the valorization of food manufacturing waste: a review. *Journal of Cleaner Production*, *65*, 28–41. https://doi.org/10.1016/J.JCLEPRO.2013.10.051

Natarajan, S., Islam, N., & Krishnan, H. B. (2020). Proteomic profiling of fast neutron-induced soybean mutant unveiled pathways associated with increased seed protein content. *Journal of Proteome Research*, *19*(10), 3936–3944. https://doi.org/10.1021/acs.jproteome.0c00160

Naylor, R., & Burke, M. (2005). AQUACULTURE AND OCEAN RESOURCES: Raising Tigers of the Sea. *Annual Review of Environment and Resources*, *30*(1), 185–218. https://doi.org/10.1146/annurev.energy.30.081804.121034

Neely, B. A., & Palmblad, M. (2021). Rewinding the Molecular Clock: Looking at Pioneering Molecular Phylogenetics Experiments in the Light of Proteomics. *Journal of Proteome Research*, acs.jproteome.1c00528. https://doi.org/10.1021/ACS.JPROTEOME.1C00528

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

Nesvizhskii, A. I., Keller, A., Kolker, E., & Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, *75*(17), 4646–4658. https://doi.org/10.1021/ac0341261

Niedzwiecka, A., Boucharef, L., Hahn, S., Zarske, M., Steinhilber, A., Poetz, O., Zagon, J., Seidler, T., Braeuning, A., & Lampen, A. (2019). A novel antibody-based enrichment and mass spectrometry approach for the detection of species-specific blood peptides in feed matrices. *Food Control*, *98*(November 2018), 141–149. https://doi.org/10.1016/j.foodcont.2018.11.036

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

Olsen, J. V., Ong, S. E., & Mann, M. (2004). Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Molecular & Cellular Proteomics*, *3*(6), 608–614. https://doi.org/10.1074/MCP.T400003-MCP200

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

Önder, Ö., Shao, W., Kemps, B. D., Lam, H., & Brisson, D. (2013). Identifying sources of tick blood meals using unidentified tandem mass spectral libraries. *Nature Communications*, *4*, 1–10. https://doi.org/10.1038/ncomms2730

Orsburn, B. C., Miller, S. D., & Jenkins, C. J. (2022). Standard Flow Multiplexed Proteomics (SFloMPro)—An Accessible Alternative to NanoFlow Based Shotgun Proteomics. *Proteomes*, *10*(1), 3. https://doi.org/10.3390/proteomes10010003

Pali-Schöll, I., Meinlschmidt, P., Larenas-Linnemann, D., Purschke, B., Hofstetter, G., Rodríguez-Monroy, F. A., Einhorn, L., Mothes-Luksch, N., Jensen-Jarolim, E., & Jäger, H. (2019a). Edible insects: Cross-recognition of IgE from crustacean- and house dust mite allergic patients, and reduction of allergenicity by food processing. *World Allergy Organization Journal*, *12*(1), 100006. https://doi.org/10.1016/j.waojou.2018.10.001

Pali-Schöll, I., Verhoeckx, K., Mafra, I., Bavaro, S. L., Clare Mills, E. N., & Monaci, L. (2019b). Allergenic and novel food proteins: State of the art and challenges in the allergenicity assessment. *Trends in Food Science & Technology*, *84*, 45–48. https://doi.org/10.1016/J.TIFS.2018.03.007

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

Pardo, M. Á., Jiménez, E., Viðarsson, J. R., Ólafsson, K., Ólafsdóttir, G., Daníelsdóttir, A. K., & Pérez-Villareal, B. (2018). DNA barcoding revealing mislabeling of seafood in European mass caterings. *Food Control*, *92*, 7–16. https://doi.org/10.1016/J.FOODCONT.2018.04.044

Perez-Riverol, Y., Wang, R., Hermjakob, H., Müller, M., Vesada, V., & Vizcaíno, J. A. (2014). Open source libraries and frameworks for mass spectrometry based proteomics: A developer's perspective. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1844*(1), 63–76. https://doi.org/10.1016/J.BBAPAP.2013.02.032

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

Ribeiro, J. C., Sousa-Pinto, B., Fonseca, J., Fonseca, S. C., & Cunha, L. M. (2021). Edible insects and food safety: allergy. *Journal of Insects as Food and Feed*, *7*(5), 833–847. https://doi.org/10.3920/jiff2020.0065

Saadat, S., Pandya, H., Dey, A., & Rawtani, D. (2022). Food forensics: Techniques for authenticity determination of food products. *Forensic Science International*, *333*, 111243. https://doi.org/10.1016/J.FORSCIINT.2022.111243

SDGs. (2022). The sustainable development goals report 2022. *United Nations Publication Issued by the Department of Economic and Social Affairs*, 64.

Sentandreu, M. A., & Sentandreu, E. (2011). Peptide biomarkers as a way to determine meat authenticity. *Meat Science*, *89*(3), 280–285. https://doi.org/10.1016/J.MEATSCI.2011.04.028

Silva, J. C. (2018). Food Forensics: Using Mass Spectrometry To Detect Foodborne Protein Contaminants, as Exemplified by Shiga Toxin Variants and Prion Strains. *Journal of Agricultural and Food Chemistry*, *66*(32), 8435–8450. https://doi.org/10.1021/acs.jafc.8b01517

Singh Malhi, G., Kaur, M., & Kaushik, P. (2021). *sustainability Impact of Climate Change on Agriculture and Its Mitigation Strategies: A Review*. https://doi.org/10.3390/su13031318

Steinhilber, A. E., Schmidt, F. F., Naboulsi, W., Planatscher, H., Niedzwiecka, A., Zagon, J., Braeuning, A., Lampen, A., Joos, T. O., & Poetz, O. (2018a). Mass Spectrometry-Based Immunoassay for the Quantification of Banned Ruminant Processed Animal Proteins in Vegetal Feeds. *Analytical Chemistry*, *90*(6), 4135–4143. https://doi.org/10.1021/acs.analchem.8b00120

Steinhilber, A. E., Schmidt, F. F., Naboulsi, W., Planatscher, H., Niedzwiecka, A., Zagon, J., Braeuning, A., Lampen, A., Joos, T. O., & Poetz, O. (2018b). 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

Steinhilber, A. E., Schmidt, F. F., Naboulsi, W., Planatscher, H., Niedzwiecka, A., Zagon, J., Braeuning, A., Lampen, A., Joos, T. O., & Poetz, O. (2019). Application of mass spectrometry-based immunoassays for the species- A nd tissue-specific quantification of banned processed animal proteins in feeds. *Analytical Chemistry*, *91*(6), 3902–3911. https://doi.org/10.1021/acs.analchem.8b04652

Toxqui Rodríguez, M. del S., Vanhollebeke, J., & Derycke, S. (2023). Evaluation of DNA metabarcoding using Oxford Nanopore sequencing for authentication of mixed seafood products. *Food Control*, *145*, 109388. https://doi.org/10.1016/j.foodcont.2022.109388

Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, *11*(12), 2301–2319. https://doi.org/10.1038/nprot.2016.136

United Nations (2022) *World Population Prospects 2022 https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf*

Vågsholm, I., Arzoomand, N. S., & Boqvist, S. (2020). Food Security, Safety, and Sustainability—Getting the Trade-Offs Right. *Frontiers in Sustainable Food Systems*, *4*, 16. https://doi.org/10.3389/fsufs.2020.00016

Van Der Plas-Duivesteijn, S. J., Klychnikov, O., Ohana, D., Dalebout, H., Van Veelen, P. A., De Keijzer, J., Nessen, M. A., Van Der Burgt, Y. E. M., Deelder, A. M., & Palmblad, M. (2016). *Differentiating samples and experimental protocols by direct comparison of tandem mass spectra*. https://doi.org/10.1002/rcm.7494

Van Der Plas-Duivesteijn, S. J., Mohammed, Y., Dalebout, H., Meijer, A., Botermans, A., Hoogendijk, J. L., Henneman, A. A., Deelder, A. M., Spaink, H. P., & Palmblad, M. (2014). Identifying proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *Journal of*

Proteome Research, *13*(3), 1537–1544. https://doi.org/10.1021/pr4010585

van der Spiegel, M., Noordam, M. Y., & van der Fels-Klerx, H. J. (2013). Safety of Novel Protein Sources (Insects, Microalgae, Seaweed, Duckweed, and Rapeseed) and Legislative Aspects for Their Application in Food and Feed Production. *Comprehensive Reviews in Food Science and Food Safety*, *12*(6), 662–678. https://doi.org/10.1111/1541-4337.12032

van Huis, A. (2020). Insects as food and feed, a new emerging agricultural sector: a review. *Journal of Insects as Food and Feed*, *6*(1), 27–44. https://doi.org/10.3920/JIFF2019.0017

van Raamsdonk, L. W. D. D., Prins, T. W., van de Rhee, N., Vliege, J. J. M. M., & Pinckaers, V. G. Z. Z. (2017). Microscopic recognition and identification of fish meal in compound feeds. *Food Additives & Contaminants: Part A*, *34*(8), 1364–1376. https://doi.org/10.1080/19440049.2017.1283711

van Raamsdonk, L.W.D., von Holst, C., Baeten, V., Berben, G., Boix, A., & de Jong, J. (2007). New developments in the detection and identification of processed animal proteins in feeds. *Animal Feed Science and Technology*, *133*(1–2), 63–83. https://doi.org/10.1016/j.anifeedsci.2006.08.004

van Raamsdonk, Leo W. D., Prins, T. W., Meijer, N., Scholtens, I. M. J., Bremer, M. G. E. G., & de Jong, J. (2019). Bridging legal requirements and analytical methods: a review of monitoring opportunities of animal proteins in feed. *Food Additives & Contaminants: Part A*, *36*(1), 46–73. https://doi.org/10.1080/19440049.2018.1543956

van Raamsdonk LWD, Scholtens IMJ, Ossenkoppele J, van Egmond H, G. M., Ossenkoppele, J. S., Egmond, H. J. Van, & Groot, M. J. (2011). Investigation into blood plasma in milk formula. *Wageningen: WUR, Report RIKILT*. https://edepot.wur.nl/184869

Varunjikar, M. S., Belghit, I., Gjerde, J., Palmblad, M., Oveland, E., & Rasinger, J. D. (2022). Shotgun proteomics approaches for authentication, biological analyses, and allergen detection in feed and food-grade insect species. *Food Control*, *137*(December 2021), 108888. https://doi.org/10.1016/J.FOODCONT.2022.108888

Visciano, P., & Schirone, M. (2021). Food frauds: Global incidents and misleading situations. *Trends in Food Science & Technology*, *114*, 424–442. https://doi.org/10.1016/J.TIFS.2021.06.010

VKM. (2022). *KM; Lene Frost Andersen; Paula Berstad; Barbara Bukhvalova; Monica; Carlsen; Lisbeth Dahl; Anders Goksøyr; Lea Sletting Jakobsen; Helle Katrine Knutsen; Ingrid; Kvestad; Inger Therese Lillegaard; Bente Mangschou; Haakon Meyer; Christine Louise Parr; Kirst*. VKM Report 2022:17, ISBN: 978-82-8259-392-2, ISSN: 2535-4019. Norwegian Scientific Committee for Food and Environment (VKM), Oslo, Norway. https://vkm.no/download/18.7ef5d6ea181166b6bb6a110c/1654589000550/Benefit%20and%20risk%20assessment%20of%20fish%20in%20the%20Norwegian%20diet%207.6.22.pdf

Walloon Agricultural Research Centre. (2014). *EURL-AP PCR Proficiency Test 2014 - Final version* (Issue September). https://www.eurl.craw.eu/wp-content/uploads/2019/10/eurl_ap_pcr_pt_2014_final_version.pdf

Welker, F., Collins, M. J., Thomas, J. A., Wadsley, M., Brace, S., Cappellini, E., Turvey, S. T., Reguero, M., Gelfo, J. N., Kramarz, A., Burger, J., Thomas-Oates, J., Ashford, D. A., Ashton, P. D., Rowsell, K., Porter, D. M., Kessler, B., Fischer, R., Baessmann, C., … MacPhee, R. D. E. (2015). Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature*, *522*(7554), 81–84. https://doi.org/10.1038/nature14249

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

Woodgate, S. L., Wan, A. H. L., Hartnett, F., Wilkinson, R. G., Simon, |, & Davies, J. (2022). *The utilisation of European processed animal proteins as safe, sustainable and circular ingredients for global aquafeeds*. https://doi.org/10.1111/raq.12663

WRI. (2020). *World resources report 2013-14 : interim findings Creating a Sustainable Food Future*. https://www.wri.org/research/creating-sustainable-food-future-interim-findings

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

Yocum, A. K., & Chinnaiyan, A. M. (2009). Current affairs in quantitative targeted proteomics: Multiple reaction monitoring - Mass spectrometry. *Briefings in Functional Genomics and Proteomics*, *8*(2), 145–157. https://doi.org/10.1093/bfgp/eln056

Zortea, R. B., Maciel, V. G., & Passuello, A. (2018). Sustainability assessment of soybean production in Southern Brazil: A life cycle approach. *Sustainable Production and Consumption*, *13*, 102–112. https://doi.org/10.1016/J.SPC.2017.11.002

Zuckerkandl, E., Jones, R. T., & Pauling, L. (1960). A Comparison of Animal Hemoglobins by Tryptic Peptide Pattern Analysis. *Proceedings of the National Academy of Sciences*, *46*(10), 1349–1360. https://doi.org/10.1073/pnas.46.10.1349

I

Paper I

Belghit, I., **Varunjikar, M.**, Lecrenier, M.-C. C., Steinhilber, A. E., Niedzwiecka, A., Wang, Y. V. V., Dieu, M., Azzollini, D., Lie, K., Lock, E.-J. J., Berntssen, M. H. G. H. G., Renard, P., Zagon, J., Fumière, O., van Loon, J. J. A. J. A., Larsen, T., Poetz, O., Braeuning, A., Palmblad, M., & Rasinger, J. D.

Future feed control – Tracing banned bovine material in insect meal

Food Control. (2021), 128, 108183

Food Control 128 (2021) 108183

Contents lists available at ScienceDirect

Food Control

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

Future feed control – Tracing banned bovine material in insect meal

I. Belghit [a,*], M. Varunjikar [a], M-C. Lecrenier [b], A. Steinhilber [c], A. Niedzwiecka [d], Y.V. Wang [e], M. Dieu [f], D. Azzollini [g,h], K. Lie [a], E-J. Lock [a], M.H.G. Berntssen [a], P. Renard [f], J. Zagon [d], O. Fumière [b], J.J.A. van Loon [i], T. Larsen [e], O. Poetz [c,j], A. Braeuning [d], M. Palmblad [k], J. D. Rasinger [a,**]

[a] Institute of Marine Research, P.O. Box 1870 Nordnes, 5817, Bergen, Norway
[b] Walloon Agricultural Research Centre (CRA-W), Knowledge and Valorization of Agricultural Products Department, Chaussée de Namur 24, 5030, Gembloux, Belgium
[c] SIGNATOPE GmbH, Reutlingen, 72770, Germany
[d] Department of Food Safety, German Federal Institute for Risk Assessment, Berlin, 10589, Germany
[e] Department of Archaeology, Max Planck Institue for the Science of Human History, Kahlaische Strasse 10, 07745, Jena, Germany
[f] University of Namur, Mass Spectrometry Facility (MaSUN), Rue de Bruxelles 61, B-5000, Namur, Belgium
[g] Food Quality & Design, Agrotechnology and Food Science Group, Wageningen University, Wageningen, the Netherlands
[h] European Food Safety Authority (EFSA), Via Carlo Magno, 1A, 43126, Parma, Italy
[i] Laboratory of Entomology, Plant Sciences Group, Wageningen University, Wageningen, the Netherlands
[j] NMI Natural and Medical Sciences Institute at the University of Tuebingen, Reutlingen, 72770, Germany
[k] Leiden University Medical Center, Leiden, the Netherlands

ARTICLE INFO

Keywords:
Feed control
BSF larvae
Proteomics
Carbon isotope fingerprinting of amino acids
qPCR
Spectral libraries

ABSTRACT

In the present study, we assessed if different legacy and novel molecular analyses approaches can detect and trace prohibited bovine material in insects reared to produce processed animal protein (PAP). Newly hatched black soldier fly (BSF) larvae were fed one of the four diets for seven days; a control feeding medium (Ctl), control feed spiked with bovine hemoglobin powder (BvHb) at 1% (wet weight, w/w) (BvHb 1%, w/w), 5% (BvHb 5%, w/w) and 10% (BvHb 10%, w/w). Another dietary group of BSF larvae, namely *BvHb 10%, was first grown on BvHb 10% (w/w), and after seven days separated from the residual material and placed in another container with control diet for seven additional days. Presence of ruminant material in insect feed and in BSF larvae was assessed in five different laboratories using (*i*) real time-PCR analysis, (*ii*) multi-target ultra-high performance liquid chromatography coupled to tandem mass spectrometry (UHPLC-MS/MS), (*iii*) protein-centric immunoaffinity-LC-MS/MS, *(iv)* peptide-centric immunoaffinity-LC-MS/MS, (*v*) tandem mass spectral library matching (SLM), and (*vi*) compound specific amino acid analysis (CSIA). All methods investigated detected ruminant DNA or BvHb in specific insect feed media and in BSF larvae, respectively. However, each method assessed, displayed distinct shortcomings, which precluded detection of prohibited material versus non-prohibited ruminant material in some instances. Taken together, these findings indicate that detection of prohibited material in the insect-PAP feed chain requires a tiered combined use of complementary molecular analysis approaches. We therefore advocate the use of a combined multi-tier molecular analysis suite for the

Abbreviations: (PAP), Processed Animal Proteins; (BvHb), Bovine Hemoglobin powder; (BSF), Black Soldier Fly; (UHPLC-MS/MS), Multi-target Ultra-High-performance Liquid Chromatography coupled to tandem Mass Spectrometry; (TSE), Transmissible Spongiform Encephalopathies; (BSE), Bovine Spongiform Encephalopathies; (SOP), Standard Operating Procedures; (EURL-AP), European Union Reference Laboratory for Animal Protein; (SLM), Spectral library matching; (ULOQ), Upper limit of quantification; (CSIA), Compound specific stable isotope patterns of amino acids; (AA), Amino acid; (MRM), multiple reaction monitoring; (GC), Gas chromatography; (PCA), Principal component analysis.

* Corresponding author.
** Corresponding author.
E-mail addresses: Ikram.Belghit@hi.no (I. Belghit), madhushri.shrikant.varunjikar@hi.no (M. Varunjikar), m.lecrenier@cra.wallonie.be (M.-C. Lecrenier), steinhilber@signatope.com (A. Steinhilber), Niedzwiecka@bfr.bund.de (A. Niedzwiecka), ywang@shh.mpg.de (Y.V. Wang), marc.dieu@unamur.be (M. Dieu), domenico.azzollini@gmail.com (D. Azzollini), kaikristoffer.lie@hi.no (K. Lie), Erik-Jan.Lock@hi.no (E.-J. Lock), Marc.Berntssen@hi.no (M.H.G. Berntssen), patsy.renard@unamur.be (P. Renard), jutta.zagon@bfr.bund.de (J. Zagon), o.fumiere@cra.wallonie.be (O. Fumière), joop.vanloon@wur.nl (J.J.A. van Loon), larsen@shh.mpg.de (T. Larsen), poetz@signatope.com, poetz@nmi.de (O. Poetz), Braeuning@bfr.bund.de (A. Braeuning), n.m.palmblad@lumc.nl (M. Palmblad), Josef.Rasinger@hi.no (J.D. Rasinger).

https://doi.org/10.1016/j.foodcont.2021.108183
Received 12 March 2021; Received in revised form 15 April 2021; Accepted 17 April 2021
Available online 28 April 2021

detection, differentiation and tracing of prohibited material in insect-PAP based feed chains and endorse ongoing efforts to extend the currently available battery of PAP detection approaches with MS based techniques and possibly $\delta^{13}C_{AA}$ fingerprinting.

1. Introduction

Research on the use of insects as feed ingredients for terrestrial and aquatic animals has developed rapidly in the last five years. By 2017, seven different insect species have been authorized for use in feed for farmed fish (EU Regulation 2017/893). Among these species, black soldier fly (BSF) (*Hermetia illucens*) is considered one of the most relevant species for the production of insect ingredients for fish feed (Belghit, Liland, et al., 2019). The production of BSF larvae yields fish feed ingredients of high nutritive qualities, and offers certain environmental benefits since these production animals have exceptionally fast growth rates, and efficiently convert low-grade organic matter into high-value protein and fat compounds (Ewald et al., 2020; Liland et al., 2017). According to EU regulation 2017/893, insects reared to produce processed animal protein (PAP) are to be considered as farmed animals. Consequently, just like any other farmed animal species in the EU, insects are subject to the same rules established for the prevention of transmissible spongiform encephalopathies (TSE).

In the EU, following an outbreak of bovine spongiform encephalopathies (BSE) in the early 90s, the use of all mammalian-derived proteins in farmed ruminants was banned in 1994. The ban was extended in 2001 to a new regulation, which generally prohibited the use of PAP (except for use in fish meal) and the use of blood products in feed for any farmed animal, respectively (EC, 2001; EC, 2003). In 2013, the EU has set out a progressive working plan for the re-authorization of non-ruminant PAP and blood product in aquafeed (EC, 2011; 2013). This partial re-authorization of PAP gave rise to new regulatory challenges and called for the development and validation of sensitive analytical approaches, which allow for both species and tissue specific differentiation of PAP in feed to differentiate authorized from non-authorized use (Lecrenier et al., 2016; Rasinger et al., 2016).

To guarantee that the use of PAP in feed is in line with current legislation, standard operating procedures (SOP) have been established by the European Union Reference Laboratory for Animal Protein (EURL-AP) for the control of feed stuffs. Optical light microscopy has been the first official method for the detection and characterization of PAP in feed (EC, 2009). However, species-specific identification of PAP is not achievable with microscopy (EC, 2013). This shortcoming led to the development of a second official method, the EURL-AP validated qualitative polymerase chain reaction (qPCR) for ruminant DNA-detection (Fumière et al., 2009; EURL-AP 2013). Even though qPCR is rapid and sensitive, this method is not tissue specific. For example, authorized milk powder cannot be differentiated from prohibited PAP or blood products from the same species (Lecrenier et al., 2020). Therefore, additional approaches have been developed which allow for the determination of both species and tissue specific origin of PAP and blood products in animal feeds (Lecrenier et al., 2018; Marbaix et al., 2016; Rasinger et al., 2016; Steinhilber et al., 2019).

Proteomic-based methods using (tandem) mass spectrometry (MS) were, in a recent scientific opinion by the European Food Safety Authority (EFSA), identified as promising tools to complement current standard techniques of PAP detection in feed (EFSA, 2018). Different laboratories specialized in feed and food safety analyses have been developing complementary MS-based approaches for identification and quantification of peptide markers as protein surrogates for the detection of prohibited PAP and blood products. Among those, targeted MS-methods have been established for detection of bovine specific PAP and blood products as well as permitted ruminant milk products in feed material (at 0.1%, w/w) (Lecrenier et al., 2018; Marchis et al., 2017). The detection of species-specific blood peptides in feed matrices (between 0.05 and 1%, w/w) has also been shown to be useful by applying antibody-based enrichment approaches prior LC-MS/MS read out (Niedzwiecka et al., 2019; Steinhilber et al., 2019). When genomic information is sparse or unavailable, untargeted MS approaches based on direct spectra comparisons and spectral library matching have been used to identify and quantify species and tissue-specific adulteration in food and feed (Belghit, Lock, et al., 2019; Ohana et al., 2016; Rasinger et al., 2016; Wulff, Nielsen, Deelder, Jessen, & Palmblad, 2013).

In addition to proteomic-based tools, the detection of stable carbon isotope patterns of amino acids (AA) (hereafter $\delta^{13}C_{AA}$ fingerprinting), has shown great promise for food and feed authentication (Wang et al., 2018; Wang, Wan, Krogdahl, Johnson, & Larsen, 2019). The $\delta^{13}C_{AA}$ fingerprinting method can trace the biosynthetic origins of proteinogenic amino acids via two different routing mechanisms of their carbon skeletons. While there is little or no changes in the $\delta^{13}C$ values of the essential amino acids during trophic transfer, shifts in $\delta^{13}C$ values for the non-essential AAs can be considerable because animals can synthesize them *de novo* from building blocks derived from dietary macromolecules (McMahon, Fogel, Elsdon, & Thorrold, 2010; McMahon, Polito, Abel, McCarthy, & Thorrold, 2015). Since the $\delta^{13}C_{AA}$ fingerprints reflect diets over a time period that depends on the particular metabolic turnover rate of the analyzed tissue, the method can in theory detect traces of feed material well after the feed sources have changed. This feature makes it highly complementary to our other tested molecular methods that are suited for detecting the most recent diets only.

The aim of this study was to compare the current official method (qPCR) to MS-based approaches and $\delta^{13}C_{AA}$ fingerprinting for detection of prohibited bovine material in BSF larvae that could be used as feed ingredients for farmed fish. BSF larvae were reared on substrate with or without added bovine hemoglobin powder at three different concentrations. Detection of ruminant material in (*i*) the feed media of BSF larvae and in (*ii*) the BSF larvae reared on the adulterated substrate were performed using (*i*) qPCR, (*ii*) multi-target UHPLC-MS/MS, (*iii*) protein-centric immunoaffinity-LC-MS/MS, *(iv)* peptide-centric immunoaffinity-LC-MS/MS, (*v*) tandem mass spectral library matching (SLM) and (*vi*) $\delta^{13}C_{AA}$ fingerprinting technique.

2. Materials and methods

2.1. Feed preparation

The control feeding medium (Ctl) for the BSF larvae consisted of a standard poultry feed (Kasper Faunafood Kuikenopfokmeel 1, Woerden, The Netherlands, 600320), used as a reference feed medium for BSF larvae by the Laboratory of Entomology (Wageningen, The Netherlands). The control feed medium was spiked with bovine hemoglobin powder (BvHb) (92 B, 06000-131-17-0705) at three different concentrations, as follows: (*i*) to 1098 g of ground poultry feed in a sampling bag was added 11.1 g of BvHb, to obtain 1% (w/w) spiked control diets (BvHb 1%), (*ii*) to 1054.5 g of ground poultry feed in a sampling bag was added 55.5 g of BvHb, to obtain 5% (w/w) spiked control diets (BvHb 5%), and (*iii*) to 999 g of ground poultry feed in a sampling bag was added 111 g of BvHb, to obtain 10% (w/w) spiked control diets (BvHb 10%). The design of the experiment is described in Table 1.

2.2. Rearing of BSF larvae and sample preparation

The experiment was carried out at the Laboratory of Entomology (Wageningen, The Netherlands) with seven-day old BSF larvae taken

Table 1
Description of the different feeding media prepared for the black soldier fly larvae growth trial.

Conditions	Ctl	BvHb 1%	BvHb 5%	BvHb 10%	*BvHb 10%
BvHb in medium (%, w/w)	0	1	5	10	10
Total feeding period (days)	7	7	7	7	14

Ctl = control diet, Kasper Faunafood Opfokmeel 1; BvHb = bovine hemoglobin powder. *BvHb 10% = BvHb 10% for 7 days followed by Ctl diet for 7 additional days.

from the stock colony of the Laboratory of Entomology. Experimental units were plastic containers (17.8 × 11.4 × 6.5 cm) to which a homogenized mixture of feed consisting of 18 g of the respective feed media (Ctl, BvHb 1%, BvHb 5% and BvHb 10% (w/w)); 36 mL of water and ~100 BSF larvae were added. The containers were closed with perforated transparent plastic lids to allow for air exchange and were placed in a climate-controlled cabinet (27 ± 1 °C and 80 ± 1% RH). In addition to the four dietary groups (Ctl, BvHb 1%, BvHb 5% and BvHb 10% (w/w)), another dietary group of BSF larvae, namely *BvHb 10%, were first grown on BvHb 10% (w/w) medium, and after seven days separated from the residual material and placed in another container with control diet for seven additional days (decontamination period). At the end of the feeding experiment with a total feeding period of seven days for larvae grown on Ctl, BvHb 1%, BvHb 5%, BvHb 10% (w/w), and a period of 14 days for the decontamination treatment (*BvHb 10% (w/w)), larvae were separated from residual material, rinsed with lukewarm tap water, dried on tissue paper and immediately frozen at – 80 °C. Frozen BSF larvae were ground to a powder using a blender (Braun Multiquick 5 (600 W), Kronberg, Germany) and freeze-dried (freezing for 24 h at −20 °C in vacuum (0.2–0.01 mBar) followed by vacuum at 25 °C until constant weight was reached. Feed media and freeze-dried BSF larvae were divided into different fractions and distributed to different laboratories (laboratories A-E) for the multi-laboratory analyses: (*i*) qPCR (laboratories A and B), (*ii*) multi-target UHPLC-MS/MS (laboratory A), (*iii*) protein-centric immunoaffinity-LC-MS/MS (laboratory B), (*iv*) peptide-centric immunoaffinity-LC-MS/MS (laboratory C), (*v*) direct comparison of tandem mass spectra (laboratory D) and (*v*) $\delta^{13}C_{AA}$ fingerprinting technique (laboratory E). The five dietary groups of BSF larvae were studied in biological duplicates at the five laboratories (n = 2).

2.3. Detection of bovine hemoglobin in the feeding media and in BSF larvae

2.3.1. Real time-PCR (laboratories A and B)

Samples were characterized by real time-PCR according to EURL-AP Standard Operating Procedures 'DNA extraction using the "Wizard® Magnetic DNA purification system for Food" kit' and 'Detection of ruminant DNA in feed using real-time PCR' (https://www.eurl.craw.eu/legal-sources-and-sops/method-of-reference-and-sops/), as laid down in European Commission (EC) Regulation No 152/2009 (Commission, 2009). At laboratory A, PCR were performed on a LightCycler® 480 (Roche Diagnostics GmbH, Rotkreuz, Switzerland). The Ct values were calculated using the "Abs Quant/2nd Derivative max" analysis type of the LightCycler® 480 Software release 1.5.1.62 (Roche Diagnostics GmbH, Rotkreuz, Switzerland). At laboratory B, PCR was performed on a QuantStudio 6 flex thermocycler (ThermoFisher Scientific, Waltham, MA, USA) with automatic baseline setting and a fixed threshold of 0.04 in all experiments. All analyses were done with universal mastermix DMML-D2-D600 from Diagenode (Liège, Belgium). All samples were analyzed in technical duplicates.

2.3.2. Multi-target UHPLC-MS/MS (laboratory A)

A multi-target UHPLC-MS/MS approach was used for the simultaneous detection of targeted ruminant blood and milk proteins. Protocols for protein extraction, digestion, peptide purification and MS analysis were based on the protocol described by Lecrenier et al. (2018) with minor changes. Before extraction, 1 μg of each heavy-labeled concatemers, used as internal standards, were spiked to 1 g of sample. Proteins were extracted in 10 mL of extraction buffer (200 mM TRIS-HCl, pH 9.2, 2 M urea) for 30 min by shaking at 20 °C followed by sonication for 15 min at 4 °C. Tubes were then centrifuged at 4660 g for 10 min at 4 °C and 5 mL of supernatant was transferred into new tubes. The protein extracts were diluted with 5 mL of 200 mM ammonium bicarbonate and reduced with 500 μL of 200 mM DTT at 20 °C for 45 min prior to alkylation with 500 μL of 400 mM IAA for 45 min in the dark at 20 °C. Subsequently, digestion was performed by adding 500 μL of trypsin (1 mg/mL in 50 mM acetic acid) for 1 h at 37 °C and trypsin action was stopped by the addition of 150 μL of 20% (v/v) formic acid in water. Tubes were then centrifuged at 4660 g at 4 °C for 10 min. Peptides were purified by reversed-phase extraction using Sep-Pak tC18 cartridges (Waters – Milford, Massachusetts, USA). Cartridge pre-conditioning was performed with 18 mL acetonitrile followed by equilibration with 18 mL of 0.1% (v/v) formic acid in water. Digested supernatant (10 mL) was loaded on the column. Next, 9 mL of 0.1% (v/v) formic acid in water was used to flush out impurities. Elution was then performed with 5 mL of acetonitrile/0.1% (v/v) formic acid in water 80/20 (v/v). Before evaporation at 45 °C using Centrivap, 15 μL of DMSO was added to each tube to prevent dryness. Finally, the pellets were resuspended in 375 μL of 0.1% (v/v) formic acid in water/-acetonitrile 95/5 (v/v) and centrifuged at 4660 g for 10 min at 4 °C. The supernatants were transferred into a new tube and stored at −20 °C before injection.

Samples were analyzed using a Xevo TQS micro triple quadrupole system with a positive electrospray and multiple reaction monitoring (MRM) mode coupled with an Acquity system (Waters – Milford, Massachusetts, USA). Peptides were separated by reverse-phase liquid chromatography using a C18 Acquity BEH Waters column (2.1 × 100 mm). A gradient (Mobile phase A = 0.1% (v/v) formic acid in water (ULC/MS grade) and mobile phase B = 0.1% (v/v) formic acid in acetonitrile) of 16 min (at 0.2 mL/min) allowed the separation of the peptide biomarkers. Elution was carried out as follows: 0–2 min: 92% A; 2–10 min: 92–58% A; 10–10.10 min: 15% A; 10.10–12.50 min: 15% A; 12.50–12.60 min: 92% A, 12.60–16 min: 92% A. The acquisition and processing of data were carried out by MassLynx software (v. 4.1, Waters). The peptides described in previous studies were selected to be used as biomarkers for the detection of bovine hemoglobin, casein and beta-lactoglobulin (Lecrenier et al., 2018). All samples were extracted and analyzed in technical triplicates.

2.3.3. Protein-centric immunoaffinity LC-MS/MS (laboratory B)

Sample preparation and semiautomatic immunoprecipitation with an antibody raised against bovine hemoglobin for the MS-based immunoassays were previously described by Niedzwiecka et al. (2019) and Steinhilber et al. (2019). For the analysis of insects, some minor changes were made to the protocols. Based on the protocol by Niedzwiecka et al. (2019), a total amount of 1 g was used for sample preparation in 10% trichloroacetic acid and 2% 2-mercaptoethanol in acetone for 2 h at −20 °C. After washing, proteins were extracted using 7 M urea, 2 M thiourea and 12.5 μg/mL α-amylase in water. For semiautomatic immunoprecipitation, the amount of protein extract was changed to 1 mL to increase the maximum amount of hemoglobin available for immunoprecipitation. The samples were then digested with trypsin and analyzed as described in the original publication using a nano-LC-ESI-MS/MS maXis Impact UHR-TOF equipped with a nanoFlow ESI sprayer interface (Bruker, Bremen, Germany) and a 1290 Infinity nano high performance LC (Agilent Technologies, Waldbronn, Germany). LC and MS parameters were used without modifications from the

protocol. All samples were extracted and analyzed in technical duplicates.

2.3.4. Peptide-centric immunoaffinity LC-MS/MS (laboratory C)

The peptide-centric immunoaffinity LC-MS/MS method was a modified version of the method previously published in Steinhilber et al. (2018). Two of the plasma protein markers (SERPINF2 and HP252) were removed from the assay to keep complement (C9) and α-2-macroglobulin (A2M), and the peptide for hemoglobin α-chain (HBA), myosin-7 (MYH7), matrilin-1 (MATN1) and osteopontin (OPN) were added. The chromatographic method was modified by using a faster trapping method (0.15 min at 150 μL/min) and a shorter separation method (8%–50% eluent B in 3.0 min followed by a washing and equilibration step for 2.0 min, 1.5 μL/min flowrate). Peptide separation was performed on an Acclaim Pepmap RSLC C18 (75 μm I.D. × 150 mm, 3 μm, Thermo Fisher Scientific). Mass spectrometric detection was performed using a Sciex QTRAP 6500+ triple quadrupole mass spectrometer operating in MRM mode. All samples were extracted and analyzed in technical duplicates.

2.3.5. Spectral library matching (laboratory D)

Protein extraction, quantification and digestion were performed as described in Belghit, Lock, et al. (2019) and in Rasinger et al. (2016) without any modifications. The protein digest was analyzed by using nano-LC-ESI-MS/MS maXis Impact UHR-TOF (Bruker, Bremen, Germany) coupled with a UPLC Dionex UltiMate 3000 (Thermo). The digests were separated by reverse-phase liquid chromatography using a 1.0 mm × 15 cm reverse phase Thermo column (Acclaim PepMap 100 C18) in an Ultimate 3000 liquid chromatography system. Mobile phase A was 98% of 0.1% formic acid in water and 2% acetonitrile. Mobile phase B was 0.1% formic acid in acetonitrile. The flow rate was 30 μL/min. Mobile phase A was 95% water, 5% acetonitrile, 0.1% formic acid. Mobile phase B was 20% water, 80% acetonitrile, 0.1% formic acid. The digest (10 μl) was injected, and the organic content of the mobile phase was increased linearly from 5% B to 40% in 75 min and from 40% B to 95% B in 10 min. The column effluent was directly connected to the MS. In survey scan, MS spectra were acquired for 0.5 s in the m/z range between 50 and 2200. The 10 most intense peptides ions 2+ or 3+ were sequenced. The collision-induced dissociation (CID) energy was automatically set according to mass to charge (m/z) ratio and charge state of the precursor ion. MaXis and Thermo systems were piloted by Compass HyStar 3.2 (Bruker). Mass spectrometry data generated were converted using DataAnalysis 4.2 (Bruker) and exported as mzML files. Bovine hemoglobin and milk data were searched against the bovine reference proteome obtained from UniProt (UP000009136; accessed on December 2020); insect data was matched against *Hermetia illucens* specific proteins (UniProtKB; accessed on December 2020) using X! Tandem (Craig & Beavis, 2004) as implemented in the Trans-Proteomics Pipeline (TPP) (Deutsch et al., 2015; Ohana et al., 2016). Spectral libraries were created using SpectraST (Version 5.0), as described in Lam (2011), and all sample spectra were searched against their respective spectral libraries for relative quantification of BvHb (Deutsch et al., 2015). Dot products above 0.8 were considered as valid matches and used for quantification. The data used in this study and spectral libraries created are available on MassIVE (ftp://MSV000087026@massive.ucsd.edu). A graphical overview of the SLM workflow and an example output of matched spectra are shown in Supplementary Figures 1 and 2, respectively.

2.3.6. Stable isotope analyses (laboratory E)

The detailed procedure for AA hydrolyses, Gas Chromatogramy (GC) settings, derivatization, carbon correction and data calibration are described in Wang et al. (2018). In short, each sample of about 3 mg was hydrolyzed with 6 N HCl at 110 °C for 20 h before derivatizing the AAs to *N*-acetyl methyl esters following the protocols by Larsen et al. (2013) and Corr, Berstan, and Evershed (2007). The AA derivatives were injected with an autosampler into a InertCap 35 column (60 m, 0.32 mm i.d., 0.50 μm film thickness, GL Sciences) in a GC and then combusted on a Combustion Isotope Ratio Mass Spectrometer (IRMS, Elementar Isoprime visION System, Langenselbold, Germany) at the Max Planck Institute for the Science of Human History, Jena Germany. Isotope data are expressed in delta (δ) notation in per mil (‰) in per mil (‰): δ (‰) = $[(R_{sample}/R_{standard}) - 1] \times 1000$, where R is the ratio of heavy to light isotope. The carbon isotope ratios are expressed relative to the international standards VPDB. Our in-house reference AA-mixture was calibrated against the n-alkane A7 mixture with well-established $\delta^{13}C$ values (available from A. Schimmelmann, Biogeochemical Laboratories, Indiana University). All samples were analyzed in technical triplicates. The average standard deviation for the internal reference standard nor-leucine (Nle) was 0.3‰ (n = 3 for each batch) and the in-house amino acid standards ranged from 0.2‰ for Pro to 0.6‰ for Ala (n = 4–7 for each batch). We obtained the well-defined peaks for the following 15 amino acids: NEAA; alanine (Ala), asparagine/aspartic acid (Asx), glutamine/glutamic acid (Glx), glycine (Gly), proline (Pro), tyrosine (Tyr) and serine (Ser), and EAA; histidine (His), isoleucine (Ile), leucine (Leu), lysine (Lys), methionine (Met), phenylalanine (Phe), threonine (Thr), and valine (Val). We also determined the bulk $\delta^{13}C$ and $\delta^{15}N$ values with the latter expressed relative to AIR. Approximately 1 mg of the dry mass of diets and BSF larvae from each treatment were analyzed in duplicates for bulk carbon and nitrogen isotopes with an EA-IRMS in the Iso Analytical Limited Inc, UK. For quality control, internal lab standards (IA-R068, IA-R038, IA-R069) and a mixture of IAEA-C7 and IA-R-R046) were analyzed in between sample runs. These standards were calibrated against international reference material IAEA–CH–6, IAEA-N-1, IAEA-C-7 for both $\delta^{13}C$ and $\delta^{15}N$. Internal standard yielded 1s = 0.03‰ and 0.03‰ for $\delta^{13}C$ and $\delta^{15}N$ respectively.

3. Results and discussion

In the EU, insects are considered farmed animals, and as such, are subject to the same legal standards as other production animals; this includes rules and regulations concerning the prevention and control of TSE. For efficient control and monitoring of compliance with current feed and food safety regulations, fast and sensitive analytical approaches complementary to the current official methods are required. To the best of our knowledge, this is the first study to compare the suitability of different legacy and novel molecular tools for the detection of prohibited blood products in insect feed and in insect larvae, respectively. The data generated here, shows that each of the six analytical approaches applied, can detect the presence of BvHb in insect feed media and/or in BSF larvae. We also found that each method suffered from some inherent shortcomings in the detection of prohibited material in insect feed and insects; these can however easily be overcome if the tools discussed below are used in unison in tiered PAP-analysis systems.

3.1. Black soldier fly larvae development

In general, adulteration of the feeding media with BvHb at 1%, 5% and 10% (w/w) prepared for the BSF growth trial supported similar larval development as Ctl-fed diets. Despite differences in non-essential $\delta\ ^{13}C_{AA}$ patterns between dietary treatment groups (see Supplementary Table 7), there were no differences in survival (>95%) or growth (mean individual larval body mass *ca.* 180 mg at day 14 of larval development) between BSF larvae fed the control or feed media spiked with BvHb at 1%, 5% and 10% (w/w, data not shown). These results confirm previous findings on the ability of the BSF larvae to grow on adulterated feed media without affecting their survival or growth performance (Bosch, Fels-Klerx, Rijk, & Oonincx, 2017; Camenzuli et al., 2018).

3.2. Detection of bovine hemoglobin powder in the feeding media and in BSF larvae

3.2.1. qPCR

Tables 2 and 3 provide a summary of qPCR results obtained for the detection of prohibited BvHb in the media used for the rearing of BSF larvae and for BSF larvae grown on these media, respectively (Tables 2 and 3). Detailed analysis outputs are presented in Supplementary Table 1. Feeding media adulterated with BvHb at the 1%, 5% and 10% (w/w) level were all correctly identified as positive for ruminant DNA (Table 2). Control feed media, which consisted of a standard poultry feed without BvHb adulteration, also were found to be positive for ruminant DNA by qPCR (laboratories A and B, Table 2 and Supplementary Table 1). As dictated by EU legislation, standard poultry feed, including feed material used in the present study, must not contain ruminant PAP or blood products. The positive result obtained by qPCR thus could indicate the presence of non-permitted ruminant material in control feed media. On the other hand, the positive finding also could be due to the presence of permitted feed ingredients of bovine origin such as milk.

At the lowest level of adulteration (1% (w/w) BvHb, Table 3, Supplementary Table 1) tested in the current study, qPCR performed by laboratory A confirmed the presence of ruminant DNA in BSF larvae. Real-time PCR, which is based on the detection of DNA, allows for amplification of minute amounts of target sequences specific to a species or group of species and in general displays very high sensitivities with respect to its target analytes (Fumière, Dubois, Baeten, von Holst, & Berben, 2006; Olsvik et al., 2017; Tanabe et al., 2007). Therefore, qPCR can detect less than 0.1% (w/w) in mass fraction of PAP or blood products in feed and in feed ingredients, respectively. However, when applying the same official qPCR assay in another laboratory (B), in the insect larvae fed the BvHb 1% (w/w) diet, ruminant DNA was not detected (Table 3, Supplementary Table 1). In cases of trace levels of ruminant DNA contamination, interlaboratory differences for ruminant PAP detection using the EURL-validated qPCR assay have been described before. For example, Olsvik et al. (2017) reports on qPCR data obtained at three different national reference laboratories, which analyzed 19 non-ruminant PAP and compared these data to results obtained using an immunoassay-based method. Ruminant PAP was detected in five out of 19 samples and in accordance with the findings of the present study, methodological and multi-laboratory differences for qPCR assay results were reported (Olsvik et al., 2017). The authors speculated that the observed differences in the results obtained might be due to a shift in the normal distribution of Ct-values close to the cut-off of the PCR assay, PCR inhibition or different process during homogenization and grinding step (Olsvik et al., 2017).

3.2.2. LC-MS/MS-based approaches

Contrary to current legislation on PAP, qPCR does not distinguish between non-authorized and authorized ruminant products such as bovine milk (EFSA, 2018). When tissue specificity is the goal, proteomics approaches can be applied to complement and refine current methods of PAP detection (Rasinger et al., 2016). In 2014, EURL-AP initiated an international laboratory network to investigate and develop alternative techniques for PAP detection including, MS-based techniques, immunoassays or spectroscopic methods to complement current standard analytic approaches (Lecrenier et al., 2020; Van Raamsdonk et al., 2019). MS-based proteomic approaches were listed among the most promising methods for complementing current standard techniques of feed PAP and blood products detection in a report published by EFSA (EFSA, 2018). The potential of MS-based methods for resolving current challenges of official regulatory PAP analyses recently was confirmed in an inter-laboratory study performed across five different European laboratories in which different MS-based protocols for detection of prohibited bovine material in feed samples were compared (Lecrenier et al., 2021). The study concluded that MS-based analyses efficiently identified non-authorized bovine protein in feed sample mixes at an adulteration level of 1% (w/w) (Lecrenier et al., 2021). The finding by Lecrenier et al. (2021) is further corroborated by results obtained in the present work in which four different MS-based analyses protocols were applied to detect BvHb in the insect-PAP feed chain. Two complementary proteomic approaches were used; (*i*) targeted MS with or without the use of stable isotope-labeled standards (laboratories A, B and C) and (*ii*) SLM (laboratory D).

Targeted MS (laboratories A, B and C) positively identified bovine haemoglobin powder in feeding media spiked with 1%, 5% or 10% (w/w) BvHb (Table 2 and Supplementary Tables 2-3). When using non-targeted SLM (laboratory D), a linear increase of bovine specific peptides was observed in the feeding media with increasing concentrations of BvHb (Supplementary Tables 4-5). Multi-target UHPLC-MS/MS (laboratory A), SLM (laboratory D) and peptide-centric immunoaffinity LC-MS/MS (laboratory C) (Table 2 and Supplementary Tables 2-5) detected the presence of bovine hemoglobin also in control feeding media. However, determined abundances of BvHb in Ctl media were very low when compared to feeding media spiked with 1%, 5% or 10% (w/w) BvHb (Supplementary Tables 3-5). For example, using quantitative peptide-centric immunoaffinity LC-MS/MS (laboratory C), in control feed, 19.0 ± 1,3 fmol of BvHb specific peptide, bovine hemoglobine α chain (HBA), were detected, whereas at the 1% (w/w) level of BvHb adulteration, over 15000 fmol of HBA were measured; at 5% and 10% (w/w) BvHb in feed, levels of HBA were above the upper limit of quantification (Supplementary Table 3). As was discussed above, control feeding media consisted of standard poultry feed, which should be free of ruminant PAP or blood, but ruminant DNA was detected in these samples by qPCR (Table 2 and Supplementary Table 1). Since three of the MS datasets obtained also were indicative of the control feeding media being contaminated with bovine hemoglobin, the positive finding of the qPCR analyses could indeed indicate that the poultry feed used as control diet in the present study was indeed contaminated with trace amounts of ruminant blood products or blood meal. In addition to bovine specific blood proteins, bovine plasma proteins were detected by peptide-centric immunoaffinity LC-MS/MS (laboratory C), presumably

Table 2
Detection of ruminant material in the feeding media used for the black soldier fly larvae growth trial.

	qPCR (labs A, B)		Targeted MS (labs A, B, C)								SLM (lab D)	
			LC-MS/MS		IA-LC-MS/MS (protein IP)	IA-LC-MS/MS (peptide IP)						
	Ruminant DNA		Hb	MP[1]	Hb	Hb	PP	MP[2]	MY	CP	Hb	MP
Ctl	+	+	+	+	–	+	–	–	–	–	+	+
BvHb 1%	+	+	+	+	+	+	–	–	–	–	+	+
BvHb 5%	+	+	+	+	+	+	+	-	-	–	+	+
BvHb 10%	+	+	+	+	+	+	+	-	-	–	+	+

Plus sign (+) indicates a positive result; minus sign a (−) negative result. Workflows: LC-MS/MS (laboratory A, triple quadrupole); immunoaffinity-LC-MS/MS (IA-LC-MS/MS), IA on protein level (laboratory B, Q-TOF); IA-LC-MS/MS, IA on peptide level (laboratory C, triple quadrupole); SLM, spectral library matching (laboratory D, Q-TOF). Bovine proteins identified: Hb, hemoglobin; PP, plasma proteins: α2 macroglobulin and complement component 9; MP, milk protein: [1] Beta-lactoglobulin, casein and [2] osteopontin; MY, muscle protein: myosin 7; CP, cartilage protein: matrilin 1. Detailed analysis outputs are presented in Supplementary Tables 1-6.

Table 3
Detection of ruminant material in the BSF larvae grown on feeding media containing bovine hemoglobin powder (n = 2).

	qPCR (labs A, B)		Targeted MS (labs A, B, C)								SLM (lab D)	
			LC-MS/MS		IA-LC-MS/MS (protein IP)	IA-LC-MS/MS (peptide IP)						
	Ruminant DNA		Hb	MP[1]	Hb	Hb	PP	MP[2]	MY	CP	Hb	MP
Ctl	−	−	-	−	−	+	−	−	−	−	−	+
	−	−	-	−	−	+	−	−	−	−	−	+
BvHb 1%	+	−	-	−	−	+	−	−	−	−	−	+
	+	−	-	−	−	+	−	−	−	−	−	+
BvHb 5%	+	−	-	−	+	+	−	−	−	−	+	+
	+	+	+	−	+	+	−	−	−	−	+	+
BvHb 10%	+	+	-	−	+	+	−	−	−	−	+	+
	+	+	-	−	+	+	−	−	−	−	+	+
*BvHb 10%	−	−	-	−	−	+	−	−	−	−	−	+
	−	−	-	−	−	+	−	−	−	−	−	+

Plus sign (+) indicates a positive result; minus sign a (−) negative result. Workflows: LC-MS/MS (laboratory A, triple quadrupole); immunoaffinity-LC-MS/MS (IA-LC-MS/MS), IA on protein level (laboratory B, Q-TOF); IA-LC-MS/MS, IA on peptide level (laboratory C, triple quadrupole); SLM, spectral library matching (laboratory D, Q-TOF). Bovine proteins identified: Hb, hemoglobin; PP, plasma proteins: α2 macroglobulin and complement component 9; MP, milk protein: [1] Beta-lactoglobulin, casein and [2] osteopontin; MY, muscle protein: myosin 7; CP, cartilage protein: matrilin 1. Detailed analysis outputs are presented in Supplementary Tables 1-6.

being plasma residues of the BvHb preparation. All MS-based methods investigated, also positively identified bovine milk peptides in the standard chicken feed, which was used as control feeding media in the present study (β-lactoglobulin, casein or osteopontin Table 2 and Supplementary Tables 2-6).

In the BSF larvae fed control feed media or feed adulterated with BvHb at 1% (w/w) level, only peptide-centric immunoaffinity LC-MS/MS detected the presence of bovine blood (Fig. 1A, Table 3). One reason as to why the remaining MS approaches failed to detect BvHb in the BSF larvae at the 1% (w/w) level might be the lower sensitivity of these methods compared to the immunoaffinity-based approach. Also, the fact that SLM method detected the presence of BvHb in the BSF larvae in a linear manner with increasing concentration of BvHb only at 5% and 10% (w/w) but not at 1% (w/w) (Fig. 1B and Supplementary Table 6) points to a lack of sensitivity of these approaches when compared to the immunoaffinity-based approach. When using multi-target UHPLC-MS/MS method (laboratory A), only one of the two replicate samples of BSF larvae fed diets adulterated with 5% was positive for BvHb (Table 3). These results are probably due to differences in homogeneity and particle size distribution between the two replicate samples. As described earlier, the heterogeneity of the samples can interfere with the correct detection of specific peptide in certain matrices (Marbraix et al., 2016). Taken together, our data indicate that, as with classic PAP, also for detection and differentiation of insect PAP, LC-MS/MS-based proteomics show great potential to resolve current analytical gaps but technical challenges remain to be addressed in the future.

3.2.3. δ $^{13}C_{AA}$ fingerprinting method

In the current study, δ $^{13}C_{AA}$ fingerprinting (laboratory E) detected BvHb contamination in BSF larvae fed 10% (w/w) for one week, when this was followed by a decontamination period during which larvae were fed control diets for an additional week (*BvHb 10%) (Fig. 2A). In addition to δ $^{13}C_{AA}$ fingerprinting, peptide-centric immunoaffinity LC-MS/MS (laboratory C) successfully detected traces of non-permitted bovine blood residues in BSF after decontamination. However, given that control-media used in the present study was found to contain traces of bovine material, it is not clear if positive MS finding in the *BvHb 10% group is result of the background contamination detected in the control diet or if this method indeed is able to detect traces of non-permitted material in larvae after decontamination. The challenge of detecting non-permitted material using MS-based assays could be due to the removal of easily detectable residual exterior BvHb contamination stemming from direct contact of BSF larvae with the 10% (w/w) BvHb diet and frass when placing larvae in clean containers during the decontamination period. In addition, after seven days feeding on Ctl-media, BvHb-exposed larvae may have effectively cleaned their gut of any internal BvHb residues. Actually, before harvesting insect larvae,

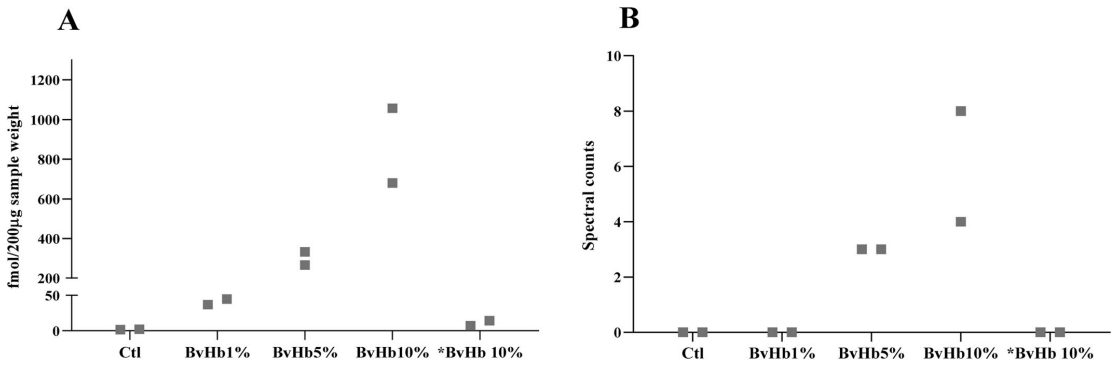


Fig. 1. (A) Quantification of hemoglobin α chain (HBA, fmol absolute/200 µg sample weight, by peptide-centric immunoaffinity LC-MS/MS (laboratory C, Y axis) in the black solider fly larvae fed the control (Ctl) or feed media spiked with BvHb at 1%, 5% and 10% (w/w); BvHb 1%, BvHb 5% and BvHb 10% (w/w), respectively; *BvHb 10%: BvHb 10% for 7 days followed by Ctl diet for 7 additional days (n = 2, X axis). **(B)** Total count of spectra matching against hemoglobin spectral library (laboratory D, Y axis) determined in the black solider fly larvae fed the control (Ctl) or feed media spiked with BvHb at 1%, 5% and 10% (w/w); BvHb 1%, BvHb5% and BvHb10% (w/w), respectively; *BvHb 10%: BvHb 10% for 7 days followed by Ctl diet for 7 additional days (n = 2, X axis).

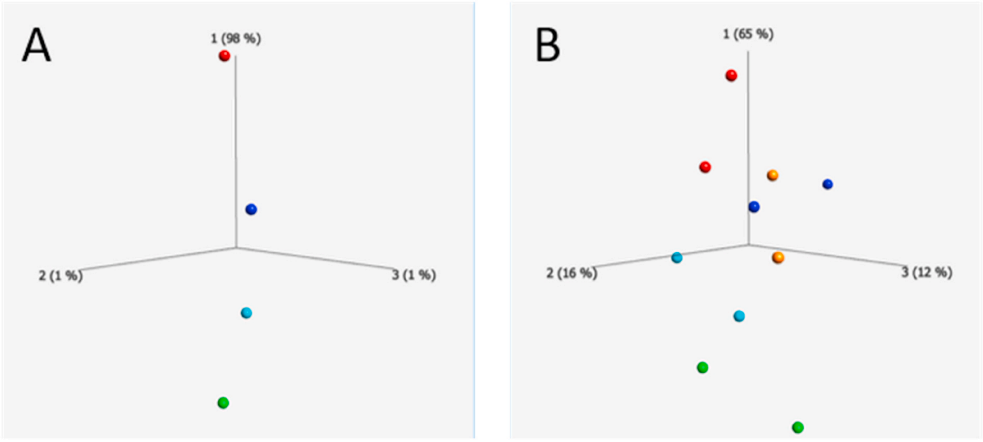


Fig. 2. Detection of bovine hemoglobin powder (BvHb) using $\delta^{13}C_{AA}$ fingerprinting. Principal component analysis (PCA) of **(A)** BvHb in feeding media and **(B)** in black soldier fly (BSF) larvae fed the control (Ctl) or feed media spiked with BvHb at 1%, 5% and 10% (w/w); BvHb 1%, BvHb5% and BvHb10% (w/w), respectively; *BvHb 10%: BvHb 10% for 7 days followed by Ctl diet for 7 additional days (n = 2). PCAs are based on δ13CAA displaying significant correlation ($p < 0.05$) in rank regression analysis in relation to concentrations of BvHb in BSF fed adulterated diets. **(A)** The green, turquoise, blue and red dots represent the control (Ctl), or feed media spiked with BvHb at 1%, 5% and 10% (w/w); BvHb 1%, BvHb 5% and BvHb 10% (w/w), respectively. **(B)** The green, turquoise, blue, red and orange dots represent BSF larvae fed on Ctl, BvHb 1%, BvHb 5%, BvHb10% and *BvHb 10% (w/w), respectively. *BvHb 10%: BvHb 10% for 7 days followed by Ctl diet for 7 additional days. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the inclusion of a starvation period, also called gut purging, of at least 24 h has been recommended, since the gut content of insects was found to contribute considerably to overall contaminant levels and the microbial loads detected in harvested larvae (Bosch et al., 2017; van Huis, 2013). Bosch et al. (2017) showed that feeding yellow mealworm larvae with poultry feed for 2 days after being fed media containing aflatoxin, considerably reduce the content of this mycotoxin in the larvae. In the current study, substitution of adulterated feeding media with clean poultry diets for seven days prior to harvest, thus allowed the larvae to significantly reduce or possibly eliminate any left-over BvHb in the gut.

Despite the hypothesized lack of internal or external BvHb residues present in BSF larvae fed control diets for a week after one-week of BvHb 10% (w/w) exposure, $\delta^{13}C_{AA}$ fingerprints detected differences in non-essential AA composition (Fig. 2, Supplementary Table 7). $\delta^{13}C_{AA}$ values for BSF larvae fed control diets (Ctl) or BvHb 10%* (w/w) were the highest for almost all AA (Fig. S3). Principal component analysis (PCA) of the most discriminative AAs (Ala, Val, Leu, Glx, Phe, Lys and Tyr) (Fig. 2A) display significant correlations ($p < 0.05$) in rank regression analysis in relation to increasing concentrations of BvHb in feeding media (Supplementary Table 7). To discern between BSF larvae fed the different feeding media, Ala, Glx, His, Ile, and Ser were identified as the most discriminative AA that explain the clustering variation (Fig. 2B). The fact we were able to discern between Ctl and the depurated larvae (*BvHb 10%) shows that AAs originating from BvHb proteins had not been replaced completely after seven days on the Ctl diet. This time period is considerably longer than the 100 min required for ingested feed to pass though the digestive system of BSF larvae (Mumcuoglu, Miller, Mumcuoglu, Friger, & Tarshis, 2001). These promising $\delta^{13}C_{AA}$ fingerprinting results warrant further sensitivity tests with depurated larvae.

The data obtained in the present study indicate that $\delta^{13}C_{AA}$ fingerprinting, while less sensitive than LC-MS-based approach discussed above, was able to cluster the BSF larvae fed *BvHb10% together with groups of insects fed BvHb at the 5% and 10% (w/w) level. $\delta^{13}C_{AA}$ fingerprinting has recently been used to address questions of food authenticity in the aquaculture sector, successfully discriminating between wild-caught, organically, and conventionally farmed salmon groups, as well as salmon fed alternative diets such as insects or macroalgae (Wang et al., 2018, 2019). In other words, based on previous studies and the findings presented here, in addition to MS-based approached, $\delta^{13}C_{AA}$ fingerprinting should also be considered for use in a multi-tier molecular analysis toolbox that can efficiently address questions of food authenticity and detect trace amounts of illegal material through the insect-PAP feed chain.

4. Conclusions

The aim of this study was to assess the suitability of legacy and novel molecular analysis tools (i.e. qPCR, MS-based approaches and $\delta^{13}C_{AA}$ fingerprinting) for detection of prohibited bovine material in the food chain when including insect PAP. The data generated here, show that each of the analytical approaches investigated is capable of detecting the presence of BvHb in insect feeding media and/or in BSF larvae. It also was found that each method displayed distinct shortcomings, which precluded detection of prohibited material in some instances. We therefore advocate the use of a combined multi-tier molecular analysis suite for the detection, differentiation and tracing of prohibited material in insect-PAP based feed chains. Taken together, the results confirmed earlier reports on the shortcomings of official monitoring methods and endorse ongoing efforts to extend the currently available battery of PAP detection approaches with MS based techniques and possibly $\delta^{13}C_{AA}$ fingerprinting.

Author contributions

Conceptualization, J.R, I.B, M.B and E-J.L.; Data curation, I.B, J.R, M. V, M-C.L, A.E.S, A.N, Y.V.W, M.D, O.F, J.V, T.L, O.P, A.B and M.P., Formal analysis, I.B, J.R, M.V, M-C.L, A.E.S, A.N, Y.V.W, M.D, O.F, J.V, T.L, O.P, A.B and M.P.,; Investigation, I.B, J.R, M.V, M-C.L, A.E.S, A.N, Y. V.W, M.D, D.A, K.L, M.B, J.Z, O.F, J.V, T.L, O.P, A.B and M.P,; Methodology, I.B, J.R, M.V, M-C.L, A.E.S, A.N, Y.V.W, M.D, E-J.L, D.A, K.L, M. B, J.Z, O.F, J.V, T.L, O.P, A.B and M.P,; Project administration, I.B, E-J.L, J.R and M.B,; Software, M.V, J.R, I.B, M-C.L, A.N, Y.V.W, M.D, O.F, T.L, O.P and M.P,; Writing-original draft, I.B and J.R,; Writing –review &original draft, I.B, J.R, M.V, M-C.L, A.E.S, A.N, Y.V.W, M.D, K.L, M.B, P.R, J,Z, O.F, J.V, T.L, O.P, A.B and M.P.

Declaration of competing interest

O.P. is shareholder of SIGNATOPE GmbH. SIGNATOPE offers assay development and service using immunoaffinity LC-MS/MS technology. D.A is currently employed with the European Food Safety Authority (EFSA) at the Nutrition Unit that provides scientific and administrative support to the NDA panel in the area of safety assessment of novel foods. However, the present article is published under the sole responsibility of the authors and may not be considered as an EFSA scientific output. The positions and opinions presented in this article are those of the author/s alone and are not intended to represent the views/any official position or scientific works of EFSA. To know about the views or scientific outputs of EFSA, please consult its website under http://www.efsa.europa.eu.

Acknowledgments

This study was supported by the Norwegian Research Council project ENTOFÔR, grant number 268344 and the Institute of Marine Research, Bergen (MultiOmicsTools, 15470). A.E.S. and O.P. were supported by funds of German Government's Special Purpose Fund held at Landwirtschaftliche Rentenbank (FKZ 28RZ6IP002).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2021.108183.

Credit author statement

On the behalf of all the co-author's, we affiliate the work performed in the manuscript "Future feed control – Tracing banned bovine material in insect meal" with the different institutions (IMR; CRA-W, Signatope, BfR, Max Planck Institue, University of Namur, WUR University, NMI Natural and Medical Sciences Institute at the University of Tuebingen and Leiden University), we have no conflicts of interest, and declare that we have contributed to the acquisition, analysis and interpretation of data.

References

Belghit, I., Liland, N. S., Gjesdal, P., Biancarosa, I., Menchetti, E., Li, Y., et al. (2019). Black soldier fly larvae meal can replace fish meal in diets of sea-water phase Atlantic salmon (*Salmo salar*). *Aquaculture, 503*, 609–619.

Belghit, I., Lock, E.-J., Fumière, O., Lecrenier, M.-C., Renard, P., Dieu, M., et al. (2019). Species-specific discrimination of insect meals for aquafeeds by direct comparison of tandem mass spectra. *Animals : An Open Access Journal from MDPI, 9*(5), 222.

Bosch, G., Fels-Klerx, H. J.v. d., Rijk, T. C.d., & Oonincx, D. G. A. B. (2017). Aflatoxin B1 tolerance and accumulation in black soldier fly larvae (Hermetia illucens) and yellow mealworms (Tenebrio molitor). *Toxins, 9*(6), 185.

Camenzuli, L., Van Dam, R., de Rijk, T., Andriessen, R., Van Schelt, J., & Van der Fels-Klerx, H. J. I. (2018). Tolerance and excretion of the mycotoxins aflatoxin B_1, zearalenone, deoxynivalenol, and ochratoxin A by Alphitobius diaperinus and *Hermetia illucens* from contaminated substrates. *Toxins, 10*(2), 91.

Commission Regulation (EC). (2001). No 999/2001 of the European Parliament and of the Council of 22 May 2001 laying down rules for the prevention, control and eradication of certain transmissible spongiform encephalopathies. In , *L 147. Official journal of the European union* (pp. 31–35). Office for Official Publications of the European Communities, 2001; p 1– 40.

Commission Regulation (EC). (2003). No. 1234/2003 of 10 July 2003 amending Annexes I, IV and XI to Regulation (EC) No. 999/2001 of the European Parliament and of the Council and Regulation (EC) No. 1326/2001 as regards the transmissible spongiform encephalopathies and animal feeding. In , *L 173. Official journal of the European union* (pp. 11–17). Office for Official Publications of the European Communities, 2003; p 6– 13.

Commission Regulation (EC). (2009). No 152/2009 of 27 January 2009 laying down the methods of sampling and analysis for the official control of feed. *Official Journal of European Union, L54*, 1–130.

Commission Regulation (EU). (2011). *No 142/2011 of 25 February 2011 implementing Regulation (EC) No 1069/2009 of the European Parliament and of the Council laying down health rules as regards animal by-products and derived products not intended for human consumption and implementing Council Directive 97/78/EC as regards certain samples and items exempt from veterinary checks at the border under that Directive Text with EEA relevance*.

Commisson regulation (EU). (2013). No 56/2013 of 16 January 2013 amending Annexes I and IV to Regulation (EC) No 999/2001 of the European Parliament and of the Council laying down rules for the prevention, control and eradication of certain transmissible spongiform encephalopathies. *Official Journal of European Union, L21*, 3–16.

Corr, L. T., Berstan, R., & Evershed, R. P. (2007). Optimisation of derivatisation procedures for the determination of δ13C values of amino acids by gas chromatography/combustion/isotope ratio mass spectrometry. *Rapid Communications in Mass Spectrometry, 21*(23), 3759–3771.

Craig, R., & Beavis, R. C. (2004). Tandem: Matching proteins with tandem mass spectra. *Bioinformatics, 20*(9), 1466–1467.

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.

EFSA, Ricci, A., Allende, A., Bolton, D., Chemaly, M., Davies, R., et al. (2018). Updated quantitative risk assessment (QRA) of the BSE risk posed by processed animal protein (PAP). *EFSA Journal, 16*(7), Article e05314.

European Union (EU). (2017). *Commission regulation (EU) 2017/893 of 24 may 2017 amending annexes I and IV to regulation (EC) No 999/2001 of the European parliament and of the Council and annexes X, XIV and XV to commission regulation (EU) No 142/2011 as regards the provisions on processed animal protein*.

European Union Reference Laboratory for Animal Proteins in feedingstuffs. (2013). EURL-AP Standard Operating Procedure - operational protocols for the combination of light microscopy and PCR. http://eurl.craw.eu/img/page/sops/EURL-AP%20SOP%20operational%20schemes%20V2.0.pdf. (Accessed 20 April 2014).

Ewald, N., Vidakovic, A., Langeland, M., Kiessling, A., Sampels, S., & Lalander, C. (2020). Fatty acid composition of black soldier fly larvae (*Hermetia illucens*) – possibilities and limitations for modification through diet. *Waste Management, 102*, 40–47.

Fumière, O., Dubois, M., Baeten, V., von Holst, C., & Berben, G. (2006). Effective PCR detection of animal species in highly processed animal byproducts and compound feeds. *Analytical and Bioanalytical Chemistry, 385*(6), 1045.

Fumière, O., Veys, P., Boix, A., Holst, C.v., Baeten, V., & Berben, G. (2009). Methods of detection, species identification and quantification of processed animal proteins in feedingstuffs. *Biotechnologie, Agronomie, Société et Environnement, 13*(Supplement), 59–70.

Lam, H. (2011). Building and searching tandem mass spectral libraries for peptide identification. *Molecular & Cellular Proteomics, 10*(12), 8565. R111.

Larsen, T., Ventura, M., Andersen, N., O'Brien, D. M., Piatkowski, U., & McCarthy, M. D. (2013). Tracing carbon sources through aquatic and terrestrial food webs using amino acid stable isotope fingerprinting. *PloS One, 8*(9), Article e73441.

Lecrenier, M. C., Marbaix, H., Dieu, M., Veys, P., Saegerman, C., Raes, M., et al. (2016). Identification of specific bovine blood biomarkers with a non-targeted approach using HPLC ESI tandem mass spectrometry. *Food Chemistry, 213*, 417–424.

Lecrenier, M.-C., Marien, A., Veys, P., Belghit, I., Dieu, M., Gillard, N., et al. (2021). Inter-laboratory study on the detection of bovine processed animal protein in feed by LC-MS/MS-based proteomics. *Food Control, 125*, 107944.

Lecrenier, M. C., Planque, M., Dieu, M., Veys, P., Saegerman, C., Gillard, N., et al. (2018). A mass spectrometry method for sensitive, specific and simultaneous detection of bovine blood meal, blood products and milk products in compound feed. *Food Chemistry, 245*, 981–988.

Lecrenier, M.-C., Veys, P., Fumière, O., Berben, G., Saegerman, C., & Baeten, V. (2020). Official feed control linked to the detection of animal byproducts: Past, present, and future. *Journal of Agricultural and Food Chemistry, 68*(31), 8093–8103.

Liland, N. S., Biancarosa, I., Araujo, P., Biemans, D., Bruckner, C. G., Waagbø, R., et al. (2017). Modulation of nutrient composition of black soldier fly (*Hermetia illucens*) larvae by feeding seaweed-enriched media. *PloS One, 12*(8). e0183188-e0183188.

Marbaix, H., Budinger, D., Dieu, M., Fumière, O., Gillard, N., Delahaut, P., et al. (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.

Marchis, D., Altomare, A., Gili, M., Ostorero, F., Khadjavi, A., Corona, C., et al. (2017). LC-MS/MS identification of species-specific muscle peptides in processed animal proteins. *Journal of Agricultural and Food Chemistry, 65*(48), 10638–10650.

McMahon, K. W., Fogel, M. L., Elsdon, T. S., & Thorrold, S. R. (2010). Carbon isotope fractionation of amino acids in fish muscle reflects biosynthesis and isotopic routing from dietary protein. *Journal of Animal Ecology, 79*(5), 1132–1141.

McMahon, K. W., Polito, M. J., Abel, S., McCarthy, M. D., & Thorrold, S. R. (2015). Carbon and nitrogen isotope fractionation of amino acids in an avian marine predator, the gentoo penguin (Pygoscelis papua). *Ecology and evolution, 5*(6), 1278–1290.

Mumcuoglu, K. Y., Miller, J., Mumcuoglu, M., Friger, M., & Tarshis, M. (2001). Destruction of bacteria in the digestive tract of the maggot of Lucilia sericata (Diptera: Calliphoridae). *Journal of Medical Entomology, 38*(2), 161–166.

Niedzwiecka, A., Boucharef, L., Hahn, S., Zarske, M., Steinhilber, A., Poetz, O., et al. (2019). A novel antibody-based enrichment and mass spectrometry approach for the detection of species-specific blood peptides in feed matrices. *Food Control, 98*, 141–149.

Ohana, D., Dalebout, H., Marissen, R. J., Wulff, T., Bergquist, J., Deelder, A. M., et al. (2016). Identification of meat products by shotgun spectral matching. *Food Chemistry, 203*, 28–34.

Olsvik, P. A., Fumière, O., Margry, R. J. C. F., Berben, G., Larsen, N., Alm, M., et al. (2017). Multi-laboratory evaluation of a PCR method for detection of ruminant DNA in commercial processed animal proteins. *Food Control, 73*, 140–146.

Rasinger, J. D., Marbaix, H., Dieu, M., Fumière, O., Mauro, S., Palmblad, M., et al. (2016). Species and tissues specific differentiation of processed animal proteins in aquafeeds using proteomics tools. *Journal of Proteomics, 147*, 125–131.

Steinhilber, A. E., Schmidt, F. F., Naboulsi, W., Planatscher, H., Niedzwiecka, A., Zagon, J., et al. (2018). Mass spectrometry-based immunoassay for the quantification of banned ruminant processed animal proteins in vegetal feeds. *Analytical Chemistry, 90*(6), 4135–4143.

Steinhilber, A. E., Schmidt, F. F., Naboulsi, W., Planatscher, H., Niedzwiecka, A., Zagon, J., et al. (2019). Application of mass spectrometry-based immunoassays for the species- and tissue-specific quantification of banned processed animal proteins in feeds. *Analytical Chemistry, 91*(6), 3902–3911.

Tanabe, S., Hase, M., Yano, T., Sato, M., Fujimura, T., & Akiyama, H. (2007). A real-time quantitative PCR detection method for pork, chicken, beef, mutton, and horseflesh in foods. *Bioscience Biotechnology and Biochemistry, 71*(12), 3131–3135.

Van Huis, A., van Itterbeeck, J., Klunder, H. C., Mertens, E., Halloran, A., Muir, G., et al. (2013). *Edible insects:future prospects for food and feed security*. Rome, Italy: Food and Agriculture Organization of The United Nations.

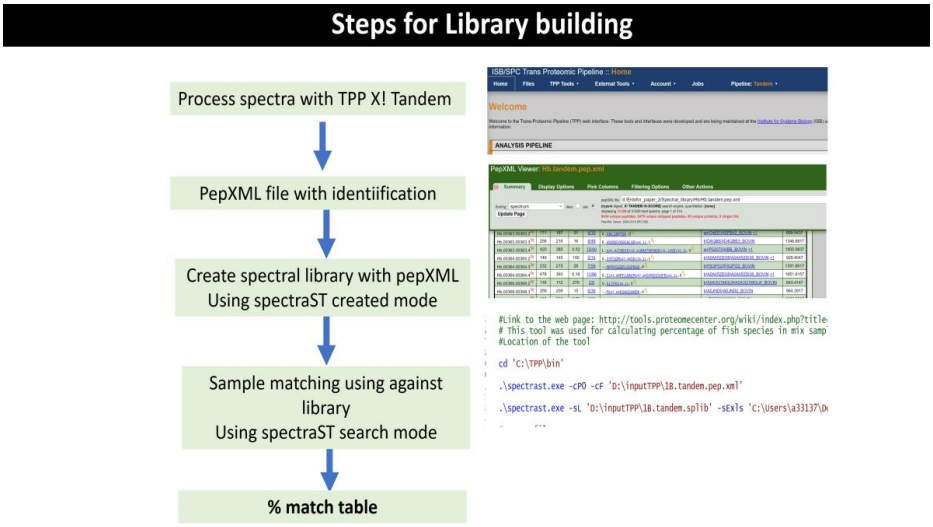
Van Raamsdonk, L. W. D., Prins, T. W., Meijer, N., Scholtens, I. M. J., Bremer, M. G. E. G., & de Jong, J. (2019). Bridging legal requirements and analytical methods: A review of monitoring opportunities of animal proteins in feed. *Food Additives & Contaminants: Part A, 36*(1), 46–73.

Wang, Y. V., Wan, A. H. L., Krogdahl, Å., Johnson, M., & Larsen, T. (2019). (13)C values of glycolytic amino acids as indicators of carbohydrate utilization in carnivorous fish. *PeerJ, 7*. e7701-e7701.

Wang, Y. V., Wan, A. H. L., Lock, E.-J., Andersen, N., Winter-Schuh, C., & Larsen, T. (2018). Know your fish: A novel compound-specific isotope approach for tracing wild and farmed salmon. *Food Chemistry, 256*, 380–389.

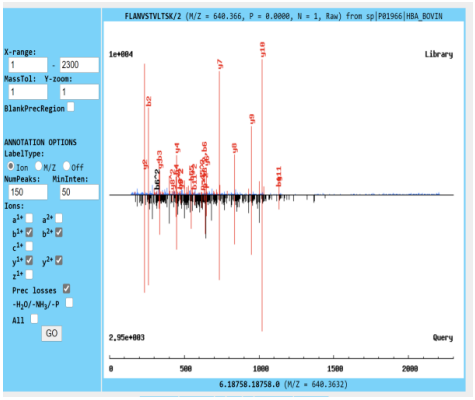
Wulff, T., Nielsen, M. E., Deelder, A. M., Jessen, F., & Palmblad, M. (2013). Authentication of fish products by large-scale comparison of tandem mass spectra. *Journal of Proteome Research, 12*(11), 5253–5259.

Supplementary Figure 1. Spectral library workflow



Supplementary Figure 2. Spectral library matching of samples to bovine hemoglobin library for dot product calculation (A) spectra matching and (B) table of ion annotation

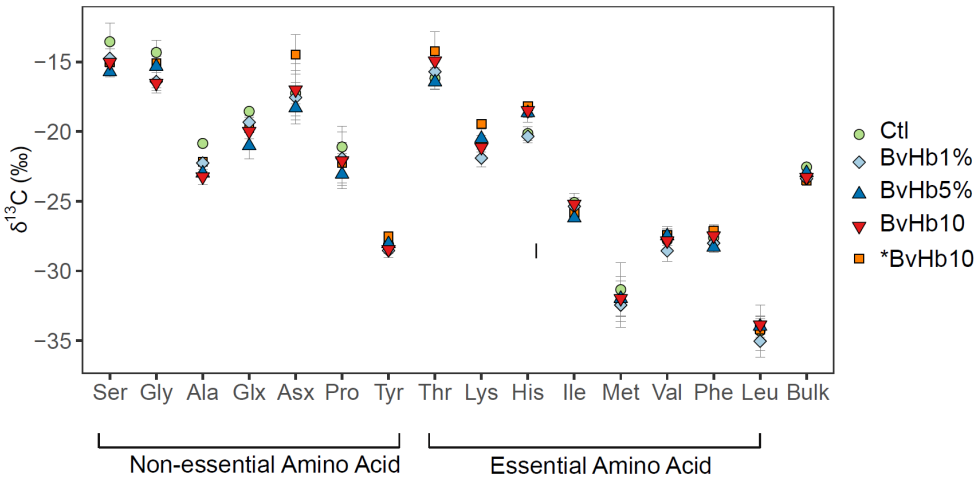
A



B

b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
148.0757	74.5415	1	F	12		
261.1598	131.0835	2	L	11	1132.6572	566.8322
332.1969	166.6021	3	A	10	1019.5732	510.2902
446.2398	223.6235	4	N	9	948.5360	474.7717
545.3082	273.1577	5	V	8	834.4931	417.7502
632.3402	316.6738	6	S	7	735.4247	368.2160
733.3879	367.1976	7	T	6	648.3927	324.7000
832.4563	416.7318	8	V	5	547.3450	274.1761
945.5404	473.2738	9	L	4	448.2766	224.6419
1046.5881	523.7977	10	T	3	335.1925	168.0999
1133.6201	567.3137	11	S	2	234.1448	117.5761
		12	K	1	147.1128	74.0600

Supplementary Figure 3. Average $\delta^{13}C_{AA}$ (mean ± SD) and $\delta^{13}C$ bulk values of BSF fed on different feeding media. See Material and Methods for the amino acid abbreviations



Paper I, Supplementary Tables can be downloaded from here https://ars.els-cdn.com/content/image/1-s2.0-S0956713521003212-mmc2.xlsx

Tables	Title	Legends
Table S1	Laboratories (A and B) qPCR results comparison of the feeding media and in BSF larvae grown on substrate containing bovine hemoglobin (n=2)	All samples extracted twice and analyzed in duplicate. The cut-off for PCR detection of ruminant DNA for Labs A and B was 36.50 and 35.89, respectively. Green = negative result and red = positive result.
Table S2	Bovine Hemoglobin peptides identified in the feeding media and in BSF larvae grown on substrate containing bovine hemoglobin (n=2) (QTOF, laboratory B)	Four bovine hemoglobin peptides (BHP 1-4), previously described in Niedzwiecka et al. (2019), were used in the present study (laboratory B) : VGGHAAEYGAEALER (BHP-1, m/z range = 510,08-511,08), EFTPVLQADFQK (BHP-2, m/z range = 711,37-712,37), AAVTAFWGK (BHP-3, m/z range = 475,26-476,26) and VVAGVANALAHR (BHP-4, 392,74-393,74). Peptides identified at estimated Rt. MS2 fragment pattern with expected intensity distribution.
Table S3	Bovine Hemoglobin peptides amount detected in the feeding media and in BSF larvae grown on substrate containing bovine hemoglobin (n=2) (triple quadrupole, laboratory C)	Values are means (fmol/200 µg sample weight) with their coefficient of variation (CV%). Bovine peptides previously described in Steinhilber et al. (2019) and in Niedzwiecka et al. (2019), were used in the present study (laboratory C); tissue-specific to blood: α-2-macroglobulin (A2M), complement component 9 (C9), hemoglobin α-chain (HBA); cartilage: myosin-7 (MYH7) and matrilin-1 (MATN1); and bone and milk: osteopontin (OPN).
Table S4	Bovine Hemoglobin spectral matching performed in the feeding media and BSF larvae grown on substrate containing bovine hemoglobin (n=2) (QTOF, laboratory D)	Hb SM; hemoglobin spectra matching, MP SM; milk protein spectra matching. A list of Hb SM and Hb specific matches peptides detected in the feeding media and in the BSF larvae fed the experimental diets are reported in Supplementary Tables 5 and 6, respectively. Hb specific matches peptides are marked in red in the Supplementary Tables 5 and 6.
Table S5	Bovine Hemoglobin and milk spectral matching performed in the feeding media detailed description (n=2) (laboratory D)	Feeding media: media condition; Query: spectra; ID: peptide sequence; Dot: dot product of spectral library and spectra comparison; protein: UniProt protein Ids.
Table S6	Bovine Hemoglobin and milk spectral matching performed in BSF larvae grown on substrate containing bovine hemoglobin (n=2) (laboratory D)	Detailed investigation of spectral matches with BvHb library to ensure specific matches; BSF larvae; BSF larvae grown on different condition; Query: spectra; ID: peptide sequence; Dot: dot product of spectral library and spectra comparison; protein: UniProt protein IDs.
Table S7	Compound specific amino acid analysis (CSIA) in feeding media and BSF larvae grown on substrate containing bovine hemoglobin (BvHb) (laboratory E)	Rank regression analysis was performed to assess if amino acid (AA) patterns correlate with concentrations of BvHb added to the feeding media and in BSF larvae, respectively.

II

Paper II

Varunjikar, M.S., Moreno-Ibarguen, C., Andrade-Martinez, J.S., Tung, H.S., Belghit, I., Palmblad, M., Olsvik, P.A., Reyes, A., Rasinger, J.D. and Lie, K.K.

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

Food Control (2022), 131, 108417

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

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

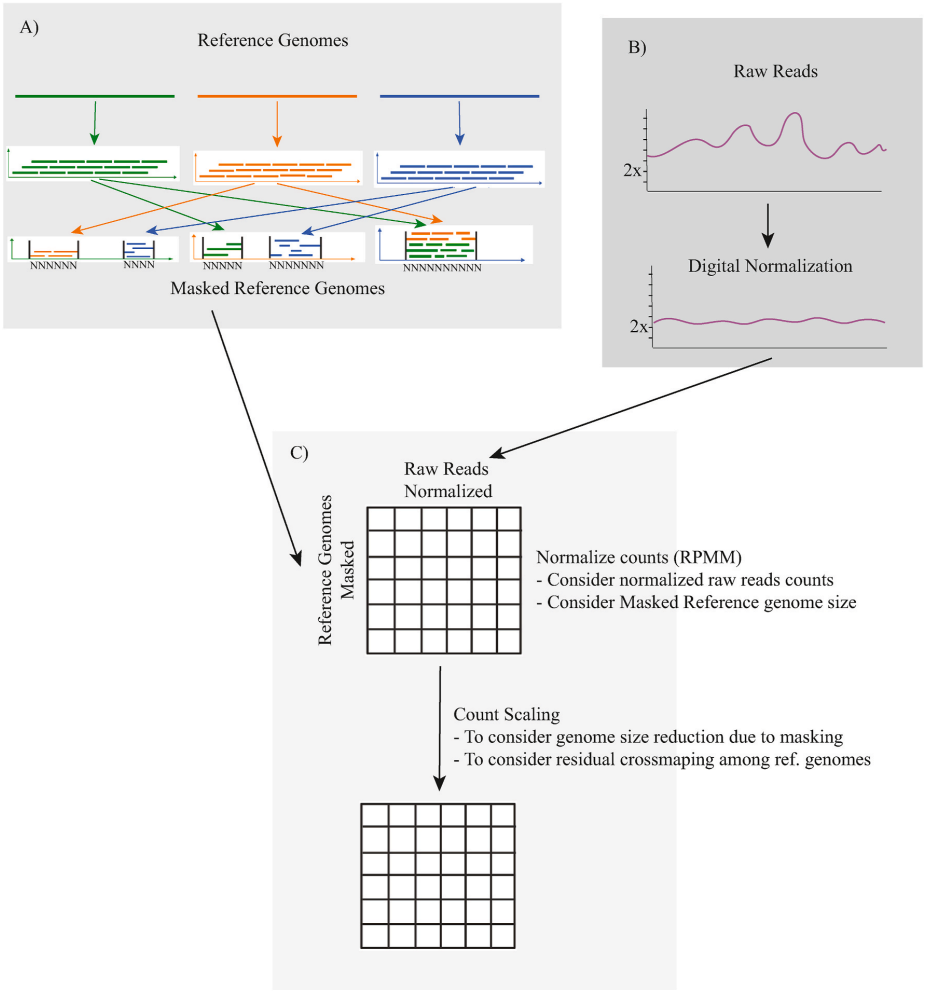


Fig. 1. Workflow of bioinformatics pipeline used for DNA sequencing analyses before calculating percentages. **(A)** Reference genome masking was conducted by generating a set of simulated reads from each genome followed by cross-mapping against all other reference genomes. Any identification of cross-mapping 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

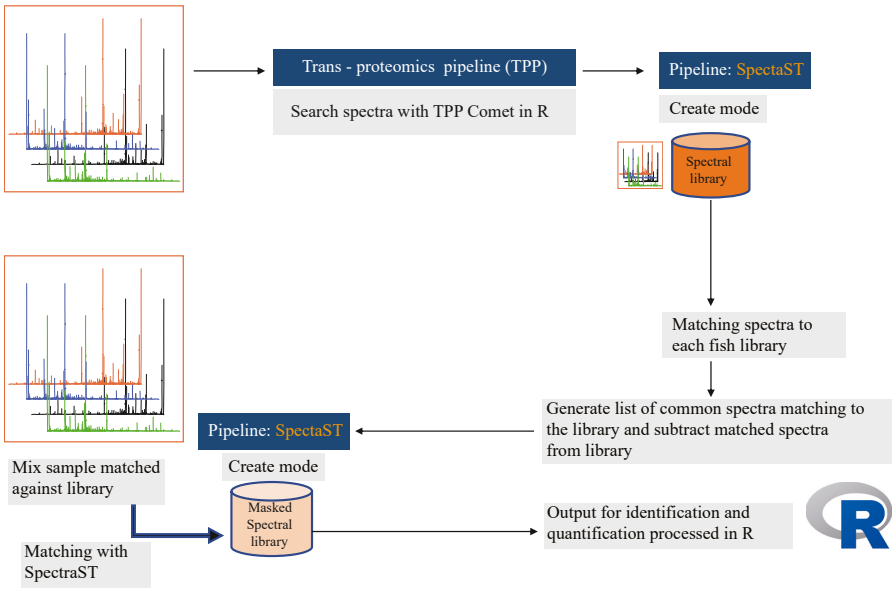


Fig. 2. Representation of proteomics bioinformatics methods used for calculation of percentages in the fish samples using spectral library workflow, where *Trans*-Proteomic Pipeline (TPP) was used for searching spectra and creating libraries as well as searching against the libraries.

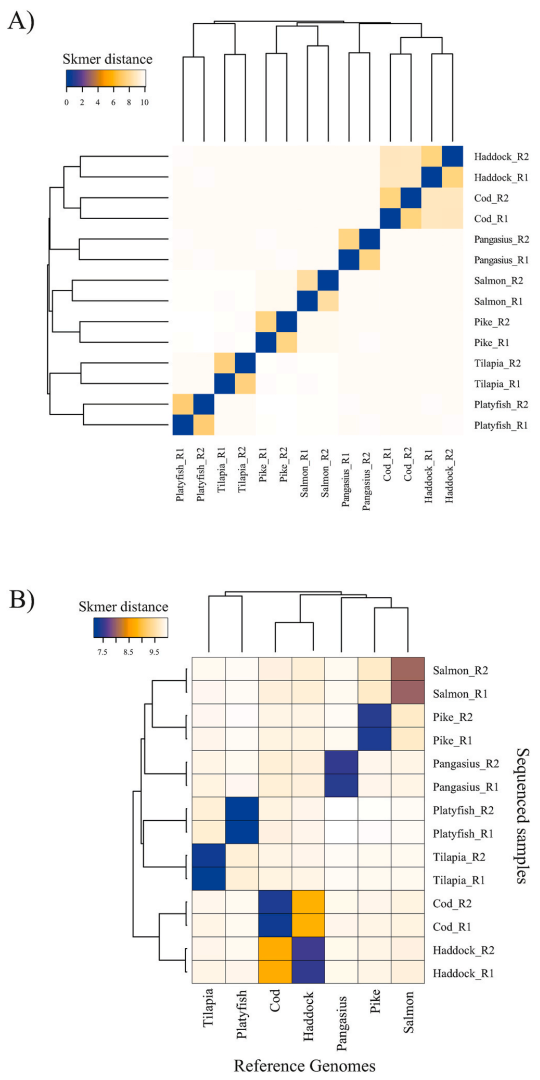


Fig. 3. Sample relatedness based on DNA sequencing using Skmer. (A) Analysis of pure samples with Skmer; note that forward (R1) and reverse (R2) reads from each sample are used, strong relatedness is identified from the corresponding paired sample. (B) Comparison of pure tissue samples against all the 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.

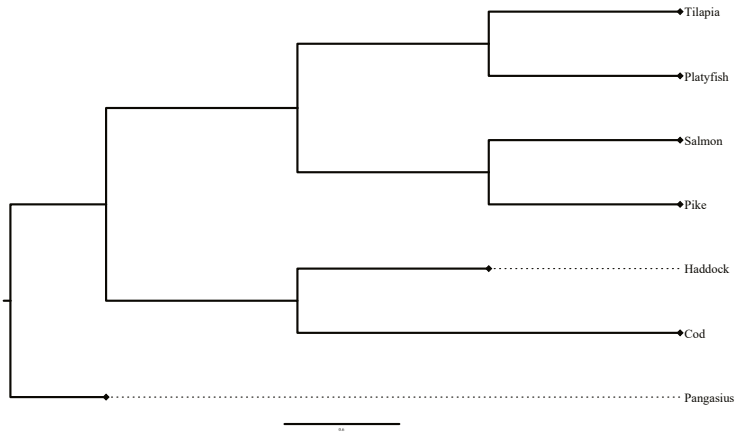


Fig. 4. Phylogenetic tree built from compareMS2 with top 500 spectra, which agrees with the phylogeny of the selected fish species. Scientific names of the species are as given here: Haddock (*Melanogrammus aeglefinus),* Tilapia (*Oreochromis niloticus),* Cod *(Gadus morhua),* Salmon *(Salmo salar),* 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). Inter-laboratory 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

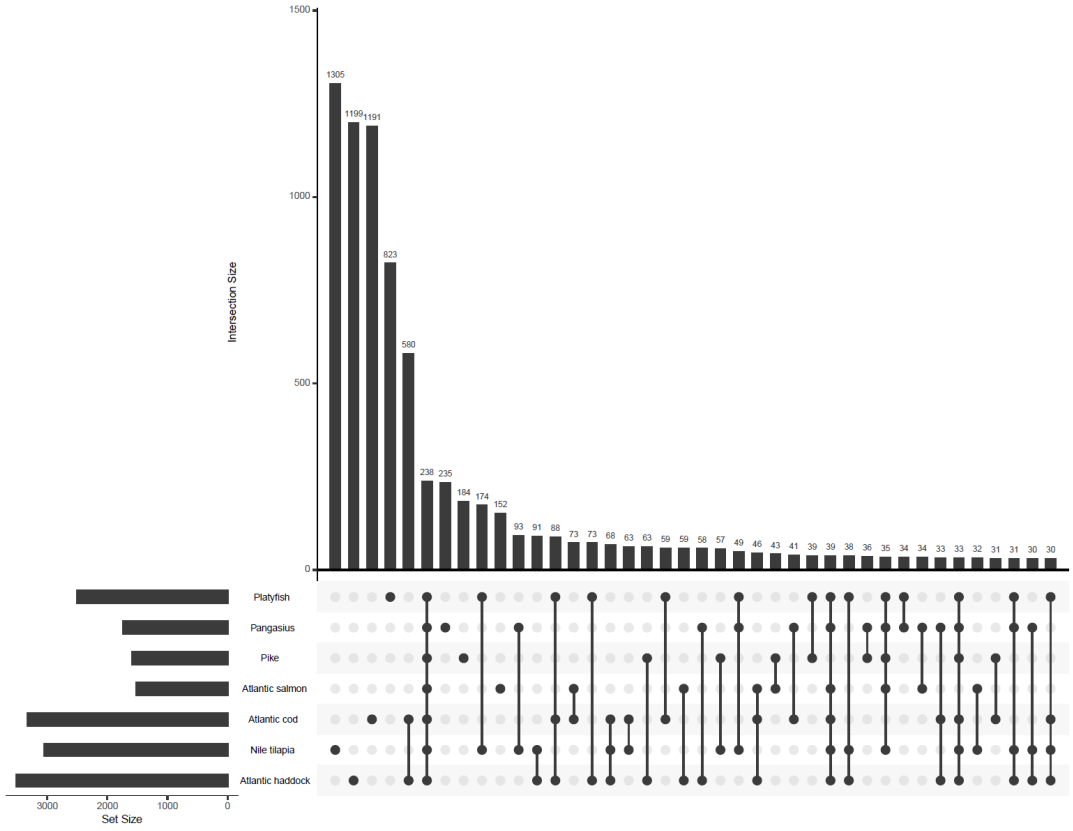
Wulff, T., Nielsen, M. E., Deelder, A. M., Jessen, F., & Palmblad, M. (2013). Authentication of fish products by large-scale comparison of tandem mass spectra. *Journal of Proteome Research, 12*(11), 5253–5259. https://doi.org/10.1021/pr4006525

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*.

Supplementary Figure 1



Paper II Supplementary Tables can be downloaded from here https://ars.els-cdn.com/content/image/1-s2.0-S0956713521005557-mmc1.xlsx

Tables	Title	Legends
Table S1	Reference genome from different species	Reference genome from fish species retrieved and from pangasius draft genome was assembled using SPADES
Table S2	Cross mapping among reference genome	Cross mapping among reference genome to estimate closeness and redundancy among targeted genome (values depicted in percentage)
Table S3	Cross mapping among reference genome	Cross mapping among reference genome to estimate closeness and redundancy among targeted genome (Actual values)
Table S4	Total and percent sequencing reads	Total and percent sequencing reads mapped against the masked reference genome
Table S5	Quantification of fish percentage	Quantitation of fish mixture (N=4) and DNA mixture in percentage (N= 3)
Table S6	Cross-maching with spectral libraries	(A) crossmatching with spectral libraries and matches obteined against each library; (B) Normalization by total number spectra to obtain percent; (C) percent matches obtained per spectra
Table S7	Example SpectraST output	Output of SpectraST; description of spectral matches is given
Table S8	Taxon identifier and number of proteins of selected fish species.	The Number of proteins, refers to the number of predicted and (reviewed) proteins of fish species listed in the UniprotKB/Swiss-Prot reference proteome; UP refers to the respective reference proteome number.

Paper III

Varunjikar, M. S., Belghit, I., Gjerde, J., Palmblad, M., Oveland, E., & Rasinger, J. D**.**

Shotgun proteomics approaches for authentication, biological analyses, and allergen detection in feed and food-grade insect species

Food Control (2022), 137, 108888

III

Food Control 137 (2022) 108888

Contents lists available at ScienceDirect

Food Control

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

Shotgun proteomics approaches for authentication, biological analyses, and allergen detection in feed and food-grade insect species

Madhushri S. Varunjikar [a,1], Ikram Belghit [a,1], Jennifer Gjerde [a], Magnus Palmblad [b], Eystein Oveland [a], Josef D. Rasinger [a,*]

[a] Institute of Marine Research, P.O. Box 1870 Nordnes, 5817, Bergen, Norway
[b] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, the Netherlands

ARTICLE INFO

Keywords:
Tandem mass spectrometry
Edible insect
Feed control
Spectral libraries
Insect allergens
Q exactive orbitrap

ABSTRACT

Untargeted proteomics can contribute to composition and authenticity analyses of highly processed mixed food and feed products. Here, we present the setup of an analytical flow tandem mass spectrometry method (AF-HPLC HR-MS) for analysis of insect meal from five different species. Data acquired were compared with previously published data employing spectra matching and standard bottom-up proteomics bioinformatics analyses. In addition, data were screened for insect species marker peptides and common allergens, respectively. The results obtained indicate that the performance of the newly established AF-HPLC HR-MS workflow is in line with previously published methods for insect species differentiation. Data obtained in the present study, also lead to the discovery of novel markers for the development of targeted MS analyses of insect species in food- and feed-mixes and highlighted that known allergen such as arginine kinase or tropomyosin were consistently detected across all five species tested.

1. Introduction

In 30 years, 9.7 billion people are estimated to live on our planet and the demand for feed and food crops is expected to increase to 25–70% above today's levels (FAO et al., 2018). To ensure food security for the growing population, novel food and feed ingredients such as insects will play an important role as future protein sources in animal feed and human nutrition (IPIFF, 2021). However, in the European Union (EU), their current and future usage in the feed and food sector is and will be regulated by strict legislative texts. To enforce and monitor regulatory guidelines robust and versatile high-throughput analytical tools will be required; in this context mass-spectrometry (MS) based proteomics approaches have shown to hold great promise (Belghit et al., 2021; Lecrenier et al., 2018; Varunjikar et al., 2022).

The most common proteomics workflow takes the bottom-up approach in which proteins in the sample are enzymatically digested by a protease (e.g., trypsin), and the resulting peptides are analysed by high-performance liquid chromatography (HPLC) coupled to a tandem mass spectrometer (HPLC-MS/MS). The data output files including both MS and MS/MS spectra are then analysed using different proteomics bioinformatics tools that allow for peptide identification and protein inference based on different algorithms. The combinations of a quadrupole with a high resolution TOF analyser (QTOF) or with an high resolution orbitrap mass spectrometer (HR-MS) are among the most widely used for shotgun proteomics analyses (Szabó et al., 2021). Untargeted proteomic workflows commonly aim to identify as many peptides and proteins as possible and usually utilise nanoflow HPLC (nano-LC) for chromatographic separation of samples. Nano-LC is more sensitive than normal flow approaches are and hence, the preferred choice in bottom-up proteomics. However, the use of nano-flow LC is technically challenging, and frequent column changes are required due to faster build-up of high back pressure when compared to normal flow HPLC. Normal flow HPLC, also referred to as analytical flow (AF) HPLC, is simpler to set up, more robust to run in routine proteomic analysis settings (Lenčo et al., 2018). Thus, in regulatory laboratories for high-throughput feed or food safety and authenticity analyses, the use of AF-HPLC-MS/MS-based proteomics can contribute to make implementation of proteomic approaches attractive and affordable for control laboratories (Sentandreu & Sentandreu, 2011).

* Corresponding author.
E-mail addresses: madhushri.shrikant.varunjikar@hi.no (M.S. Varunjikar), ikram.belghit@hi.no (I. Belghit), jennifer.gjerde@hi.no (J. Gjerde), n.m.palmblad@lumc.nl (M. Palmblad), eystein.oveland@hi.no (E. Oveland), josef.rasinger@hi.no (J.D. Rasinger).
[1] Equal contribution; shared first authorship.

https://doi.org/10.1016/j.foodcont.2022.108888
Received 10 December 2021; Received in revised form 8 February 2022; Accepted 10 February 2022
Available online 23 February 2022

Abbreviations

AF	Analytical flow
BSF	Black Soldier Fly
HC	House Cricket
HR-MS	High Resolution-Mass Spectrometry
LW	Lesser mealworm
MGF	Mascot Generic Format
mzML	open standard data format for mass spectrometry data
MF	Microflow
MW	Morio worms
PSM	Peptide-Spectrum Match
SLM	Spectral Library Matching
TPP	Trans-Proteomic Pipeline
QTOF	Quadrupole time-of-flight
HPLC-MS/MS	High-Performance Liquid Chromatography coupled to Tandem Mass Spectrometry
YW	Yellow meal Worm

Proteomic-based methods using HPLC-MS/MS were recently identified as promising tools to complement current standard techniques of processed animal protein (PAP) detection in feed in a scientific opinion by the European Food Safety Authority (EFSA) (Aguilera et al., 2018). According to European regulation of animal protein (European Commission, 2013/51, European Commission, 2017/893), insects reared to produce PAP are to be considered farmed animals. In 2017, the European Commission (EC) allowed the use of insect meal processed from seven different black soldier fly (BSF) (*Hermetia illucens*), common housefly (*Musca domestica*), yellow mealworm (*Tenebrio molitor*), lesser mealworm (*Alphitobius diaperinus*), house cricket (*Acheta domesticus*), banded cricket (*Gryllodes sigillatus*) and field cricket (*Gryllus assimilis)* (European Commission, 2017/893). Silkworm (*Bombyx mori*) was recently added to the approved list of insect species in aquaculture (European Comission, 2021/1372), resulting in a total of eight insect species allowed in aquafeed. Recently, in August 2021, the EC adopted the decision to allow the use of insect PAP in formulated pig and poultry feeds (European Comission, 2021/1372). At the time of writing, a draft bill for implementing the regulation to authorise the commercialisation of frozen and dried migratory locust (*Locusta migratoria)* on the EU market was issued (IPIFF, 2021), and following a favourable opinion of the EFSA Panel on Nutrition, Novel Foods and Food Allergens (NDA) (EFSA NDA panel, 2021a), the placing on the market of dried yellow mealworm (*Tenebrio molitor*) larva as a novel food under Regulation (EU) 2015/2283 was authorized (European Comission, 2021/882). Also, a favourable opinion on the draft legal act authorising the placing on the market of frozen, dried and powder forms of house cricket (*Acheta domesticus*) as a novel food was issued (EFSA NDA panel, 2021b). Concerning house cricket, EFSA highlighted that the consumption of the evaluated insect proteins may potentially lead to allergic symptoms and that in addition, allergens present in substrate fed to insects may end up in the insect consumed (EFSA NDA panel, 2021b). Therefore, analytical approaches must be developed which allow for an unambiguous detection and identification of white-listed insect species in insect-protein containing feed or food products. Among the five insect species used in this study, four are white-listed insect species whereas one species, morio worm (*Zophobas morio*), is not officially approved in the EU for use in feed or food but is considered a potential future feed or food ingredient (Rumbos & Athanassiou, 2021). This species was not included in previously published Belghit et al., 2019 but as it might be used as a food and feed in future, we included it in the current study.

For safe use of insects in feed and food real-time polymerase chain reaction (qPCR) assays are being developed (Daniso et al., 2020; Debode et al., 2017; Garino et al., 2021; Köppel et al., 2019). In parallel, targeted and non-targeted HPLC-MS based proteomics methods are being developed by several laboratories. Analyses of MS/MS spectra were shown to be suitable for the identification, quantification and tracing of processed animal protein (PAP) in feed (Belghit et al., 2019, 2021; Marbaix et al., 2016; Rasinger et al., 2016; Steinhilber et al., 2018a,b), the detection of allergens in edible insects (Bose et al., 2021; Francis et al., 2019), and the identification of species origin. When genomic information is scarce (Belghit et al., 2019, 2021; Nessen et al., 2016; Ohana et al., 2016; Rasinger et al., 2016; Varunjikar et al., 2022; Wulff et al., 2013).

The objectives of the present study were to (*i*) set up and optimise an analytical flow LC-MS/MS proteomics assay for insect species authentication, (*ii*) compare data obtained from two different proteomics workflows, microflow HPLC (MF-HPLC) QTOF and the optimized AF-HPLC HR-MS, using spectra matching approaches, and (*iii*) based on both MF-HPLC QTOF data and AF-HPLC HR-MS data, identify common, and unique insect species-specific proteins, and potential allergens.

2. Materials and methods

2.1. Samples

HeLa Protein Digest Standards were purchased from Thermo Fisher Scientific Pierce™ (Thermo Scientific, San Jose, CA) and was used for standardisation of the instrument and optimising the HPLC and MS conditions with the HR-MS orbitrap instrument. Eight samples from species of the Diptera order; black soldier fly larvae (BSF) (*H. illucens*), nine samples from species of the Coleoptera order, including the yellow mealworm (YW) (*T. molitor*) and the lesser mealworm (LW) (*A. diaperinus*), and two samples from the Orthoptera order; house cricket (HC) (*A. domesticus*) were collected from different insect food and feed companies. The eighteen insect meal samples have been reported in more detail elsewhere (Belghit et al., 2019). Additionally, one morio worm (MW) (*Z. morio)* sample was included in the current study (Supplementary Table 1).

2.2. Protein extraction

Insect samples were weighed into a test tube of the One Plus Grinding kit (GE Healthcare Life Science, 80648337, Piscataway, NJ, USA) and lysis buffer (4% SDS, 0.1 M Tris-HCl, pH 7.6). Samples were kept on ice and homogenised in the tube containing resins with a pestle. To this homogenate, freshly prepared, 1 M Dithiothreitol was added to obtain a final concentration of 0.1 M, further, these tubes were centrifuged for 10 min at 15,000 g to remove resin and other debris. The supernatant was collected and heated at 95 °C on the heat-block for 5 min. After this, samples were centrifuged again, and the supernatant was collected in new tubes to store at −20 °C until further processing. The protein concentration of extracted samples was determined by the Pierce 660 assay as described in Rasinger et al. (2016) using BSA for the standard curve (Thermo Scientific, San Jose, CA).

2.3. Protein digestion and purification

Protein extracts from insect samples were digested with filter-aided sample preparation method as described in Belghit et al. (2019), where 150 mg of extracted protein was diluted with 200 μL of 8 M urea solution prepared in Tris-HCl (100 mM, pH 8.5). This solution was transferred to an ultrafiltration spin column (Microcon 30, Millipore, Burlington, MA, USA). Further, these proteins were alkylated as described in Belghit et al. (2019) with 50 mM of iodoacetamide for 20 min for incubation in darkness at room temperature. After incubation, the protein mixture in the column was washed with 200 μL of 8 M urea solution along with 100 μL of 50 mM ammonium bicarbonate solution.

After this step trypsin was added to the filters in 1:50 enzyme to protein ratio and tubes were incubated for 16 h at 37 °C. After incubation filters were centrifuged and washed with 40 μL of 50 mM ammonium bicarbonate solution with the same molarity as mentioned above and later with NaCl (0.5 M). Following centrifugation, the digested tryptic peptides were purified with Pierce™ C18 spin column (ThermoFisher, 89870). The columns were first washed with methanol/water (50/50, v/v), and then equilibrated with wash solvent (acetonitrile/trifluoroacetic acid/water, 5/0.5/94.5, v/v/v). Digested samples were diluted with acetonitrile/trifluoroacetic acid/water (20/2/78, v/v/v) and loaded into the columns. Peptides were eluted with acetonitrile/water (30/70, v/v) and subsequently evaporated in a speed vacuum dryer (LABCONCO CentriVap micro IR). Peptide pellets were dissolved in acetonitrile/formic acid/water (2/0.1/97.9, v/v/v) and kept at −20 °C until mass spectrometric analyses.

2.4. LC-MS/MS analyses

2.4.1. QTOF

For the ESI-MS/MS maXis Impact UHR-TOF (Bruker, Bremen Germany), the method is described in Belghit et al. (2019). Briefly, HPLC analyses were performed using the UltiMate 3000 HPLC system (Thermo Scientific, San Jose, CA). Approximately 5.0 μg samples were separated using 2.0 μm Acclaim PepMap 100 C18, 1 × 150 mm (Thermo Scientific, San Jose, CA). The flow rate was 40 μL/min. Mobile phase A was 95% water, 5% acetonitrile, 0.1% formic acid. Mobile phase B was 20% water, 80% acetonitrile, 0.1% formic acid. The digest was injected, and the organic content of the mobile phase was increased linearly from 4% B to 40% B in 60 min and from 40% B to 90% B in 10 min, and then washed with 90% B for 10 min and with 4% B for 10 min, for a total of 90 min. The column effluent was directly connected to the maXis UHR-TOF coupled with electrospray ionisation (ESI) (Bruker, Billerica, MA, USA). In the survey scan, MS spectra were acquired for 0.5 s in the mass to charge (m/z) range between 50 and 2200. The 10 most intense peptides ions 2+ to 4+ were fragmented during a cycle time of 3 s. The collision-induced dissociation (CID) energy was automatically set according to the m/z ratio and charge state of the precursor ion. The mass spectrometer and HPLC systems are controlled by Compass HyStar 3.2 (Bruker, Billerica, MA, USA). Regarded as micro flow-HPLC QTOF (MF-HPLC QTOF) here onwards in the text.

2.4.2. HR-MS Orbitrap

For the optimisation, HPLC analyses were performed using Vanquish Horizon binary HPLC (Thermo Scientific, San Jose, CA). Separations were performed using 2.2 μm Acclaim Vanquish C18, 2.1 × 250 mm (Thermo Scientific, San Jose, CA). The column temperature was maintained at 50 °C. The solvents A and B were 0.1% (v/v) formic acid in high purity water (18.2 MΩ × cm) and 0.1% formic acid (v/v) in 100% acetonitrile, respectively. Gradient conditions are described in Supplementary Table 2, with different gradient lengths varying from 60 to 80 min. The flow rate varied between 300 and 400 μL/min (Supplementary Table 2). Different amounts of HeLa cells digest were loaded (0.5–40 μg, Supplementary Table 3).

Eluting peptides were analysed on HR-MS Q Exactive Orbitrap (Thermo Scientific, San Jose, CA). MS instrumental tune parameters were set as follows: ESI spray voltage was 3.5 kV, sheath gas flow rate was 40 AU, the auxiliary gas flow rate was 10 AU, the capillary temperature was 320 °C, probe heater temperature was 400 °C and S-lens RF level was set to 50. Data-dependent acquisition (DDA) MS2 method with full MS scans in positive polarity was obtained at resolution settings of 17,500, 35,000, and 70,000 (Supplementary Table 2). Mass range was set at 200–2000 m/z and an AGC target was 5.0×10^5 up to 3.0×10^6 with a maximum injection time of 50 ms. For MS2, the resolution settings were 17,500 and 35,000 at a fixed first mass of 140 m/z with an AGC target value of 5.0×10^5 and an isolation window of 1.2 m/z. The normalised collision energy set was 32 and the top 10 precursors were selected for fragmentation. The signal intensity threshold was 2.0×10^4 with dynamic exclusion of 10, 20 and 30 s (Supplementary Table 2). This is regarded as analytical flow- HPLC HR-MS (AF-HPLC HR-MS) here onwards in the text.

After the optimisation of the HPLC and MS parameters with the HeLa Digest, the developed HR-MS workflow was implemented to analyse the nineteen insect meal samples. Gradient conditions were as follows: 2% B to 35% B in 62 min, hold at 95% B until 5 min and 2% B from 67.1 until 80 min. The flow rate was 400 μL/min flow rate (test number 19 in Supplementary Table 2). MS scans were obtained at a resolution of 70,000. Mass range was set at 350–2000 m/z and an AGC target was 3.0×10^6 with a maximum injection time of 50 ms. For MS2, the resolution was 35,000 at a fixed first mass of 140 m/z with an AGC target value of 3.0×10^6 and an isolation window of 1.2 m/z. The normalised collision energy set was 32 and the top 10 precursors were selected for fragmentation. The signal intensity threshold was 2.0×10^4 with dynamic exclusion of 30 s.

2.5. Bioinformatic analyses

2.5.1. Direct spectral comparison and Spectral library building with SpectraST

Proteomic-based phylogenetic data analysis was performed as described in Varunjikar et al. (2022). In short for direct spectral comparison of tandem mass spectra using compareMS2 (compareMS2, 2021; GUI, 2021; Palmblad & Deelder, 2012) MGF files containing the top 500 most intense tandem mass spectra were created using msConvert (version: 3.0., ProteoWizard). CompareMS2 was used to create distance matrices and phylogenetic trees. Overview of bioinformatics analyses is given in Supplementary Fig. 1.

Using the mzML and pepXML files generated from MF-HPLC QTOF and AF-HPLC HR-MS data and search output, spectral libraries (SLs) were created for each of all the five insect species using SpectraST (version 5.0) as previously described (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. Post-processing of the results was done in R (version 4.0.3) Outputs were recorded using tidyverse functions (version 1.3.0) and UpSetR (version 1.4.0).

2.5.2. Protein identification and data analysis

For analyses of acquired spectra from HeLa cell digest MSGFplus (V.1.26.0 (Pedersen, 2021).) search engine was used in R interphase to match the spectra to the UniProt human reference proteome (up000005640). Post analyses were done in R (version 4.0.3).

For identification of PSM, peptides, and proteins and to compare percentage identification from MF-HPLC QTOF and AF-HPLC HR-MS, tandem mass spectra were searched against proteomes of respective species from UniProt databased as described in 2.1 (accessed on June 2021).

For proteome analyses and marker detection, acquired data were matched against reviewed sequences (12, 976) from Arthropoda species (accessed July 2021) using Comet search as implemented in the TransProteomics Pipeline (TPP) (version 5.2.0 (Deutsch et al., 2015)). 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 cysteine and oxidation of methionine were set as fixed and variable modification, respectively. Generated pepXML files were further analysed using PeptideProphet and ProteinProphet using 1% level false discovery rate (FDR) (Keller et al., 2002). Post-processing of the acquired data was done in R (version 4.0.3). Data processing and statistical comparison of proteomics samples were performed in Omics Explorer V 3.6 (Qlucore AB, Lund, Sweden). The data were analysed using two-way ANOVA of the involved insect species (groups were sample species), unsupervised principal component analyses (PCA) and hierarchical cluster analysis (HCA). For

comparing the detected protein Venn diagrams were created using www.biovenn.nl.

2.5.3. Allergen detection

For allergen detection, a list of food allergens was downloaded from (www.allergen.org) along with allergen families and biochemical names (48 sequences) and these allergen sequences were downloaded from UniProt to create a database. The collected data from each instrument were searched against the database using TPP to evaluate allergen detection ability. Data processing and statistical comparison of detected allergenic proteins from samples were performed in Omics Explorer.

3. Results and discussion

3.1. Set up and optimisation of an AF-HPLC HR-MS system for untargeted proteomics

In the present study, the performance of an AF-HPLC coupled to a standard HR-MS was tested by injecting different amounts of HeLa cell digests using different combinations of HPLC and MS and MS/MS2 settings. Since the objective of the present work was to develop a time-efficient method suitable for regulatory use, only three relatively short HPLC run-time lengths (60, 70 and 80 min) with an increasing gradient of 4% (v/v) to 50% (v/v) mobile phase B were tested; run-time lengths of 90 min and longer, which commonly are employed in non-targeted expression proteomics analyses (Kelstrup et al., 2014; Varunjikar et al., 2022), were considered impractical for use in routine regular analyses settings.

As expected, increasing the gradient time resulted in an increased number of tandem mass spectra (Supplementary Table 2). Using a run-time length of 80-min and 20 μg of HeLa digest, yielded a total of 13562 of spectra. When matched against the HeLa cell reference proteome (up000005640) using the MSGFplus search engine this resulted in 8946 peptide-spectrum matches (PSM's) and the identification of 7553 and 1951 unique peptides and proteins, respectively (Supplementary Table 2). Similar results were obtained for the analysis of 5 μg of HeLa digest over a 90 min gradient, with a nanoflow HPLC instrument coupled to Linear Trap Quadrupole (LTQ) Orbitrap Velos mass spectrometer (Michalski et al., 2011) and when analysing 20 μg of HeLa digest using a Standard flow multiplexed Proteomics (SFloMPro) system coupled with a HR-MS Classic using a 90 min gradient (Jenkins & Orsburn, 2020).

Peptide and protein identification on an AF-HPLC HR-MS (as well as on any other HPLC-MS/MS systems), in addition to gradient length, are also dependent on injected sample amounts, which must be optimised for each respective system (Jenkins & Orsburn, 2020). Recently, Lenčo et al. (2018), analysed 0.5 and 2 μg of HeLa digest and observed an increase in protein and peptide identification of up to 14% with 2 μg compared to 0.5 μg of HeLa digest. In the aforementioned study, the authors optimised a standard-flow HPLC-MS system with the aim to identify a sample loading amount that yielded a comparable number of proteins and peptides that usually can be identified when using nano LC-MS systems (Lenčo et al., 2018). In the present study, loading amounts of 0.5–40 μg HeLa digest were analysed. As can be seen in Supplementary Fig. 2 and Supplementary Table 3, the PSM, unique peptide and protein counts increased linearly with increasing amounts of HeLa digest up to 5 μg when a plateau was reached. A 10-fold increase in sample load in the column (in the range of 0.5–5 μg peptide) increased the identification rate of peptides and proteins to 23% and 11%, respectively (Supplementary Fig. 2 and Supplementary Table 3). No further increase in the number of features detected was observed when up to 40 μg peptide (the highest amount of HeLa digest tested in the present study) were injected. Hence, 5 μg were selected for further analyses of the insect meal samples.

Taken together, the data generated here suggest that, given that the sample quantity is not a limiting factor (Jenkins & Orsburn, 2020), using an AF-HPLC HR-MS, could be a viable alternative for use in regulatory laboratories to the more conventional nanoflow HPLC workflow routinely used in MS-based proteomics.

3.2. Quality control and insect species identification

Following setup and optimisation of the AF-HPLC HR-MS setup, the assay settings shown in test 19, Supplementary Table 2, were applied for comparing analysis outputs of insect-based MS data generated in the present study with data previously published by our group (Belghit et al., 2019). Insect MS data acquired previously on an MF-HPLC QTOF instrument (massIVE ID: MSV000083737) were reanalysed using compareMS2 and a TPP-based bioinformatics workflow to compare with data generated here (AF-HPLC HR-MS, massIVE ID: MSV000088034). In both studies (MF-HPLC QTOF and AF-HPLC HR-MS based workflows), similar gradient lengths (varying flow rates) and loading amounts of insect meal samples (80 min and 5.0 μg, respectively) were applied.

Analysis outputs from compareMS2 have previously been found to be useful in the determination of the effects of sample preparation and analysis approaches on the data acquired by mass spectrometry (Van Der Plas-Duivesteijn et al., 2016). Using compareMS2, in the present study distance matrixes were calculated for both insect datasets and two representative dendrograms were constructed. As shown in Fig. 1, mass spectra from both datasets were successfully arranged according to the insect species and molecular phylogeny of insects (Supplementary Fig. 3), respectively. The spectral clustering of insects reflects the relatedness of insect species at the taxonomic level and is in line with data shown previously where insect grouping based on MS data was found to be based on the orders Diptera, Coleoptera, and Orthoptera (Belghit et al. (2019). In other words, overall, all insect species analysed in the present study were well separated using compareMS2, indicating that even with only 500 spectra collected, using AF-HPLC and a routine MS instrument sufficient data can be generated to allow for a species-level differentiation of protein sources in food- and feed samples. The spectral distances obtained by pairwise spectra comparison of data acquired with the HR-MS also were comparable with those obtained using the previously published MF-HPLC QTOF data (Belghit et al., 2019). This was consistent with previous molecular phylogenetic studies conducted using compareMS2 in selected species of interest in relation to food- and feed authenticity and adulteration analyses, respectively (Ohana et al., 2016; Rasinger et al., 2016; Varunjikar et al., 2022; Wulff et al., 2013).

In addition to the compareMS2 analyses, we subjected the previously published data instrument (MF-HPLC QTOF; massIVE ID: MSV000083737) and the data obtained in the present study (AF-HPLC HR-MS; massIVE ID: MSV000088034) to a standard bottom-up proteomics data analysis workflow as described in Belghit et al. (2019), with the exception that the Comet search engine in TPP was used instead of X! Tandem in proteoQC. The spectra identification output of the Comet search engine of both the datasets (MF-HPLC QTOF and AF-HPLC HR-MS) is given in Table 1. The results showed that with the exception of one species (BSF), the number of PSMs, peptides, and proteins were twice as high when running MF-HPLC QTOF based analysis workflow compared to the newly developed AF-HPLC HR-MS -based approach., Approximatively, the same number of PSMs, peptides, and proteins were detected with the MF-HPLC QTOF and AF-HPLC HR-MS for other insect species (YM, HC, LW, and MW, Table 1). Contrary to the raw number of spectra obtained, the percentage of identified spectra was consistently higher using the AF-HPLC HR-MS workflow when compared to the MF-HPLC QTOF workflow. A total of 30% more spectra were identified for BSF, YM, and HC samples when using the AF-HPLC HR-MS workflow (Table 1).

In summary, based on the bioinformatic analysis of the insect samples data published earlier (massIVE ID: MSV000083737) and data generated in the present study (massIVE ID: MSV000088034), the results obtained, indicate that AF-HPLC HR-MS provides data of sufficient

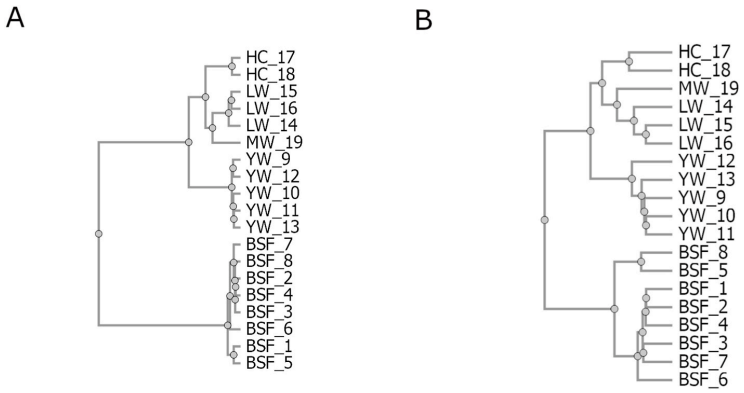


Fig. 1. Species-specific insect meal samples differentiation with direct comparison of spectra obtained by tandem mass spectrometry using compareMS2. Data obtained with (A) MF-HPLC QTOF (described in Belghit et al., 2019) and (B) the developed AF-HPLC HR-MS workflow. BSF = black soldier fly; YM = yellow mealworm; LM = lesser mealworm; HC = house cricket; MW = morio worm.

Table 1
Total numbers of spectra, identified proteins, and peptides using Comet search engine from 19 insect meal samples.

Species	MF-HPLC QTOF Belghit et al. (2019)					AF-HPLC HR-MS (newly developed)				
	tSpectra	PSM	Peptides	Proteins	% id	tSpectra	PSM	Peptides	Proteins	% id
BSF1	28176	16927	16860	11158	60%	9656	8656	8530	5838	90%
BSF2	28497	16857	16761	10817	59%	9089	8101	7960	5314	89%
BSF3	27201	15133	15049	10034	56%	10117	8901	5724	4485	88%
BSF4	28151	17011	16899	10903	60%	10049	9000	8812	5729	90%
BSF5	21910	12672	12616	9010	58%	9272	8484	8427	6047	92%
BSF6	22043	12595	12525	8705	57%	9811	8905	8796	5897	91%
BSF7	25050	12663	12583	8758	51%	10105	9019	8858	5647	89%
BSF8	28171	16283	16199	10677	58%	8749	8041	8015	5993	92%
YW9	30051	6927	6644	899	23%	10171	7228	6989	900	71%
YW10	26590	10490	9944	960	39%	10735	7654	7411	900	71%
YW11	28145	11637	10991	972	41%	10403	7472	7244	910	72%
YW12	26888	9509	9075	938	35%	9190	6263	6110	872	68%
YW13	29566	12470	11770	985	42%	10316	7426	7206	909	72%
LW14	27434	680	570	92	2%	9908	412	387	73	4%
LW15	24166	564	496	83	2%	10283	439	404	72	4%
LW16	25740	566	494	82	2%	10278	408	378	70	4%
HC17	26423	14085	13881	5060	53%	10620	9657	9556	4271	91%
HC18	24762	12507	12358	4851	51%	10121	9274	9172	4196	92%
MW19	24044	416	369	13	2%	10258	368	330	13	4%

* tSpectra - total spectra in the file; PSM– protein spectra matches; Peptides – number of identified peptides; Proteins – number of identified proteins; % id – percentage of number of identified spectra divided by total number of spectra; BSF – black soldier fly; YW – yellow mealworm; LW – lesser mealworm; HC– house cricket; MW – Morio Worms (data used from QTOF instrument Belghit et al. (2019) and HR-MS instrument).

quality to perform non-targeted species identifications of insects intended for use in food and feed. Having established that the performance of the AF-HPLC HR-MS workflow established in this study is in line with previously published assays developed for the untargeted feed- and food authenticity analyses (Belghit et al., 2019), in the next step, we assessed if this approach also is suitable for the targeted identification of insect samples using spectral library matching (SLM) and insect-specific marker peptides and marker proteins, respectively.

For creating the SLs, both MF-HPLC QTOF and AF-HPLC HR-MS data were used; each library contained an average of 12,617 spectra (MF-HPLC QTOF workflow) and 9433 spectra (AF-HPLC HR-MS workflow). Samples from both datasets were matched against these insect spectra reference libraries (cross-matching). After spectra matching of the samples to both libraries, it was found that the best matching spectra originated from samples of the same insect species as the respective library (Supplementary Table 4). As previously shown for mammals and fish (Nessen et al., 2016; Ohana et al., 2016; Varunjikar et al., 2022; Wulff et al., 2013), the spectra library against which the highest number of matching spectra are acquired can be used to determine the identity of the samples (Supplementary Table 4). In both datasets (MF-HPLC QTOF and AF-HPLC HR-MS), BSF libraries yielded the highest number of spectra when matching against spectra from BSF samples (Fig. 2A and B and Supplementary Table 4). Similarly, for HC, LW, and YM libraries, the best matches were obtained from HC, LW and YM samples, respectively in both datasets (Fig. 2A and B and Supplementary Table 4). Surprisingly, all LW samples showed relatively high spectral hits against the YW library and *vice versa*; this could be explained by the relatedness of the insect species belonging to the same order and family (Coleoptera-Tenebrionidae). A single MW sample included in the presented study had relatively low hits against any of the other libraries, with the most hits against the LW library; this could be explained in parts by the "phylogeny" obtained by compareMS2 in which the MW-19 sample clustered closely with other LW samples (Fig. 1A and B). When working with the detection of closely related fish-species, in mixes, we found that using SLM, it was difficult to distinguish cod and haddock, which both are members of the Gadidae family (Varunjikar et al., 2022). We

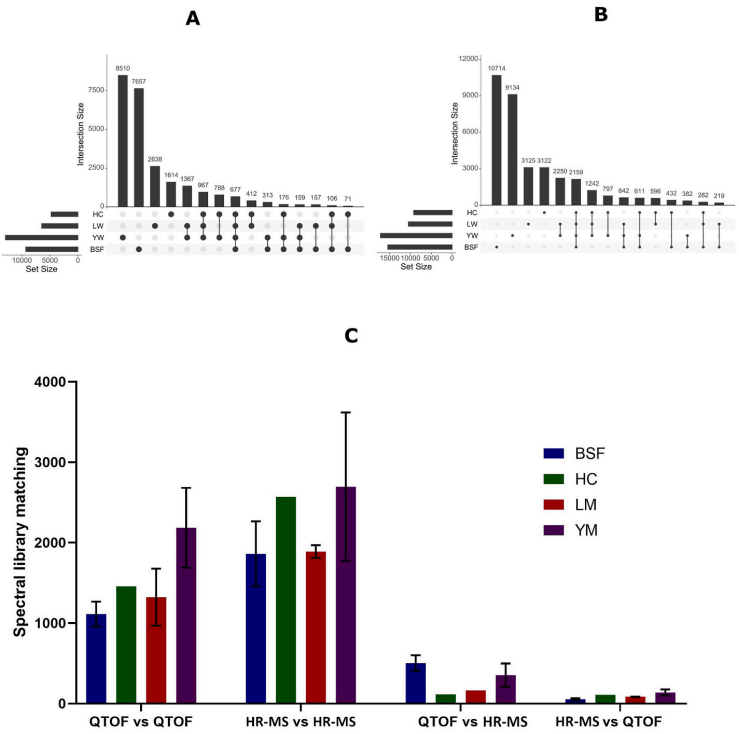


Fig. 2. (A) SpectraST output of library matching indicating MF-HPLC QTOF spectra-specific to insect species BSF, YW, LW, and HC. The detected species-specific spectra can be potential markers for species identification using the untargeted approach. Note: for Morio warm SL was created but to further evaluate species-specific marker additional samples were not available. (B) SpectraST output of library matching indicating HR-MS spectra specific to insect species BSF, YW, LW, and HC. The detected species-specific spectra can be potential markers for species identification using the untargeted approach. Note: for Morio warm SL was created but to further evaluate species-specific marker additional samples were not available. (C) Average number of SLs matching for each analysis; n = 8, 1, 4, and 5 for BSF, HC, LM, and YM, respectively. QTOF vs QTOF: data collected from QTOF and library created on QTOF; HR-MS vs HR-MS: data collected from HR-MS and library created on HR-MS; QTOF vs HR-MS: data collected from QTOF and library created on HR-MS; HR-MS vs QTOF: data collected from HR-MS and library created on QTOF.

therefore speculate that when using SLM, also for closely related insect species from e.g., the Coleoptera-Tenebrionidae family (i.e., YW, LW, and MW), this might be the case in mixed samples.

The compatibility of MF-HPLC QTOF and AF-HPLC HR-MS for building SLs and matching was evaluated by matching acquired data (Fig. 2C). The results of SL matching indicated that the highest number of matches (10–20%) to the SLs were acquired when the libraries were built on the same MS instrument as the query sample. An overview of the spectral matching in Fig. 2C, suggests that the higher number of spectral hits were reported when libraries were built on HR-MS and query data were run on QTOF instrument compared to libraries built on QTOF and queries ran on HR-MS matching. Overall, these findings are consistent with previous work performed on flatfish and other fish species where the highest match was with respective species and closely related species (Nessen et al., 2016; Ohana et al., 2016; Varunjikar et al., 2022; Wulff et al., 2013).

Taken together our data indicate that SL created based on data obtained on MF and AF -HPLC coupled to HR-MS or QTOF instruments can be used for the detection and identification of insect species in food and feed mixtures. The data underlying the analyses presented here were made publicly available on massIVE (massIVE ID: MSV000083737 and massIVE ID: MSV000088034) and can in future be further tested with a larger number of samples for evaluating the robustness of the method.

3.3. Insect protein identification and marker detection

In addition to the spectra matching approaches presented in the previous section, in the present study, we also performed a classic reference proteome dependent bottom-up proteomics data analysis. While this approach is commonly used, is important to note that to date only a few insect-specific reference proteomes exist in public databases and most of the entries are unreviewed (Table 1). Especially for BSF, the UniProtKB database comprises exclusively of unreviewed sequences (1 reviewed and 17,593 unreviewed sequences, accessed on July 2021) and protein identifications at this moment in time might not be very precise. Due to these challenges, the SL-based approach presented above would be beneficial for insect species identification, as previously proposed (Belghit et al., 2019). Regardless, for a comparison with other insect focused studies in the literature, in addition to the SL-based insect identification, we also performed a classic protein identification analysis using both the previously published MF-HPLC QTOF data (massIVE ID: MSV000083737) and the AF-HPLC HR-MS data created in the present study (massIVE ID: MSV000088034).

Originally, a proteoQC based workflow was used to analyse MF-HPLC QTOF data (massIVE ID: MSV000083737). As was the case in the present study, the analyses by (Belghit et al., 2019) also revealed that the rate of protein identification is directly dependent on the size of the UniProtKB database for the insect species in question. For protein identification and species-specific marker detection spectra acquired from both MF-HPLC QTOF and AF-HPLC HR-MS workflows were searched against reviewed sequences from all species of arthropods. A similar approach has been used previously for analyses of non-model species whose reference proteomes are incomplete or not yet available (Francis et al., 2020; Nessen et al., 2016; Varunjikar et al., 2022; Wulff et al., 2013).

Re-analyses of the MF-HPLC QTOF insect dataset (massIVE ID: MSV000083737) generated by Belghit et al. (2019) identified 4745 proteins. The AF-HPLC HR-MS data generated in the present study (massIVE ID: MSV000088034) yielded 4147 protein identifications suggesting that the AF-HPLC HR-MS setup established here, yields result comparable to those obtained previously (Fig. 3A). While comparable in relation to the total number of proteins identified, further analysis of the

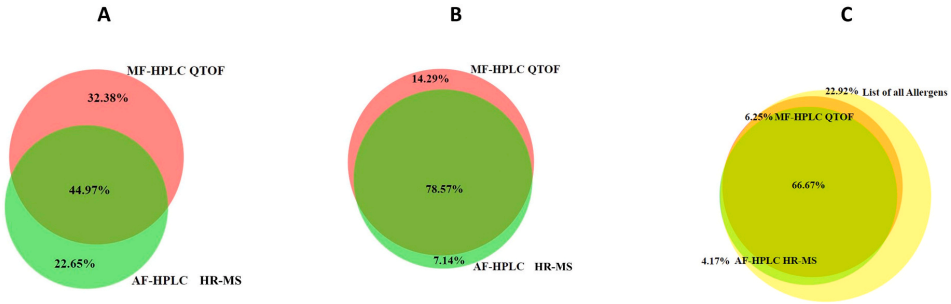


Fig. 3. (A) Insect protein identification; Venn diagrams comparing the percentages of proteins detected in 19 insect meal sample using AF-HPLC HR-MS and MF-HPLC QTOF. (B) Insect marker detection; Venn diagrams comparing the percentages of species-specific proteins detected in 19 insect meal sample using AF-HPLC HR-MS and MF-HPLC TOF (78% proteins were common in both the dataset) (C) Insect allergen detection; Venn diagrams comparing the number of allergens detected in 19 insect meal sample using MF-HPLC QTOF and AF-HPLC HR-MS workflows. Heatmaps illustrating the allergens identification using.

protein data revealed that less than half of the identified proteins (a total of 2758; ~45%) were consistently detected in both datasets; 1986 (32%) and 1389 (22%) proteins were specific to the MF-HPLC QTOF and AF-HPLC HR-MS datasets, respectively (Fig. 3A). A possible reason for the observed difference in protein identification between the two sample analysis workflows can, as was shown previously (Kalli et al., 2014; Rasinger et al., 2016), be the different type of instrument or the different HPLC and MS parameters used. Furthermore, protein extraction protocols also have been shown to affect proteomic profile descriptions (Belghit et al., 2019; Bose et al., 2021; Marbaix et al., 2016; Rasinger et al., 2016). Therefore, to minimize effects of sample preparation, instrument and analysis settings, for future MS-based analyses and differentiation of insects in feed and food, standardized procedures should be established and ideally, be made available in standard operating procedures (SOP) as is the case for example, for the qPCR-based analyses of processed animal proteins (PAP) (European Commission, 2013/51; Olsvik et al., 2017).

Following protein identification, AF-HPLC HR-MS and MF-HPLC QTOF data were compared on species levels (Fig. 4A and B). The results show that samples from the same insect species were grouped together in hierarchical clustering analyses which were performed on MS data passing statistical significance thresholds in a grouped comparison analysis (Qlucore Omics Explorer, $q < 0.1$, Supplementary Table 5). Most of the samples from the Coleoptera family were grouped in the heatmap except for two YW samples analysed using HR-MS workflow; unlike in the compareMS2 output, these were placed on a separate branch of the heatmap (Fig. 4B). Some insect samples used in this study were defatted and processed differently which could have affected protein extractions and protein inference. The heatmap shown in Fig. 4A and B also suggest that there were ~19 proteins with high expression levels in BSF samples when compared to other samples (i.e., LW, YW, MW, and HC). Also, from the Coleoptera family, 21 proteins were displaying different expression levels in YW, LW, and MW compared to BSF. A possible explanation for the overrepresentation of BSF specific proteins could be that the database used for spectra peptide matching and protein inference comprises Arthropoda protein sequences which are dominated by *Drosophila melanogaster* (fruit fly) entries. The latter belongs to the same order as BSF (i.e., Diptera) and therefore the protein matches might be higher; this also was observed in a study by Francis et al., 2020 where proteomics analyses were used for edible insect fingerprinting in novel food. .

To mine for potential species-specific marker proteins for the detection of insects in food and feed, we focused on proteins consistently detected in both AF-HPLC HR-MS and MF-HPLC QTOF data (Fig. 3B). The analysis of the AF-HPLC HR-MS and MF-HPLC QTOF data suggests that for YW, the larval cuticle protein A2B could be a potential marker for species identification (Supplementary Figs. 4A and B; Supplementary Table 6). For HC, cytochrome *c* oxidase (mitochondrial) could be a potential marker protein for species identification given that it was detected only in HC samples (i.e. HC-17 in MF-HPLC QTOF data and both HC-17 and 18 in AF-HPLC HR-MS data) (Supplementary Figs. 4A and B). While further analyses are warranted to confirm that the proteins described here indeed are species-specific, the data provided in this study can be used as the basis to explore the development of quantitative standard reaction monitoring (SRM) assays for the species-specific identification of insects in food and feed as recently demonstrated for PAP identification in animal feed (Lecrenier et al., 2021; Marbaix et al., 2016; Steinhilber et al., 2018a,b; 2019). This work could complement efforts recently reported in a study using a peptidomics approach based on a combination of high-resolution untargeted and targeted species-specific markers for BSF and LM (Leni, Prandi, et al., 2020).

3.4. Detecting allergen in insect species of interest

In addition to the eight insect species permitted to be used as PAP in feed (European Commission, 2017/893), the European Commission recently authorised the marketing of dried yellow mealworms for human consumption (European Comission, 2021/882) and a favourable opinion the placing on the market of house cricket (*Acheta domesticus*) as a novel food was issued by EFSA (EFSA NDA panel, 2021b). Concerning the consumption of the house cricket, EFSA identified no other safety concerns than allergenicity and in a recent review on edible insects and food safety, it was highlighted that extensive allergenic risk assessments would be required before the safe introduction edible insects in the food market were (Ribeiro et al., 2021). In the light of the potential allergenic risk insects mays pose, it was assessed in the present study if untargeted proteomics data acquired from both MF-HPLC QTOF and AF-HPLC HR-MS also can be successfully screened for the presence of relevant known food allergens (Supplementary Table 7).

From the list of 48 allergenic proteins, 37 were detected in both datasets and 32 were consistently detected in both MF-HPLC QTOF and AF-HPLC HR-MS data (Fig. 3C). Using a proteomic and bioinformatic approach, Barre et al., 2021 identified a comparable number of pan-allergens (46 proteins) in house crickets (*Acheta domesticus*) (EFSA NDA panel, 2021b). Among the four families of allergens in silk moth (*Bombyx mori*) which is a close relative of the selected insect species in this study, arginine kinase (Q2F5T5), low molecular mass lipoproteins (Q00802 and Q00801), and tropomyosin 1 and 2 (Q1HPU0 and Q1HPQ0) proteins were detected in the acquired data from both instruments. Tropomyosin is a known IgE-binding protein and cross-reactivity of HC tropomyosin with shrimp tropomyosin was demonstrated with ELISA in a recent study (De Marchi, Wangorsch, &

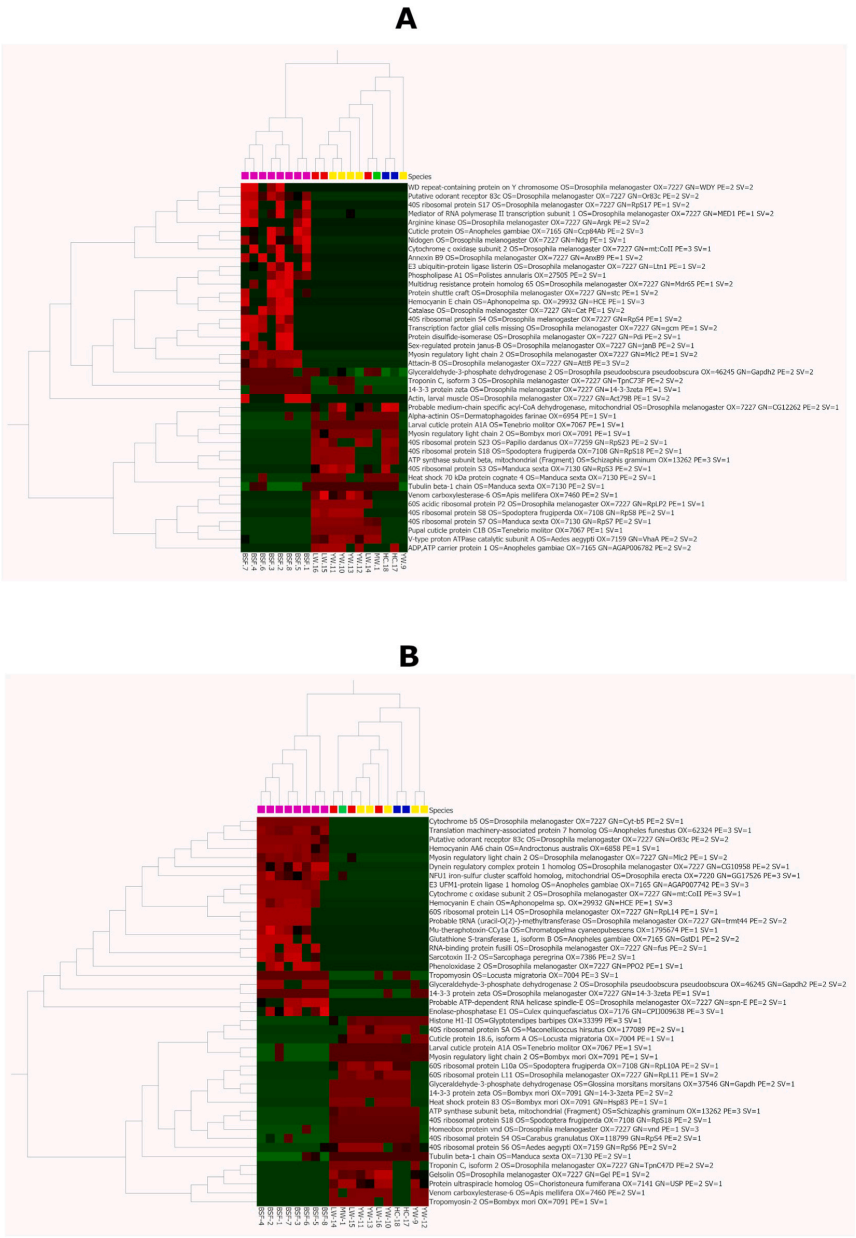


Fig. 4. Insect protein identification; Heatmaps illustrating the protein identification using (A) MF-HPLC QTOF and (B) AF-HPLC HR-MS workflows, based on TPP identification using Comet search engine and Arthropoda reviewed protein as reference database. Hierarchical clustering (HC) of samples and differentially expressed proteins where group comparison was performed using Omics Explorer V3.6. The heatmap represents expressed proteins withing each measured samples; red represents expressed proteins and green represent absent or unexpressed proteins. Note that the proteins might not be from the same species as studied given that most of the proteins for the species of interest in this study were unreviewed. So, they are from different insect species but exhibit similarity to the species BSF, YW, LW, HC, and MW. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

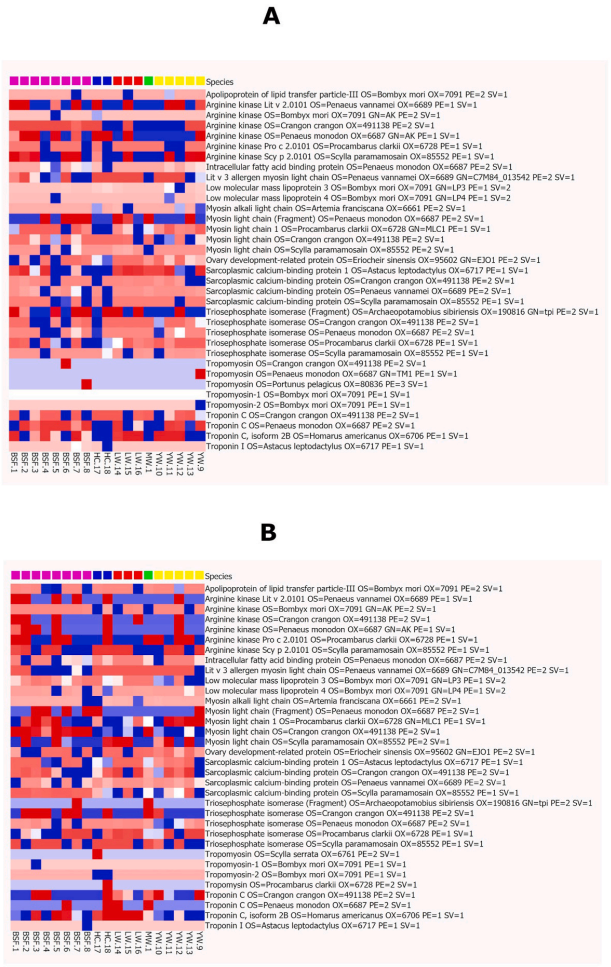


Fig. 5. Insect allergen detection. Heatmaps illustrating the allergens identification using (A) MF-HPLC QTOF and (B) AF-HPLC HR-MS workflows. Heat map representation of 37 allergens across the 19 insect samples. As explained in the insert the pink, blue, red, green, and yellow rectangles represent BSF, HC, LW, MW and YM, respectively. Each line in the heat map represents an allergen. The deeper red colour, the higher is the allergen present in the respective sample; similarly, the deeper the blue colour, the lower is the allergen present in the respective sample as illustrated in the figure insert. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Zoccatelli, 2021). In the current study, tropomyosin-2 from silk moth (Q1HPQ0) was consistently detected in both datasets (Fig. 5A and B) across all species with one exception; in HC and a single replicate of YW (YW-9) samples, Q1HPQ0 was detected only in MF-HPLC QTOF and AF-HPLC HR-MS data, respectively. Interestingly, tropomyosin also was flagged as a key pan-allergen present in the house cricket when highlighting safety concerns related to the consumption of this novel food (EFSA NDA panel, 2021b). Other allergenic proteins detected in the insect samples were arginine kinase and troponin C (from different Arthropoda species) that were present in BSF, HC, LW and YW samples. In a recent study focusing on arginine kinase (Bose et al., 2021), it was shown that protein extraction protocols can affect the quantitation of allergens from cricket samples. It could therefore be possible that the varying profile of allergens detected in the selected insect samples presented here can be attributed to differences in sample processing, instrument selection, and protein extraction protocol (Broekman et al., 2015; De Marchi, Wangorsch, & Zoccatelli, 2021; Pali-Schöll et al., 2019; Van Broekhoven et al., 2016). In other words, like proteomics-based marker detection for insect species differentiation in food and feed, also allergen detection could benefit from standardized procedures summarized in SOPs for the respective purpose (Bose et al., 2021; Marbaix et al., 2016).

The tentative screening for predicted allergens in data obtained from basic MF- and AF- HPLC HR-MS workflows commonly used in regulatory laboratories highlighted the potential of these routine tools for ensuring the safety of novel foods and feeds. What is more, the data created here (massIVE ID: MSV000088034) and by Belghit et al. (2019) (massIVE ID: MSV000083737), lays the foundation for future work focusing on spectra matching in which SL and *in-silico* assessments can be combined for allergen detection as recently exemplified by (FitzGerald et al., 2020; Leni et al., 2020).

4. Conclusion

The combination of standard MS instruments commonly available in regulatory laboratories combined with freely available open-source data analysis approaches allow for implementation of untargeted proteomics assays for food and feed safety research in routine settings. The AF-HPLC HR-MS workflow and associated bioinformatics approaches presented here can be a useful toolset suitable for the detection and differentiation

of insects in feed and food and complement existing methods currently used in the market. The approaches presented and the data generated in the present study and made available in a public repository (massIVE ID: MSV000088034) also were found to be suitable for allergen detection in insect species.

CRediT authorship contribution statement

Madhushri S. Varunjikar: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing –review & original draft. **Ikram Belghit:** Conceptualization, Supervision, Investigation, Methodology, Project administration, Writing – original draft, Writing –review & original draft. **Jennifer Gjerde:** Conceptualization, Investigation, Methodology, review & original draft. **Magnus Palmblad:** Conceptualization, Data curation, Investigation, Methodology, review &original draft. **Eystein Oveland:** Conceptualization, Methodology, Data curation, Investigation, Writing –review & original draft. **Josef D. Rasinger:** Conceptualization, Supervision, Data curation, Investigation, Methodology, Project administration, Software, Writing –review & original draft.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

We are very much thankful for support and guidance of Dr. Marc Dieu from University of Namur, mass spectrometry facility (MaSUN), rue de Bruxelles 61, B-5000 Namur, Belgium and for the compareMS2 graphical user interface by Rob Marissen (CPM). The contents of the graphical abstract were created using BioRender.com.

Funding was provided by the Institute of Marine Research (Programledelse Fiskeernæring, MultiOmicsTools project).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodcont.2022.108888.

References

Aguilera, J., Aguilera-Gomez, M., Barrucci, F., Cocconcelli, P. S., Davies, H., Denslow, N., Lou Dorne, J., Grohmann, L., Herman, L., Hogstrand, C., Kass, G. E. N., Kille, P., Kleter, G., Nogué, F., Plant, N. J., Ramon, M., Schoonjans, R., Waigmann, E., & Wright, M. C. (2018). EFSA scientific colloquium 24 – 'omics in risk assessment: State of the art and next steps. *EFSA Supporting Publications, 15*(11). https://doi.org/10.2903/sp.efsa.2018.EN-1512

Barre, A., Pichereaux, C., Simplicien, M., Burlet-Schiltz, O., Benoist, H., & Rougé, P. (2021). A proteomic-and bioinformatic-based identification of specific allergens from edible insects: Probes for future detection as food ingredients. https://doi.org/10.3390/foods10020280.

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, 128*, 108183. https://doi.org/10.1016/j.foodcont.2021.108183

Bose, U., Broadbent, J. A., Juhász, A., Karnaneedi, S., Johnston, E. B., Stockwell, S., Byrne, K., Limviphuvadh, V., Maurer-Stroh, S., Lopata, A. L., & Colgrave, M. L. (2021). Protein extraction protocols for optimal proteome measurement and arginine kinase quantitation from cricket Acheta domesticus for food safety assessment. *Food Chemistry, 348*, 129110. https://doi.org/10.1016/J.FOODCHEM.2021.129110

Broekman, H., Knulst, A., den Hartog Jager, S., Monteleone, F., Gaspari, M., de Jong, G., Houben, G., & Verhoeckx, K. (2015). Effect of thermal processing on mealworm allergenicity. *Molecular Nutrition & Food Research, 59*, 1855–1864. https://doi.org/10.1002/mnfr.201500138

European Commission Regulation 2021/1372. (2021a). Commission Regulation (EU) 2021/1372 of 17 August 2021 amending Annex IV to Regulation (EC) No 999/2001 of the European Parliament and of the Council as regards the prohibition to feed non-ruminant farmed animals, other than Fur animals, with protein deri. *Official Journal of the European Union, 64*, 1–21. http://data.europa.eu/eli/reg/2021/1372/oj.

European Commission Regulation 2021/882. (2021b). Commission implementing regulation (EU) 2021/882 of 1 June 2021 authorising the placing on the market of dried Tenebrio molitor larva as a novel food under regulation (EU) 2015/2283 of the European parliament and of the council, and amending commission im. *Official Journal of the European Union, 64*(L194), 16–20. http://data.europa.eu/eli/reg_impl/2021/882/oj.

compareMS2 GUI. https://github.com/524D/compareMS2, (2021).

Daniso, E., Tulli, F., Cardinaletti, G., Cerri, R., & Tibaldi, E. (2020). Molecular approach for insect detection in feed and food: The case of gryllodes sigillatus. *European Food Research and Technology, 246*(12), 2373–2381. https://doi.org/10.1007/S00217-020-03573-1

De Marchi, L., Mainente, F., Leonardi, M., Scheurer, S., Wangorsch, A., Mahler, V., Pilolli, R., Sorio, D., & Zoccatelli, G. (2021). Allergenicity assessment of the edible cricket Acheta domesticus in terms of thermal and gastrointestinal processing and IgE cross-reactivity with shrimp. *Food Chemistry, 359*, 129878. https://doi.org/10.1016/J.FOODCHEM.2021.129878

De Marchi, L., Wangorsch, A., & Zoccatelli, G. (2021). Allergens from edible insects: Cross-reactivity and effects of processing. https://doi.org/10.1007/s11882-021-01012-z/.

Debode, F., Marien, A., Gérard, A., Francis, F., Fumière, O., & Berben, G. (2017). Development of real-time PCR tests for the detection of Tenebrio molitor in food and feed. *Food Additives & Contaminants: Part A, 34*(8), 1421–1426. https://doi.org/10.1080/19440049.2017.1320811

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

EFSA NDA panel. (2021). Safety of dried yellow mealworm (Tenebrio molitor larva) as a novel food pursuant to Regulation (EU) 2015/2283. *EFSA Journal, 19*(1), 1–29. https://doi.org/10.2903/j.efsa.2021.6343

EFSA NDA panel. (2021b). Safety of frozen and dried formulations from whole house crickets (Acheta domesticus) as a Novel food pursuant to Regulation (EU) 2015/2283. *EFSA Journal, 19*(8). https://doi.org/10.2903/j.efsa.2021.6779

European Commission. (2013). COMMISSION REGULATION (EU) No 51/2013 of 16 January 2013 amending Regulation (EC) No 152/2009 as regards the methods of analysis for the determination of constituents of animal origin for the official control of feed. *Official Journal of the European Union*, 33–43, 2013, L 20/33) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R0051&from=EN.

European Commission. (2017). *Commission regulation (EU) 2017/893* (pp. 1–25). Official Journal of the European Union, 2017 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0893&rid=1.

Fao, I. F. A. D., Unicef, W. F. P., & Who. (2018). The state of food security and nutrition in the world 2018. Building climate resilience for food security and nutrition. Rome, FAO. Licence: CC BY-NC-SA 3.0 IGO. In *Building climate resilience for food security and nutrition*. https://doi.org/10.1093/cjres/rst006

FitzGerald, R. J., Cermeño, M., Khalesi, M., Kleekayai, T., & Amigo-Benavent, M. (2020). Application of in silico approaches for the generation of milk protein-derived bioactive peptides. *Journal of Functional Foods, 64*, 103636. https://doi.org/10.1016/J.JFF.2019.103636

Francis, F., Doyen, V., Debaugnies, F., Mazzucchelli, G., Caparros, R., Alabi, T., Blecker, C., Haubruge, E., & Corazza, F. (2019). Limited cross reactivity among arginine kinase allergens from mealworm and cricket edible insects. *Food Chemistry, 276*, 714–718. https://doi.org/10.1016/J.FOODCHEM.2018.10.082

Francis, F., Mazzucchelli, G., Baiwir, D., Debode, F., Berben, G., & Caparros Megido, R. (2020). Proteomics based approach for edible insect fingerprinting in novel food: Differential efficiency according to selected model species. *Food Control, 112*, 107135. https://doi.org/10.1016/J.FOODCONT.2020.107135

Garino, C., Zagon, J., & Nesic, K. (2021). Novel real-time PCR protocol for the detection of house cricket (Acheta domesticus) in feed. *Animal Feed Science and Technology, 280*, 115057. https://doi.org/10.1016/J.ANIFEEDSCI.2021.115057

IPIFF. (2021). An overview of the European market of insects as feed. 2020–2021. https://ipiff.org/.

Jenkins, C., & Orsburn, B. (2020). *Standard flow multiplexed proteomics (SFloMPro) – an accessible and cost-effective alternative to NanoLC workflows*. BioRxiv. https://doi.org/10.1101/2020.02.25.964379, 2020, 02.25.964379.

Kalli, A., Smith, G. T., Sweredoski, M. J., & Hess, S. (2014). Evaluation and optimization of mass spectrometric mode: Focus on LTQ-Orbitrap Mass analyzers. *Journal of Proteome Research, 12*(7), 3071–3086. https://doi.org/10.1021/pr3011588.Evaluation

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

Kelstrup, C. D., Jersie-Christensen, R. R., Batth, T. S., Arrey, T. N., Kuehn, A., Kellmann, M., & Olsen, J. V. (2014). Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field orbitrap mass spectrometer. *Journal of Proteome Research, 13*(12), 6187–6195. https://doi.org/10.1021/pr500985w

Köppel, R., Schum, · R.afael, Habermacher, · M., Sester, C., Lucia, ·, Piller, E., Meissner, · S., & Pietsch, K. (2019). Multiplex real-time PCR for the detection of insect DNA and

determination of contents of Tenebrio molitor, Locusta migratoria and Achaeta domestica in food. https://doi.org/10.1007/s00217-018-03225-5, 245, 559-567.

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). Inter-laboratory study on the detection of bovine processed animal protein in feed by LC-MS/MS-based proteomics. *Food Control, 125*(November 2020), 1–7. https://doi.org/10.1016/j.foodcont.2021.107944

Lecrenier, M. C., Planque, M., Dieu, M., Veys, P., Saegerman, C., Gillard, N., & Baeten, V. (2018). A mass spectrometry method for sensitive, specific and simultaneous detection of bovine blood meal, blood products and milk products in compound feed. *Food Chemistry, 245*(September), 981–988. https://doi.org/10.1016/j.foodchem.2017.11.074

Lenčo, J., Vajrychová, M., Pimková, K., Prokšová, M., Benková, M., Klimentová, J., Tambor, V., & Soukup, O. (2018). Conventional-flow liquid chromatography-mass spectrometry for exploratory bottom-up proteomic analyses. *Analytical Chemistry, 90*(8), 5381–5389. https://doi.org/10.1021/acs.analchem.8b00525

Leni, G., Prandi, B., Varani, M., Faccini, A., Caligiani, A., & Sforza, S. (2020). Peptide fingerprinting of Hermetia illucens and Alphitobius diaperinus: Identification of insect species-specific marker peptides for authentication in food and feed. *Food Chemistry, 320*, 126681. https://doi.org/10.1016/J.FOODCHEM.2020.126681

Leni, G., tedeschi, tullia, faccini, A., pratesi, federico, folli, claudia, puxeddu, ilaria, Migliorini, paola, Gianotten, natasja, Jacobs, J., Depraetere, S., caligiani, A., & Sforza, S. (2020). Shotgun proteomics, in-silico evaluation and immunoblotting assays for allergenicity assessment of lesser mealworm, black soldier fly and their protein hydrolysates. https://doi.org/10.1038/s41598-020-57863-5.

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. https://doi.org/10.1021/acs.jafc.6b00064

Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., & Horning, S. (2011). Mass spectrometry-based proteomics using Q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Molecular & Cellular Proteomics, 10*(9), 1–11. https://doi.org/10.1074/mcp.M111.011015

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

Pali-Schöll, I., Meinlschmidt, P., Larenas-Linnemann, D., Purschke, B., Hofstetter, G., Rodríguez-Monroy, F. A., Einhorn, L., Mothes-Luksch, N., Jensen-Jarolim, E., & Jäger, H. (2019). Edible insects: Cross-recognition of IgE from crustacean- and house dust mite allergic patients, and reduction of allergenicity by food processing. *World Allergy Organization Journal, 12*(1), 100006. https://doi.org/10.1016/J.WAOJOU.2018.10.001

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

Pedersen. (2021). MSGFplus: An interface between R and MS-GF+. https://doi.org/10.18129/B9.bioc.MSGFplus.

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

Ribeiro, J. C., Sousa-Pinto, B., Fonseca, J., Fonseca, S. C., & Cunha, L. M. (2021). Edible insects and food safety: Allergy. *Journal of Insects as Food and Feed, 7*(5), 833–847. https://doi.org/10.3920/jiff2020.0065

Rumbos, C. I., & Athanassiou, C. G. (2021). The superworm, zophobas morio (Coleoptera:tenebrionidae): A 'sleeping giant' in nutrient sources. *Journal of Insect Science, 21*(2). https://doi.org/10.1093/jisesa/ieab014

Sentandreu, M. A., & Sentandreu, E. (2011). Peptide biomarkers as a way to determine meat authenticity. *Meat Science, 89*(3), 280–285. https://doi.org/10.1016/J.MEATSCI.2011.04.028

Steinhilber, A. E., Schmidt, F. F., Naboulsi, W., Planatscher, H., Niedzwiecka, A., Zagon, J., Braeuning, A., Lampen, A., Joos, T. O., & Poetz, O. (2018a). Mass spectrometry-based immunoassay for the quantification of banned ruminant processed animal proteins in vegetal feeds. *Analytical Chemistry, 90*(6), 4135–4143. https://doi.org/10.1021/acs.analchem.8b00120

Steinhilber, A. E., Schmidt, F. F., Naboulsi, W., Planatscher, H., Niedzwiecka, A., Zagon, J., Braeuning, A., Lampen, A., Joos, T. O., & Poetz, O. (2018b). 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

Steinhilber, A. E., Schmidt, F. F., Naboulsi, W., Planatscher, H., Niedzwiecka, A., Zagon, J., Braeuning, A., Lampen, A., Joos, T. O., & Poetz, O. (2019). Application of mass spectrometry-based immunoassays for the species- A nd tissue-specific quantification of banned processed animal proteins in feeds [Research-article]. *Analytical Chemistry, 91*(6), 3902–3911. https://doi.org/10.1021/acs.analchem.8b04652

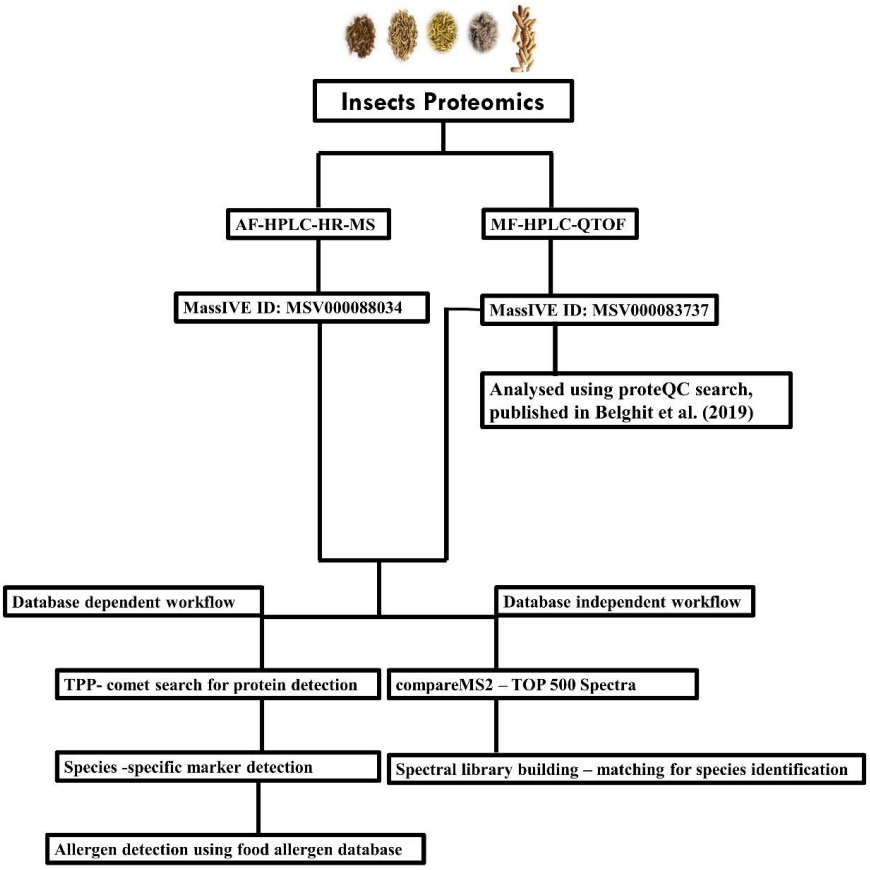
Szabó, D., Schlosser, G., Vékey, K., Drahos, L., & Révész, Á. (2021). Collision energies on QTof and Orbitrap instruments: How to make proteomics measurements comparable? *Journal of Mass Spectrometry, 56*(1), 1–12. https://doi.org/10.1002/jms.4693

Van Broekhoven, S., Bastiaan-Net, S., De Jong, N. W., & Wichers, H. J. (2016). Influence of processing and in vitro digestion on the allergic cross-reactivity of three mealworm species. *Food Chemistry, 196*, 1075–1083. https://doi.org/10.1016/J.FOODCHEM.2015.10.033

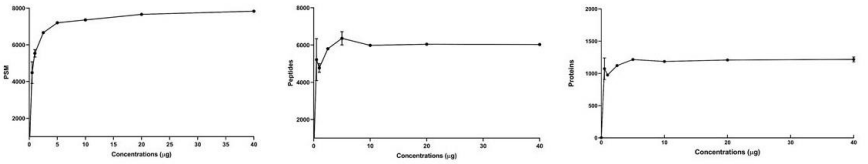
Van Der Plas-Duivesteijn, S. J., Klychnikov, O., Ohana, D., Dalebout, H., Van Veelen, P. A., De Keijzer, J., Nessen, M. A., Van Der Burgt, Y. E. M., Deelder, A. M., & Palmblad, M. (2016). *Differentiating samples and experimental protocols by direct comparison of tandem mass spectra*. https://doi.org/10.1002/rcm.7494

Varunjikar, M. S., Moreno-Ibarguen, C., Andrade-Martinez, J. S., Tung, H.-S., Belghit, I., Palmblad, M., Olsvik, P. A., Reyes, A., Rasinger, J. D., & 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*, 108417. https://doi.org/10.1016/j.foodcont.2021.108417

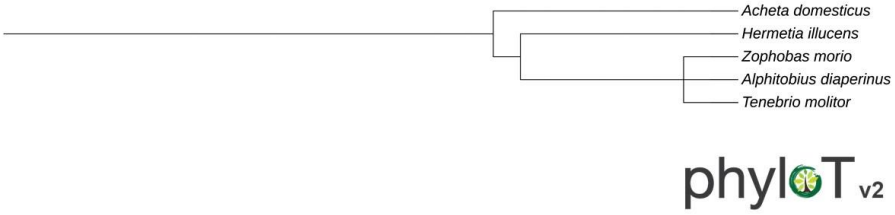
Wulff, T., Nielsen, M. E., Deelder, A. M., Jessen, F., & Palmblad, M. (2013). Authentication of fish products by large-scale comparison of tandem mass spectra. *Journal of Proteome Research, 12*(11), 5253–5259. https://doi.org/10.1021/pr4006525



Supplementary Figure 1: Insect proteomics workflow in this study and previous study of Belghit et al. (2019). Abbreviations: AF-HPLC: analytical flow High performance liquid chromatography; HR-MS: High resolution- mass spectrometry Orbitrap; MF-HPLC: microflow High performance liquid chromatography; TOF: time-of-flight; TPP: Trans-proteomics pipeline

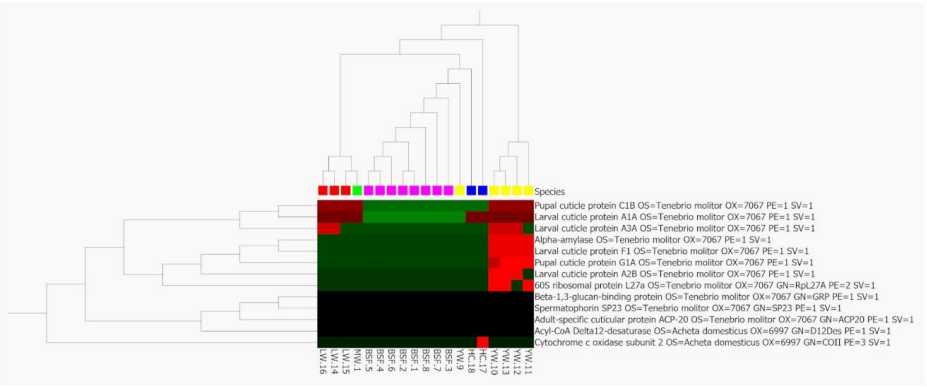


Supplementary Figure 2: Number of peptide spectrum matches (PSMs), unique identified peptides, and identified proteins with increasing concentrations of HeLa digests using the optimised AF-HPLC QE method. Values are means, with their standard deviation represented by vertical bars (n = 2).

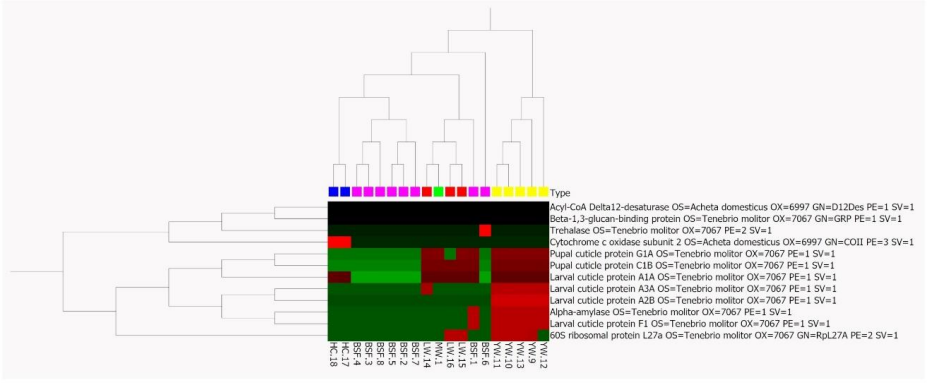


Supplementary Figure 3: Phylogeny of insect species used in the study constructed using phyloT v2 phylogenetic tree generator based on NCBI or GTD taxonomy)

A



B



Supplementary Figure 4: Insect marker detection. Heatmaps illustrating the identified proteins which were from the species involved in this study, using (A) MF-HPLC QTOF and (B) AF-HPLC QE workflows. Reviewed proteins were from YW (Tenebrio molitor) and HC (Acheta domesticus). Two species-specific proteins were detected in the dataset and peptides of this proteins can be used as potential marker for species identification.

Paper III Supplementary Tables can be downloaded from here https://ars.els-cdn.com/content/image/1-s2.0-S0956713522000810-mmc1.xlsx

Tables-S	Title	Legends
1	Insect protein samples included in the study, with the *Latin name*, order and the family belongings	BSF= black soldier fly; YM= yellow mealworm; LM= lesser mealworm; HC= house cricket; MW= morio worm, No = number of samples (previously published in Belghit et al., 2019)
2	Optimisation of HPLC-gradient length and MS2 parameters on HR-MS	settings used for LC-MS/MS HR-MS optimization to obtain highest number of MS2 spectra and matches; MS2 parameters were base on Kalli et al. 2013 and modified as per the instrument suitability to obtain higher MS2 spectra
3	Optimisation of Hela cell concentration	various concentrations of Hela cell digest were injected to find optimal concentration of injection
4	SpectraST output table indicating spectral matches of samples to four insect species spectral libraries	Number of matches to libraries created on both instruments were reported (data used for figure 3C); TOF vs TOF: data collected from TOF and library created on TOF; HR-MS vs HR-MS: data collected from HR-MS and library created on HR-MS; TOF vs HR-MS: data collected from TOF and library created on HR-MS; HR-MS vs TOF: data collected from HR-MS and library created on TOF.
5	Proteins detected in the obtained proteomics data	Proteins were detected in the insect samples along with probability per sample (used statistics and heatmap in figure 4)
6	Proteins specific to insect species	Reviewed proteins specific to insect species of interest as potential marker (used statistics and heatmap in figure 5)
7	List of food Allergen detection	Allergens detection output of TPP from both TOF and HR-MS data (data used for figure 6)

Paper IV

Varunjikar M.S., Bøhn T., Sanden M., Pineda-Pampliega J., Belghit I., Palmblad M., Rasinger J.D.

Proteomics analyses of herbicide-tolerant genetically modified, conventionally, and organically farmed soybean seeds

Submitted (2022)

IV

Paper V

Marissen, R., **Varunjikar, M. S.**, Laros, J. F., Rasinger, J. D., Neely, B. A., & Palmblad, M.

compareMS2 2.0: An Improved Software for Comparing Tandem Mass Spectrometry Datasets

Journal of Proteome Research (2022)

V

Journal of proteome research

pubs.acs.org/jpr

Technical Note

compareMS2 2.0: An Improved Software for Comparing Tandem Mass Spectrometry Datasets

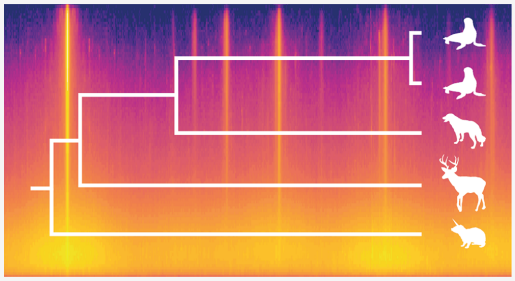
Rob Marissen, Madhushri S. Varunjikar, Jeroen F. J. Laros, Josef D. Rasinger, Benjamin A. Neely, and Magnus Palmblad*

Cite This: https://doi.org/10.1021/acs.jproteome.2c00457

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: It has long been known that biological species can be identified from mass spectrometry data alone. Ten years ago, we described a method and software tool, compareMS2, for calculating a distance between sets of tandem mass spectra, as routinely collected in proteomics. This method has seen use in species identification and mixture characterization in food and feed products, as well as other applications. Here, we present the first major update of this software, including a new metric, a graphical user interface and additional functionality. The data have been deposited to ProteomeXchange with dataset identifier PXD034932.



KEYWORDS: *compareMS2, distance metric, molecular phylogenetics, tandem mass spectrometry, quality control*

INTRODUCTION

A decade ago, Palmblad and Deelder[1] first described a method for molecular phylogenetics based on direct comparison of tandem mass spectra. The method has since seen a range of applications, including food[2,3] and feed[4−7] species identification, quality control,[8] and experimental design.[9] Similar works include the DISMS2 library by Rieder and colleagues[10] and MS1-only methods for "sequence-free" phylogenetics reviewed by Downard.[11] Neely and Palmblad[12] recently placed these methods in a larger historical context, going all the way back to the seminal comparison of separated tryptic peptides across species by Zuckerkandl, Jones, and Pauling in 1960.[13] Here, we describe a new and significantly updated version of the original compareMS2 software, with several improvements, including a graphical user interface (GUI) controlling all steps of the analysis and dynamic phylogenetic tree display, a fully symmetric distance metric, and many additional filters and output options, which we describe in this technical note.

METHODS

Symmetric Distance Measure

The original compareMS2 compared two sets of tandem mass spectra, e.g., those resulting from liquid chromatography−tandem mass spectrometry, by scanning one set and for each spectrum finding the best match in the other set (within precursor m/z and retention time tolerances). The results depended on which set was scanned, and the distance metric was only approximately symmetric. compareMS2 2.0 has a perfectly symmetric measure of the distance between sets of tandem mass spectra regardless of order of input. In this section, we describe this modified measure and some of its properties.

Comparing Pairs of Spectra

The comparison between sets of tandem mass spectra starts with the comparison of pairs of spectra. There are many measures of spectral similarity. compareMS2 supports the cosine score (dot product) and spectral angle. By default, compareMS2 uses the cosine score, i.e., the cosine of the angle between the vector representations of the spectra, after normalizing both spectra to unit length:

$$s(a, b) = \frac{a \cdot b}{\|a\| \|b\|} = \cos\theta \quad (1)$$

where θ is the angle between the vector representations of the two spectra. Equation 1 is symmetric in a and b.

Optionally, compareMS2 can first scale spectra to reduce the influence of very intense peaks, e.g., by taking the square or cube root of all intensities. All peaks below a user-defined or

Special Issue: Software Tools and Resources 2023

Received: July 28, 2022

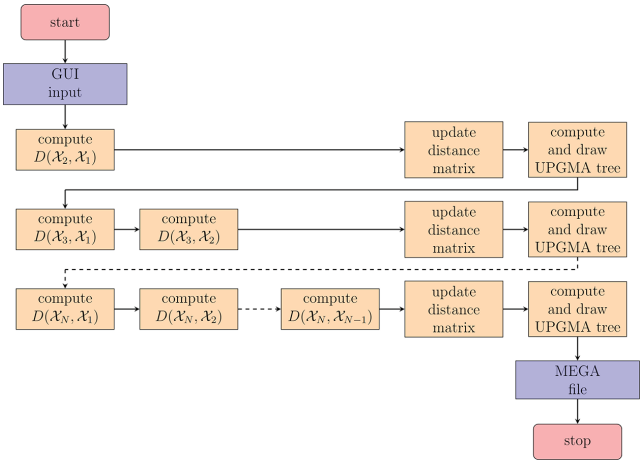


Figure 1. CompareMS2 2.0 workflow, orchestrated by the graphical user interface. After parsing and checking the input parameters, ensuring all files are present and in the correct format, compareMS2 performs $(N^2 - N)/2$ pairwise comparisons of N datasets using the symmetric distance measure described below, or $N^2 - N$ comparisons if the original measure is used. After each row is completed, compareMS2 updates the (strictly triangular) distance matrix and generates a new tree. This allows the user to monitor progress and terminate and restart the run if necessary. If the original measure is used, compareMS2 by default creates both the strictly upper and lower triangular distance matrices (these can be averaged in phylogenetics software such as MEGA).

automatically detected relative or absolute background can also be excluded from the similarity calculation.

Comparing Sets of Spectra

compareMS2 2.0 defines the similarity between two sets of tandem mass spectra, $\mathcal{A}$ and $\mathcal{B}$ as follows. If for a spectrum $a \in \mathcal{A}$ we find a spectrum $b \in \mathcal{B}$ with $s(a,b)$ greater than or equal to a minimum similarity threshold s_{min}, we say that a has a similar spectrum in $\mathcal{B}$. We then define a subset $S_{\mathcal{A}|\mathcal{B}} \subset \mathcal{A}$, given $\mathcal{B}$, of all spectra in $\mathcal{A}$ with at least one similar spectrum in $\mathcal{B}$ as

$$\mathcal{S}_{\mathcal{A}|\mathcal{B}} = \{a \in \mathcal{A} \wedge \exists\ b \in \mathcal{B} | s(a, b) \geq s_{\min}\} \tag{2}$$

and a corresponding subset $S_{\mathcal{B}|\mathcal{A}} \subset$ B as

$$\mathcal{S}_{\mathcal{B}|\mathcal{A}} = \{b \in \mathcal{B} \wedge \exists\ a \in \mathcal{A} | s(b, a) \geq s_{\min}\} \tag{3}$$

We then define a global similarity between sets $\mathcal{A} \neq \emptyset$ and $\mathcal{B} \neq \emptyset$, $S(\mathcal{A},\mathcal{B})$, as the average of the fraction of spectra in $\mathcal{A}$ with at least one similar spectrum in $\mathcal{B}$ and the fraction of spectra in $\mathcal{B}$ with at least one similar spectrum in $\mathcal{A}$:

$$S(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{S}_{\mathcal{A}|\mathcal{B}}|}{2|\mathcal{A}|} + \frac{|\mathcal{S}_{\mathcal{B}|\mathcal{A}}|}{2|\mathcal{B}|} \tag{4}$$

where $|\mathcal{X}|$ denotes the cardinality, the number of elements, in a set $\mathcal{X}$. Though in some use cases it may be meaningful to define the similarity between two empty sets, i.e., LC-MS/MS datasets without tandem mass spectra, or the similarity between an empty and a non-empty set, we have chosen to leave these undefined and have the compareMS2 output reflect this. We believe this makes sense as a dataset without tandem mass spectra usually suggests something went wrong during measurement. Values can always be imputed after the compareMS2 runs, and rows with undefined values in the distance matrix can be excluded in subsequent analyses in most phylogenetic software.

From the symmetry of eq 4, we see that $S(\mathcal{A},\mathcal{B}) = S(\mathcal{B},\mathcal{A})$. We also note that both terms in eq 4 are non-negative, therefore $S(\mathcal{A},\mathcal{B}) \geq 0$. The maximum value of $S(\mathcal{A},\mathcal{B})$ is 1 when all spectra in $\mathcal{A}$ have a similar spectrum in $\mathcal{B}$ and vice versa. The minimum value is 0 when $\mathcal{A}$ and $\mathcal{B}$ have no similar spectra. The smallest positive value of $S(\mathcal{A},\mathcal{B})$ occurs when there is exactly one pair of similar spectra in $\mathcal{A}$ and $\mathcal{B}$:

$$\min\{S(\mathcal{A}, \mathcal{B}) | S(\mathcal{A}, \mathcal{B}) > 0\} = \frac{1}{2|\mathcal{A}|} + \frac{1}{2|\mathcal{B}|} = \frac{|\mathcal{A}| + |\mathcal{B}|}{2|\mathcal{A}||\mathcal{B}|} \tag{5}$$

Finally, we arrive at the global distance measure, $D(\mathcal{A},\mathcal{B})$, which we define as the inverse of $S(\mathcal{A},\mathcal{B})$ minus one when $S(\mathcal{A},\mathcal{B})$ is positive, and as the inverse of half of the smallest positive value of $S(\mathcal{A},\mathcal{B})$ minus one when $S(\mathcal{A},\mathcal{B})$ is zero:

$$D(\mathcal{A}, \mathcal{B}) = \begin{cases} \dfrac{1}{S(\mathcal{A}, \mathcal{B})} - 1 & \text{if } S(\mathcal{A}, \mathcal{B}) > 0 \\ \dfrac{4|\mathcal{A}||\mathcal{B}|}{|\mathcal{A}| + |\mathcal{B}|} - 1 & \text{if } S(\mathcal{A}, \mathcal{B}) = 0 \end{cases} \tag{6}$$

Since $S(\mathcal{A},\mathcal{B})$ is symmetric, $D(\mathcal{A},\mathcal{B})$ is also symmetric. Note that $D(\mathcal{A},\mathcal{B}) \to \infty$ as $|\mathcal{A}| \to \infty \wedge |\mathcal{B}| \to \infty$,and there are no similar spectra in $\mathcal{A}$ and $\mathcal{B}$. In the special case of $\mathcal{A}$ and $\mathcal{B}$ both containing a single spectrum, $D(\mathcal{A},\mathcal{B})$ is 0 if the spectra are similar and 1 otherwise. The definition of the distance between sets with $S(\mathcal{A},\mathcal{B}) = 0$ correspond to $\mathcal{A}$ and $\mathcal{B}$ having a hypothetical half matching spectrum. In most real-world use cases, both $\mathcal{A}$ and $\mathcal{B}$ would contain thousands of spectra.

Two co-directional spectra—spectra whose vector representations differ only by a factor—are considered identical by s. Therefore, datasets containing perfectly co-directional spectra would have a global similarity $S = 1$ and distance $D = 0$. Strictly speaking, D is not a metric in the mathematical sense, as the identity of indiscernibles ($D(\mathcal{A},\mathcal{B}) = 0 \Leftrightarrow \mathcal{A} = \mathcal{B}$) no longer holds after normalizing the spectra. This is by design, as the absolute intensities in a tandem mass spectrum depend not only

https://doi.org/10.1021/acs.jproteome.2c00457
J. Proteome Res. XXXX, XXX, XXX−XXX

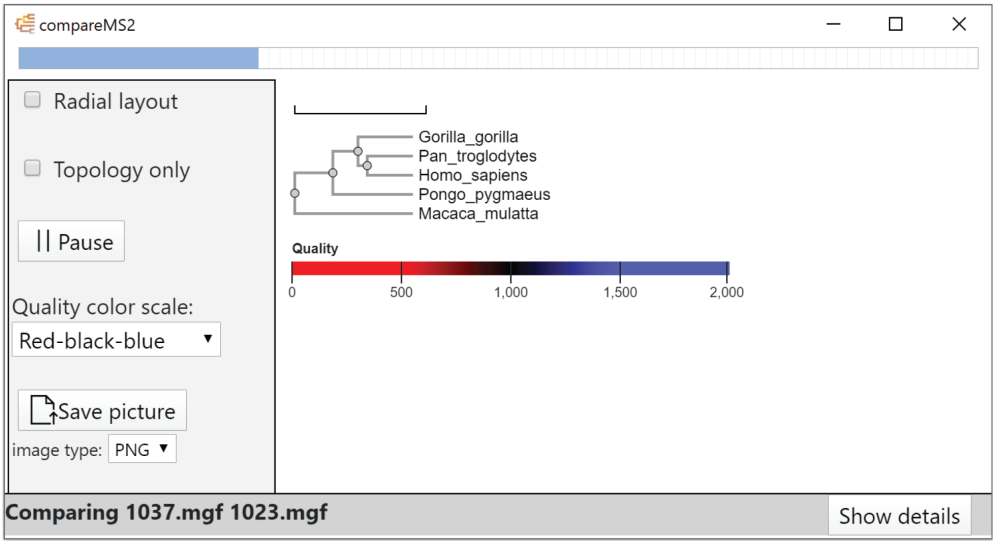


Figure 2. compareMS2 2.0 GUI, showing the output panel from the beginning of an analysis of 24 datasets, each containing 1000 tandem mass spectra, from six primate species for a total of $(24^2 - 24)/2 = 276$ comparisons. With default parameters, these comparisons take 3 min on a PC with an Intel Xeon W-2135 CPU running at 3.70 GHz. The node text color in the tree represents data quality, the default metric being the number of tandem mass spectra per species.

on the peptide sequence and abundance, but also at which point or points during the chromatographic peak the peptide was selected for MS/MS, which is generally not reproducible.

As comparing all tandem mass spectra is computationally expensive, especially for large datasets. compareMS2 allows approximation of $D(\mathcal{A},\mathcal{B})$ by only comparing a spectrum $a \in \mathcal{A}$ with those spectra $b \in \mathcal{B}$ that fall within user-defined windows of retention time or scan number, and precursor m/z.

compareMS2 Pipeline

compareMS2 takes as minimum input a directory of MGF files to be compared. We choose MGF as the default input format, as it is convenient for storing MS2-only data and the MGF files can easily be filtered, split or combined, which may be useful in some applications of compareMS2, such as when fractionating samples or removing nonpeptide spectra. Most vendor software as well as msconvert[14] can convert raw data or mzML files to MGF. To provide faster feedback to the user, compareMS2 2.0 interleaves distance matrix calculations, updates and displays a phylogenetic tree as each row of the distance matrix is completed (Figure 1). With the default symmetric metric, this matrix is triangular, hence the tree is updated rapidly in the beginning, after the first comparison, and then again after the next two comparisons etc. Version 2.0 also provides additional functionality, such as recording a quality control metric for each dataset (by default the number of tandem mass spectra in the dataset) and a filter to compare only the top-N most intense tandem mass spectra from each dataset. The datasets can be compared in alphabetical, size or random order. By default, compareMS2 outputs a MEGA (.meg) file, but Newick and NEXUS formats are also supported.

compareMS2 GUI

Technically, compareMS2 2.0 combines two software tools, which can also be run individually on the command line. The first component compares two datasets, e.g., from LC-MS/MS. The second component takes several such comparisons, combines samples from the same biological species, and computes a distance matrix. The graphical user interface (Figure 2) was designed to be simple to use, hiding most of the internal complexity of compareMS2, including the interleaved execution order of the two components (Figure 1).

Source Code and Availability

The compareMS2 source code can freely be downloaded from https://github.com/524D/compareMS2. On Windows, the software can be installed using a simple installer. compareMS has been tested on Windows 10, Ubuntu 20.04 Linux and MacOS 12. The GUI is based on Electron (https://www.electronjs.org/) and is written in Javascript, HTML, and CSS. It uses the phylotree.js library[15] to render the graphical tree representation. Conversion of the distance matrix into Newick format uses the UPGMA method is and is also implemented in JavaScript. The distance computation and distance matrix creation are performed by two command-line programs written in C. These can be used to run compareMS without the GUI. Source code and prebuild executables of the command-line tools can be found in the external_binaries directory of the compareMS2 repository.

Experimental Features

As compareMS2 provides a basic framework for comparing tandem mass spectra across datasets, we have begun to add experimental features to help visualize such comparisons. The first of these experimental outputs is a two-dimensional histogram of precursor m/z difference and spectral similarity for all comparisons of spectra between two datasets. These features will only be available on the command-line, and require additional software such as R to generate figures, but allow for example correlating spectral similarity with precursor mass

https://doi.org/10.1021/acs.jproteome.2c00457
J. Proteome Res. XXXX, XXX, XXX−XXX

difference. Scripting examples in R are available on https://osf.io/jey28/.

Testing

To demonstrate the features and performance of compareMS2 2.0, we used previously published amaZon ion trap (Bruker Daltonics) and Orbitrap Fusion Lumos (Thermo Fisher Scientific) data from primate sera and an *E. coli* lysate.[1,12] In addition, we used new data acquired on the same Orbitrap instrument and as described in[12] from California sea lion (*Zalophus californianus*), dog (*Canis lupus familiaris*), rock hyrax (*Procavia capensis*), and white-tailed deer (*Odocoileus virginianus*) sera. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[16] partner repository with the dataset identifier PXD034932 and 10.6019/PXD034932. Phylogenetic trees were generated by compareMS2 and MEGA11[17] using default parameters for both (for compareMS2 maximum precursor mass difference 2.05, score cutoff 0.8, minimum basepeak intensity 10000, minimum total ion current 0, maximum retention time difference 60, start retention time 0, end retention time 100000, maximum scan number difference 10000, start scan 1, end scan 1000000, scaling 0.5, noise 10, version of set distance metric 2, version of QC metric 0, compare only the N most intense spectra set to "All", output format "MEGA",and compare order "Smallest-largest first", and for MEGA11 "Lower Left Matrix" and "Pairwise Distance" input data for UPGMA Phylogeny Analysis).

RESULTS AND DISCUSSION

The compareMS2 2.0 GUI (Figure 2) displays a phylogenetic tree with a quality metric mapped to a continuous or divergent color gradient, the tree being continuously updated to provide real-time feedback to the user. This allows executions to be paused or terminated at any stage, which may be useful for large jobs. For example, comparing 100 LC-MS/MS datasets require 4950 pairwise comparisons, taking several hours. But already after six pairwise comparisons of four datasets, trees can be quite informative and reveal if there is an issue with the input files or parameters.

Using the five new serum datasets, each containing between 42,629 and 47,626 tandem mass spectra, we could reconstruct the correct phylogenetic tree in compareMS2 and MEGA11 (Figure 3). The 10 pairwise comparisons in compareMS2 took 40 min with default parameters on a PC with an Intel Xeon W-2135 CPU running at 3.70 GHz. The analyses can be accelerated by comparing spectra within a more narrow m/z window than the default value of 2.05. Each comparison is independent, so in principle the problem is embarrassingly parallel.

To test one of the experimental features, we compared the similarity between tandem mass spectra as a function of precursor m/z difference for comparisons between two closely related species - human and chimpanzee - as well as two species with few shared tryptic peptides—human and *E. coli* (Figure 4). These comparisons reveal information on spectral similarity, but also on mass measurement precision, charge states and isotope errors *before* and independent of any database search, where such parameters typically have to be provided. In these datasets, charge states up to $[M + 6H]^{6+}$ and isotope errors up to at least 3 Da are observed. The analysis can also be used to estimate suitable parameters for compareMS2, e.g., m/z windows and spectral similarity thresholds. We also observe some unexpected side bands most noticeable at 1/2 and 1/3 Da, but not near zero,

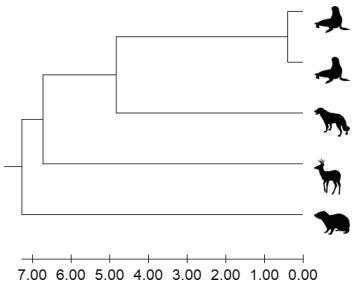


Figure 3. Phylogenetic analysis in MEGA11 based on Orbitrap Fusion Lumos LC-MS/MS datasets of sera from (top to bottom) two California sea lions (*Zalophus californianus*), dog (*Canis lupus familiaris*), white-tailed deer (*Odocoileus virginianus*), and rock hyrax (*Procavia capensis*). The evolutionary history was inferred using the UPGMA method.[18] The optimal tree is shown and drawn to scale, with branch lengths in the same units as those generated by compareMS2 and used to infer the phylogenetic tree. Taxon images are from PhyloPic.

in the Orbitrap data. These bands are also seen in comparisons of spectra within individual datasets.

When combined with posterior error probability estimators such as PeptideProphet[19] or Percolator,[20] spectral similarity measures can in principle be converted into probabilities for any pair of spectra being derived from the same or closely related analytes. When searching spectral libraries, the probability that a query spectrum matches the library spectrum is multiplied with the original probability that library spectrum was correctly identified to estimate the probability the query spectrum is correctly matched to a peptide or other analyte. The compareMS2 software uses the spectral similarity in eq 1 to calculate the overlap between sets of tandem mass spectra without regard to their identification to a specific analyte.

Naïvely, one may attempt to use something like the Jaccard similarity, J, defined as the cardinality of the intersection divided by the cardinality of the union

$$J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \tag{7}$$

However, no two spectra are exactly the same. If the criterion for considering two spectra identical (as in derived from the same peptide) for the purpose of calculating $|\mathcal{A}\cap\mathcal{B}|$ and $|\mathcal{A}\cup\mathcal{B}|$ is too strict, then one will underestimate $|\mathcal{A}\cap\mathcal{B}|$ and overestimate $|\mathcal{A}\cup\mathcal{B}|$. If the criterion is too lax, then one overestimates $|\mathcal{A}\cap\mathcal{B}|$ and underestimates $|\mathcal{A}\cup\mathcal{B}|$. In either case, the errors would multiply, making the Jaccard similarity very sensitive to the precise definition of when two spectra are considered identical. Even more problematic is the intransitive nature of this identity, which is exacerbated by chimeric spectra—spectra that are superpositions of two or more peptide tandem mass spectra. Briefly, a pure spectrum from peptide P can be considered identical to a chimeric spectrum with a small contribution from a second, cofragmenting peptide Q, which in turn is identical to a chimeric spectrum with slightly larger contribution from peptide Q, and so on, eventually ending up with the pure spectrum of peptide Q, which can be very different from the original spectrum from peptide P, just like messages in a game of telephone. This is why exercises clustering large numbers of tandem mass spectra based on spectral similarity tend to produce large globs of spectra rather than a distinct cluster for each peptide.

https://doi.org/10.1021/acs.jproteome.2c00457
J. Proteome Res. XXXX, XXX, XXX−XXX

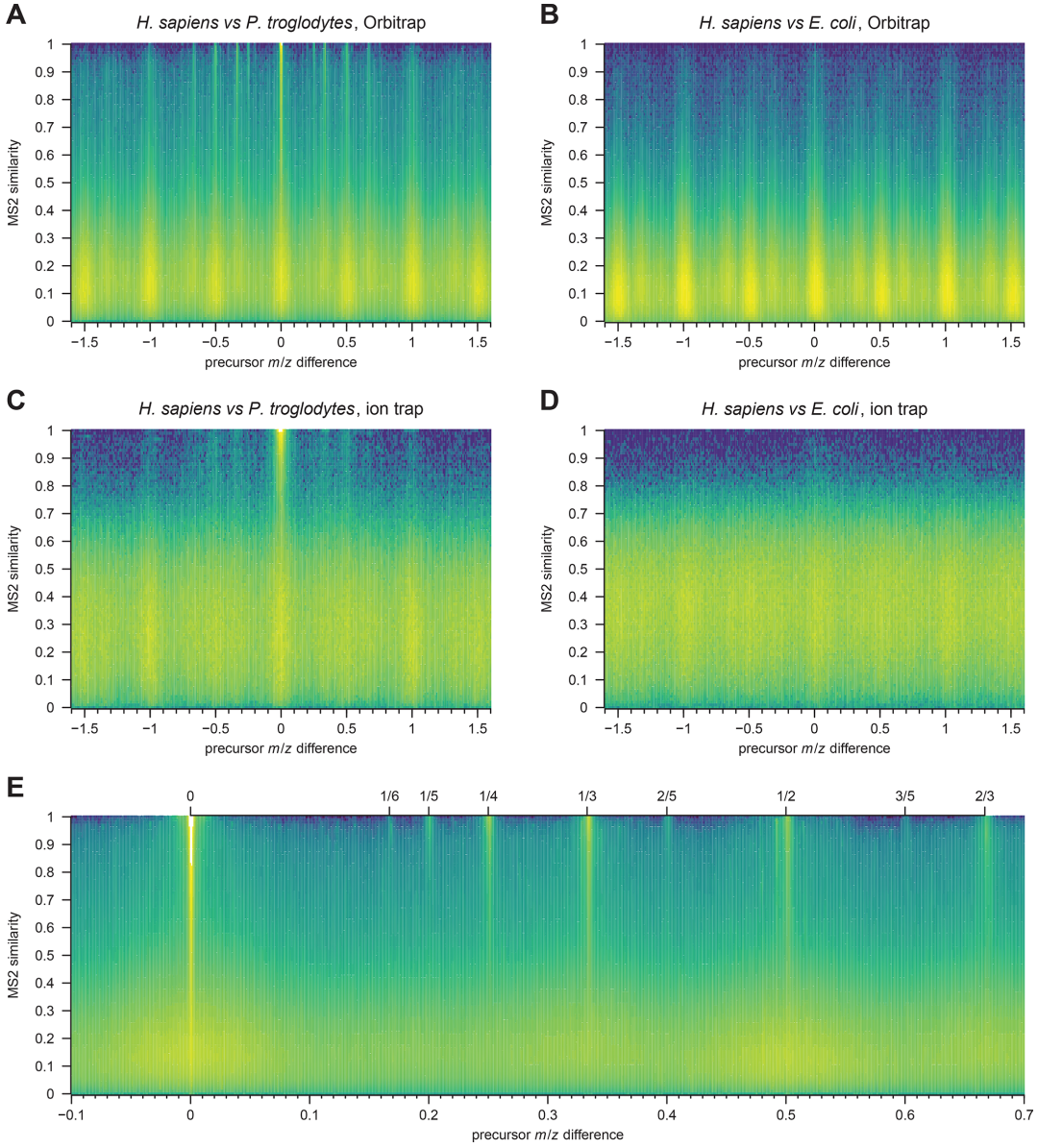


Figure 4. Similarity of tandem mass spectra as a function of precursor m/z difference in Orbitrap Fusion Lumos (A,B) and amaZon ion trap data (C,D), comparing similar (human and chimpanzee sera) and dissimilar (human serum and *E. coli*) samples. Panels A and B compare two LC-MS/MS runs, and panels C and D compare four runs per species (16 comparisons). Similar spectra have precursor m/z differences near zero or a near a rational number corresponding to the isotope error at a specific charge state (shown more clearly in panel E, generated from 8 Orbitrap human serum datasets).

CONCLUSIONS

compareMS2 compares sets of tandem mass spectra to each other rather than to predicted spectra of specific peptides as when identifying proteins from tandem mass spectra. We have used examples from molecular phylogenetics, but many other uses have been demonstrated, including food and feed identification, mixture analysis and experimental design. compareMS2 may also be used data quality control - comparing large numbers of datasets prior to database search and protein quantitation to detect outliers and possible batch effects. The visualization of spectral similarity as a function of precursor mass difference gives another window into the data, and can suggest

https://doi.org/10.1021/acs.jproteome.2c00457
J. Proteome Res. XXXX, XXX, XXX−XXX

parameters for database searches a priori. We make compareMS2 freely available as open source and provide an automatic installer for Microsoft Windows in hope that it may be as useful to others as it has been for us.

AUTHOR INFORMATION

Corresponding Author

Magnus Palmblad − *Center for Proteomics and Metabolomics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands;* orcid.org/0000-0002-5865-8994; Phone: +31 71 5266969; Email: n.m.palmblad@lumc.nl

Authors

Rob Marissen − *Center for Proteomics and Metabolomics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands;* orcid.org/0000-0002-1220-9173

Madhushri S. Varunjikar − *Institute of Marine Research, 5817 Bergen, Norway;* orcid.org/0000-0001-9011-5642

Jeroen F. J. Laros − *National Institute for Public Health and the Environment, 3720 BA Bilthoven, The Netherlands; Department of Human Genetics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands*

Josef D. Rasinger − *Institute of Marine Research, 5817 Bergen, Norway*

Benjamin A. Neely − *National Institute of Standards and Technology, Charleston, South Carolina 29412, United States;* orcid.org/0000-0001-6120-7695

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00457

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Tandem mass spectrometry data used for clustering of California sea lion, chimpanzee, dog, human, rock hyrax, and white-tailed deer serum were graciously provided with permission from an ongoing collaboration with Dr. Michael G. Janech (College of Charleston) as part of the CoMPARe Program (Comparative Mammalian Proteome Aggregator Resource). Specifically, the California sea lion sera were provided by The Marine Mammal Center (Sausalito, CA), the chimpanzee, rock hyrax, and white-tailed deer sera were provided by The Chattanooga Zoo, and the dog serum from Gus (Ohlandt Veterinary Clinic, Charleston, SC). In addition to institutional and NMFS permits and approval, data collection was performed under NIST ACUC MML-AR20-0001. The identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose. The authors also thank all users of previous versions compareMS2 for providing valuable feedback on the software and its installation on different systems.

REFERENCES

(1) Palmblad, M.; Deelder, A. M. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 728−732.

(2) Wulff, T.; Nielsen, M. E.; Deelder, A. M.; Jessen, F.; Palmblad, M. Authentication of fish products by large-scale comparison of tandem mass spectra. *J. Proteome Res.* **2013**, *12*, 5253−5259.

(3) Ohana, D.; Dalebout, H.; Marissen, R.; Wulff, T.; Bergquist, J.; Deelder, A.; Palmblad, M. Identification of meat products by shotgun spectral matching. *Food chemistry* **2016**, *203*, 28−34.

(4) Rasinger, J.; Marbaix, H.; Dieu, M.; Fumière, O.; Mauro, S.; Palmblad, M.; Raes, M.; Berntssen, M. Species and tissues specific differentiation of processed animal proteins in aquafeeds using proteomics tools. *Journal of proteomics* **2016**, *147*, 125−131.

(5) Belghit, I.; Lock, E.-J.; Fumière, O.; Lecrenier, M.-C.; Renard, P.; Dieu, M.; Berntssen, M. H. G.; Palmblad, M.; Rasinger, J. D. Species-Specific Discrimination of Insect Meals for Aquafeeds by Direct Comparison of Tandem Mass Spectra. *Animals: an open access journal from MDPI* **2019**, *9*, 222.

(6) Belghit, I.; et al. Future feed control − Tracing banned bovine material in insect meal. *Food Control* **2021**, *128*, 108183.

(7) Varunjikar, M. S.; Belghit, I.; Gjerde, J.; Palmblad, M.; Oveland, E.; Rasinger, J. D. Shotgun proteomics approaches for authentication, biological analyses, and allergen detection in feed and food-grade insect species. *Food Control* **2022**, *137*, 108888.

(8) van der Plas-Duivesteijn, S. J.; Mohammed, Y.; Dalebout, H.; Meijer, A.; Botermans, A.; Hoogendijk, J. L.; Henneman, A. A.; Deelder, A. M.; Spaink, H. P.; Palmblad, M. Identifying proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *J. Proteome Res.* **2014**, *13*, 1537−1544.

(9) van der Plas-Duivesteijn, S. J.; Wulff, T.; Klychnikov, O.; Ohana, D.; Dalebout, H.; van Veelen, P. A.; de Keijzer, J.; Nessen, M. A.; van der Burgt, Y. E. M.; Deelder, A. M.; Palmblad, M. Differentiating samples and experimental protocols by direct comparison of tandem mass spectra. *Rapid communications in mass spectrometry: RCM* **2016**, *30*, 731−738.

(10) Rieder, V.; Blank-Landeshammer, B.; Stuhr, M.; Schell, T.; Biß, K.; Kollipara, L.; Meyer, A.; Pfenninger, M.; Westphal, H.; Sickmann, A.; Rahnenführer, J. DISMS2: A flexible algorithm for direct proteome-wide distance calculation of LC-MS/MS runs. *BMC bioinformatics* **2017**, *18*, 148.

(11) Downard, K. M. Sequence-Free Phylogenetics with Mass Spectrometry. *Mass Spectrom. Rev.* **2020**, *41*, 3−14.

(12) Neely, B. A.; Palmblad, M. Rewinding the Molecular Clock: Looking at Pioneering Molecular Phylogenetics Experiments in the Light of Proteomics. *J. Proteome Res.* **2021**, *20*, 4640−4645.

(13) Zuckerkandl, E.; Jones, R.; Pauling, L. A Comparison of Animal Hemoglobins by Tryptic Peptide Pattern Analysis. *Proc. Natl. Acad. Sci. U.S.A.* **1960**, *46*, 1349−1360.

(14) Chambers, M. C.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology* **2012**, *30*, 918−920.

(15) Shank, S. D.; Weaver, S.; Kosakovsky Pond, S. L. phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC bioinformatics* **2018**, *19*, 276.

(16) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaíno, J. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **2022**, *50*, D543−D552.

(17) Tamura, K.; Stecher, G.; Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* **2021**, *38*, 3022−3027.

(18) Sneath, P.; Sokal, R. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*; Freeman: San Francisco, 1973.

(19) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002**, *74*, 5383−5392.

(20) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−925.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230867587 (print)
9788230867808 (PDF)