



Universiteit
Leiden
The Netherlands

Advancing zebrafish models in proteomics

Duivesteijn, S.J.

Citation

Duivesteijn, S. J. (2018, October 30). *Advancing zebrafish models in proteomics*. Retrieved from <https://hdl.handle.net/1887/66790>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/66790>

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

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/66790> holds various files of this Leiden University dissertation.

Author: Duivesteijn, S.J.

Title: Advancing zebrafish models in proteomics

Issue Date: 2018-10-30

Advancing Zebrafish Models in Proteomics



Suzanne J. van der Plas-Duivesteijn

Advancing Zebrafish Models in Proteomics

Suzanne J. van der Plas - Duivesteijn

ISBN: 978-94-6375-057-8

Copyright © 2018 S.J. van der Plas – Duivesteijn. All rights reserved. No part of this thesis may be reproduced in any form or by any means without permission from the author.

The work described in this thesis was performed at the Center for Proteomics and Metabolomics of the Leiden University Medical Center in the Netherlands, and was funded by the Netherlands Organization for Scientific Research (NWO), Vidi grant 917.11.398

Printing: Ridderprint BV, Ridderkerk

Cover design: Design Your Thesis, www.designyourthesis.com

Advancing Zebrafish Models in Proteomics

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 30 oktober 2018
klokke 11:15 uur

door

Suzanne Johanna Duivesteijn

geboren te 's-Gravenhage

in 1985

Promotores:

Prof. dr. A.M. Deelder

Prof. dr. A.H. Meijer

Co-promotor:

Dr. N.M. Palmblad

Leden promotiecommissie:

Prof. dr .ir. J.B. Prins

Prof. dr. T.H.M. Ottenhoff

Prof. dr. H.P. Spaink

Prof. dr. ir. F.J. Verbeek

Dr. A. van der Sar (VU University Medical Center Amsterdam)

Waar een wil is

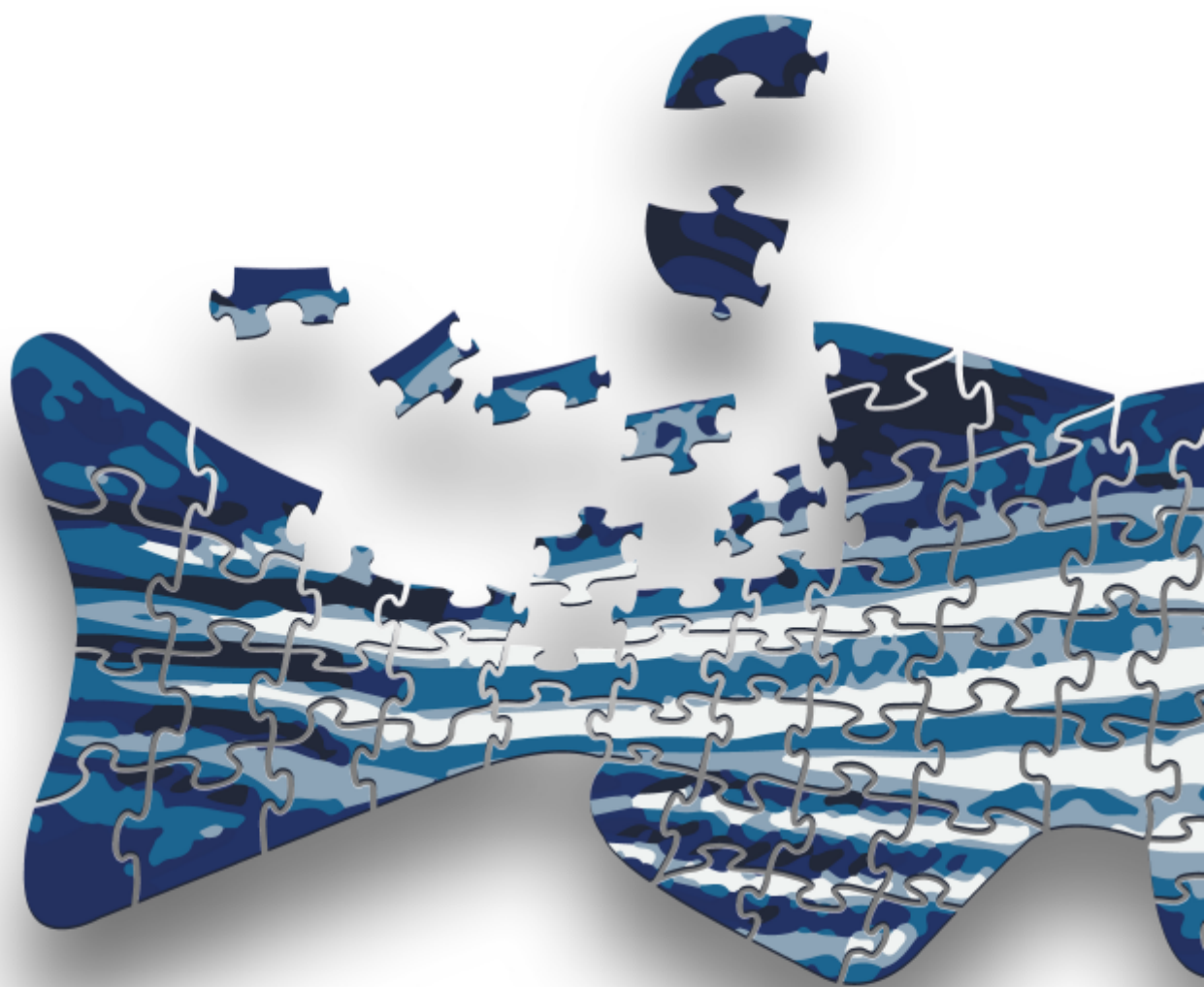
Wat je allemaal kunt bereiken met een grote dosis doorzettingsvermogen

Suzanne

Voor Peter

TABLE OF CONTENTS

Chapter 1	
General Introduction	9
Chapter 2	33
Identifying Proteins in Zebrafish Embryos Using Spectral Libraries Generated from Dissected Adult Organs and Tissues	
Chapter 3	51
Differentiating Samples and Experimental Protocols by Direct Comparison of Tandem Mass Spectra	
Chapter 4	67
A full-body transcriptome and proteome resource for the European common carp	
Chapter 5	89
Zebrafish Models in Proteomics	
Chapter 6	109
Discussion and Summary	
Appendices	
Nederlandse samenvatting	124
List of abbreviations	129
List of publications	134
Curriculum Vitae	137
Dankwoord	138



Chapter 1

General Introduction

Understanding the nature of a given disease within biomedical research (including discovering the nature of disease, drug discovery, and drug development), is impossible by simply focussing on one aspect. Instead, addressing such challenges often requires a multidisciplinary approach where the knowledge of DNA (genomics; what can happen), mRNA (transcriptomics; what will happen), proteins (proteomics; what is happening) and any relevant metabolites (metabolomics; what has happened) is combined. Although each is a field of science on its own, these “omics” obviously complement each other (1-5). Within the field of drug discovery and development, the “omics” are often complemented by using animal models. This thesis brings together proteomics (and to a lesser degree genomics) with an animal model (zebrafish), to develop new tools for the study of infectious diseases and to use of this proteomics information in drug discovery and development.

PROTEOMICS

The study of the proteome (all proteins of an organism or part of an organism) is commonly referred to as *proteomics*. Proteomics provides information about proteins present in (cells of) an organism or part of an organism at a certain time point, and so provides knowledge about the phenotype. Since this thesis is based on proteomics experiments, this topic is introduced first.

This thesis describes the study of proteins, but what is a protein? In a cell nucleus, DNA is transcribed into mRNA, which will be transported outside the nucleus. mRNA migrates to ribosomes (inside the cell's cytoplasm), which translate the mRNA into different amino acids to form a polypeptide chain (sequence). The formation of covalent bonds between amino acids produces a final folded protein. The resulting 3-dimensional structure is unique for each different protein. There are 20 different amino acids known in nature. Each amino acid consists of a carboxyl group (-COOH), an amine group (-NH₂), and a side chain (named R-group). The R-groups make the difference in a protein, as these have different characteristics that affect their protein function, like positive or negative charge (in the amino acids: Arginine, Lysine, Aspartic acid and Glutamic acid), polar or water soluble properties (Glutamine, Asparagine, Histidine, Serine, Threonine, Tyrosine, Cysteine, Methionine, and Tryptophan), or hydrophobic properties (Alanine, Isoleucine, Leucine, Phenylalanine, Valine, Proline, Glycine). Proteins are responsible for a variety of functions within a cell, e.g. organization, maintenance, waste clean-up, intra- and inter-cellular responses, and are therefore located in each part of the cell. In summary, a protein is a folded polypeptide chain consisting of different amino acids, which is formed by ribosomes in the cell's cytoplasm. Proteins can be described as the workhorses of a cell.

Protein production is dependent on different factors. For example the regulatory mechanisms controlling mRNA translation can block the production of proteins. Therefore, genomics can give us information about what can happen, but proteomics provides us information on what is happening. In addition, proteins can be affected after translation by chemical modifications, called post-translational modifications (PTMs) (6). There are different types of PTMs e.g. phosphorylation, ubiquitination, and

additional modifications. Phosphorylation occurs when a phosphoryl group is added to amino acids (most commonly serine and threonine), generating an active or inactive protein. Phosphorylation is mediated by enzymes called kinases, and proteins can be dephosphorylated by phosphatases. During ubiquitination, a small protein called ubiquitin is transferred to a protein substrate by ubiquitin ligase. In this way proteins can be targeted for degradation, but ubiquitination can also serve to modulate protein function. Therefore, investigating ubiquitination can help in understanding protein pathway regulation. Additional modifications are a number of well-known modifications (e.g. methylation, acetylation, glycosylation, oxidation and nitrosylation) all reflecting on different proteins structures and consequently functions. Since PTMs cause variation in the functionality of proteins, examination of PTMs is an important aspect of proteomics.

With such a variety in proteins and protein forms, the analysis of the proteome is complex. Therefore, there are many aspects on which proteins can be examined, e.g. in structural proteomics, expression proteomics, and interaction proteomics. As the term indicates its explanation, structural proteomics refers to the discovery of the structure of proteins, which helps in identifying their functions. In expression proteomics, the main common proteins of cells, organs or organisms are identified, and those proteins are used to examine differential expression (e.g. during disease or in response to treatment), which can be of great use in the field of drug discovery and development as well as in finding new diagnostic markers. In interaction proteomics, the formation of protein complexes is characterized to understand protein function and complex pathways. This thesis focusses on expression proteomics, investigating the presence and quantification of the most abundant proteins in a sample.

In all proteomics experiments described in this thesis a so-called shotgun bottom-up approach was employed. In a standard shotgun bottom-up approach (figure 1 A), proteins are extracted from the samples of interest (e.g. human tissue samples, infected zebrafish embryos) and digested with trypsin, an enzyme that cleaves the protein backbone at the C-terminal end of arginine and lysine (7). Since this adds further complexity to the samples, the tryptic digestion of these samples is followed by separation of the resulting peptides with reversed phase liquid chromatography (LC). This is a separation technique based on hydrophobicity, using a column. The column is for example packed with silica coated with aliphatic carbon chains providing a hydrophobic environment. This is called the stationary phase. Through hydrophobic interaction the peptides will "stick" to this stationary phase. A gradient in the mobile phase (running buffers) will change from e.g. 100% buffer A into 100% buffer B over a certain period of time. Due to the changing hydrophilic (polarity) environment of the mobile phase, proteins will elute from the column. More hydrophobic (apolar) proteins will "stick" to the column and thus its elute time increases, and therefore have a higher so called retention time.

Peptides eluted from the LC column at different times (retention times), are first ionized and transferred into gas phase e.g. by using electrospray ionization (ESI), and finally analysed in the mass analyser (figure 1 B). Mass spectrometry is in essence the measurement of the mass-to-charge (m/z) ratio of charged molecules. In tandem

mass spectrometry, a technique used in this thesis, analysis scans occur in two steps. A first scan (MS^1) registers the m/z ratio and intensities of a batch of peptides (the precursor ions). Then, the most intense precursor ions (observed in the MS^1 spectrum) are selectively isolated and fragmented into smaller pieces in the collision cell (8). The fragmentation – technique we used is called collision induced dissociation (CID), which results in fragments that are displayed in MS/MS (or MS^2) spectra. Using computational tools, the resulting fragmentation pattern, encoded by the MS^2 spectrum, allows for identification of the amino acid sequence of the detected peptide (9, 10).

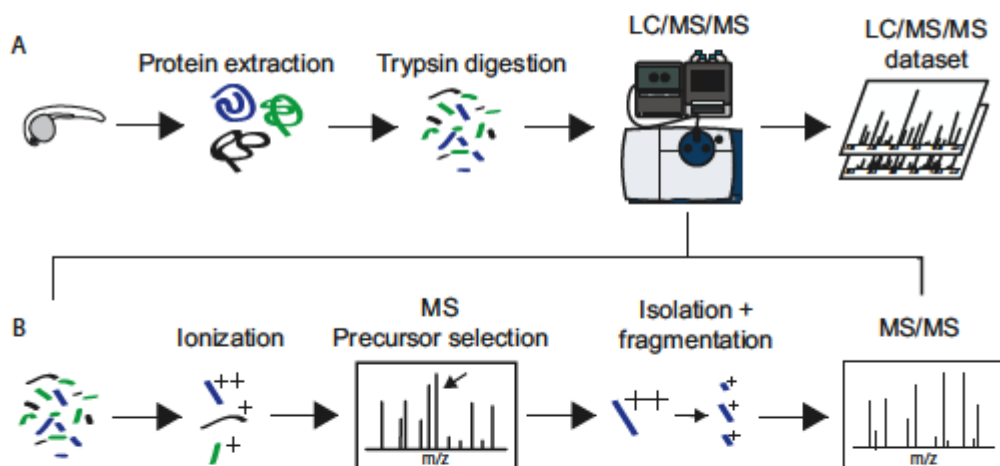


Figure 1: Standard proteomics workflow. A: A standard shotgun bottom-up proteomics experiment. Proteins of a sample (e.g. zebrafish (*Danio rerio*) embryos) are extracted and later digested enzymatically. Tryptic peptides (digested proteins) are separated by LC and analysed on a mass spectrometer of which the outcome is a dataset. **B:** A workflow of what happens inside the mass spectrometer (an IonTrap for this example). After LC separation, peptides enter the mass spectrometer via the electro spray where they are positively charged by ionisation. The mass spectrometer selects a precursor ion from the MS measurement, isolates this peptide and fragments it into smaller pieces, and creates an MS/MS spectrum. From precursor selection onwards, some steps will be repeated according to instrumental settings.

Peptide identification through analysis of MS^2 spectra is usually performed via three different approaches, namely searching against sequence databases, matching spectra in spectral libraries, and *de novo* sequencing of the peptide directly from the MS and MS^2 spectra. In database searches, acquired experimental MS^2 spectra are correlated with *in silico* generated (theoretical) spectra for each peptide derived from a protein sequence database. A peptide spectral library is a collection of annotated (identified) peptide spectra, which serve as templates for peptide/protein identification of the experimental spectra. In spectral library searching, identification is based on the correlation between the query spectral data and previously acquired spectra, taking into account the different fragment ion peaks, and their relative intensities. These intensity patterns are sufficiently reproducible, and as long as the observed peptide has been observed before and included in the library, this method is generally more sensitive than matching theoretical spectra (database searches). The third approach, peptide *de novo* sequencing,

Identifies peptides and proteins using the mass differences (between spectrum peaks) derived from the fragmented peptide ions to reconstruct the sequence of amino acids (11). *De novo* sequencing can be applied both manually and computationally. The main advantage of this technique is the ability of identifying any peptide, also those of which the sequence is not present in a sequence database (e.g. incomplete sequenced genomes). Its main disadvantage is that it requires very high quality MS² spectra and therefore can only be successfully applied to a minority of the acquired tandem mass spectra (11, 12). In this thesis we primarily identified spectra by searching sequence databases and spectral libraries. All databases used in this study were obtained from Uniprot, which in turn consists of sequences from Swiss-Prot and TrEMBL(13). The search score calculated by the database search program measures the degree of similarity between an experimental MS² spectrum and a number of candidate theoretical spectra from the database. Only the best scoring peptide-spectrum match (PSM) for each experimental spectrum is considered as the potential peptide identification and used for further statistical validation steps (10). As described above, a spectral library is a large collection of experimentally observed MS² spectra from previous experiments. A query MS² spectrum is compared to library spectra (using a certain mass tolerance window to restrict the set of candidate spectra) to determine the best match (14). A spectral library search outperforms conventional sequence database searching in terms of speed, error rates, and sensitivity of peptide identification (9, 15).

All data analysis pipelines used in this thesis relied on one or several components of the Trans Proteomic Pipeline (TPP) (16). The TPP is free and open source software, which can analyse mass spectrometry data and aggregate information into peptide or protein lists, quantify the peptides and proteins, and supports stable-isotope labelling schemes as well as label-free quantitation. The TPP is highly accessible and therefore chosen as a starting point for our analysis platform. It can be used for both protein database searches (using X!Tandem) as well as for spectral library searching (using SpectraST) (9, 17, 18). The TPP consists of several integrated software tools, such as Peptide Prophet and ProteinProphet (16, 19). After database and/or spectral library searches, peptide sequences are assigned to MS² spectra. Each assignment of a peptide to a spectrum is a peptide-spectrum match or PSM. To generate a statistical validation for these assignments the PSMs are introduced to PeptideProphet, which estimates false discovery rates (FDR) at the PSM level (16). FDR (a statistical confidence measurement) is defined as "the expected proportion of incorrect PSMs among all accepted PSMs (a global property of a filtered list), or in other words a summary statistics for the entire collection of PSMs (10). After FDR estimation, ProteinProphet generates probabilities for peptide to protein assignments and protein identifications (19). Finally, the output of the Trans Proteomic Pipeline is a list of proteins with their corresponding peptides and additional information (i.e. protein web link, percent sequence coverage, number of unique peptides, total number of independent spectra). Such lists can be used, for example, to identify presence (high abundance) or absence (low or no abundance) of proteins in diseased versus healthy organisms or in comparisons between treated and non-treated organisms.

ANIMAL MODELS AND PHYLOGENETICS

To study disease mechanisms and to identify pathways within biomedical research, *in vivo* experiments are often required, which mostly rely on the use of animal models. For example, studies into disease processes that involve multiple cell types and/or signalling pathways, or studies into the effect of a new drug on a biological system, cannot readily be mimicked in cell cultures and require the use of animal models. An animal model is a non-human model organism used to study and understand biological processes. An animal experiment is defined as an experiment on an animal model by which the animal can experience discomfort. Besides pain, discomfort also includes stress, loneliness, fear and sickness. Within Europe, a specified group of model organisms fall under specific regulations of the animal experiment law; namely living vertebrates and cephalopoda. Due to discomfort and ethical issues on using this specified group of animal models for research purposes, the experimental design needs to be approved by a committee (the Animal Experiments Committee) in a document prior to the start of the experiment. The document needs to describe every detail of the experiment concerning the animals' health and handling, the number of animals needed to be able to statistically test the hypothesis, and always includes the "Three R's" principle (Replacement, Reduction, and Refinement) of Russel and Burch (20). European regulations ensure that experiments using animal models are validated on forehand, and comply with the three R's principle and the relevant animal welfare regulations.

When applying an animal model in an experiment, the primary assignment is to test the fitness of the model to study the experimental hypothesis. This fitness is often based on the equivalency to human biology or chemistry, starting with a general examination of taxonomy and comparative biology of the human and the animal model. Taxonomic equivalence is based on the evolutionary principle that all living organism share a common ancestor and therefore have some (often significant) degree of similarity. This common ancestry is often represented by an evolutionary phylogeny "tree of life". In general, the more closely related the organism is to human, the more similar its biology and biochemistry, and so the more "suitable" it would be as a model system. Knowledge of such relationships and divergent times of origin can be of great importance when designing experiments concerning e.g. drug discovery and development. Comparisons across species to locate conserved gene sequences are fundamental. For example, identifying the closest relatives of a bacteria causing disease can assist in unravelling bacterial properties, which can contribute to a disease phenotype (4). As chimpanzees (*Pan troglodytes*) are our closest living relatives (21) (figure 2 A), they would in theory be the most clinically relevant model to study mechanisms of human disease. However, ethical considerations, cost, time, logistics, and the risk of zoonosis when using this model only allows limited use (22). Moreover, chimpanzees and other non-human primates are protected species, and therefore not allowed to be used as animal models in Dutch laboratories. Therefore, the most common animal models used in the Netherlands are mice (*Mus musculus*), rats (*Rattus norvegicus*), pigs (*Sus scrofa*), chickens (*Gallus gallus*), zebrafish (*Danio rerio*), fruit flies (*Drosophila melanogaster*), and nematodes (e.g. *Caenorhabditis elegans*) (23). Phylogenetic analysis suggests that e.g.

rodents and humans share a last common ancestor that lived approximately 91 million years ago (mya) (figure 2 A) (4). Interestingly, even with this phylogenetic distance (and their considerable phenotypic difference), humans and mice share 90% of their genome (24). Also when comparing other animal models to human, a similar trend can be found where phylogenetic distance seems large, but when looking at genome comparison a close genetic relation can be found, e.g. *P. troglodytes* share 99% of their genome with human (25, 26), *D. rerio* 70% (27), and *C. elegans* 60–80% (difference in *C. elegans* % depends on the type of comparison made) (28). Shared ancestry and similarities (between humans and animal models) down to the molecular level are crucial to determine the expected type of response, and therefore important to apply when planning strategies for measuring this in an experimental design in biomedical research.

The most currently used methods to study molecular evolution/phylogenetics make use of DNA sequences (4). However, before the development of high-throughput DNA sequencing methods, patterns of proteins or enzymatically cleaved peptides (which were separated by chromatographic or electrophoretic methods) were used for molecular phylogenetics (29, 30). A recent proteomics phylogenetics study showed that by using stable, abundant proteins, it is possible to analyse much older specimens than with DNA-based methods (31). The comparison between DNA and protein is not trivial, since one gene, via post-translational modifications of the protein sequence, can give rise to more than one proteoform (and thus finally to more than one biological function). This in turn may affect the phylogenetic distance as measured by shared peptide or protein features, independent of the DNA sequence of the same protein. A recent study explored phylogenetic relationships of primates, using a purely proteomics approach, with an algorithm that compares tryptic peptide tandem mass spectra between species directly (32). This method, based on a common bottom-up or “shotgun” proteomics method, may also be applicable in the design of proteomics experiments, and can be informative in choosing an animal model system. In contrast to evolutionary biology studies, in proteomics studies one does not care about how many million years ago our last common ancestry lived on earth, but we do care about the frequency of shared tryptic peptides between organisms, as these can cause interferences in some experiments while being beneficial in other studies. For example, shared peptides synthesized as stable isotope labelled standards can possibly be used for both human and animal studies, which will reduce the need to synthesize and validate such peptides for each (model) species separately. Or in xenotransplantation experiments, prediction of possible interferences between the two proteomes can be of great value when designing a proteomics study. Such an experiment might give ambiguous protein identifications due to shared peptides and therefore requires proper filtering to exclude those uncertain identifications and their corresponding quantitation. However, an experimental design involving for example *E.coli* and human, which only share a few peptides, will not require such a filtering step and proteomes from these distant species could be measured simultaneously. Furthermore it may be interesting to know whether spectra can be used across (model) species when identifying peptides using spectral libraries and no library for the species of interest is available. For example, one may want to use the extensive zebrafish spectral library for an experiment involving carp (a closely

related species) before a carp spectral library becomes available.

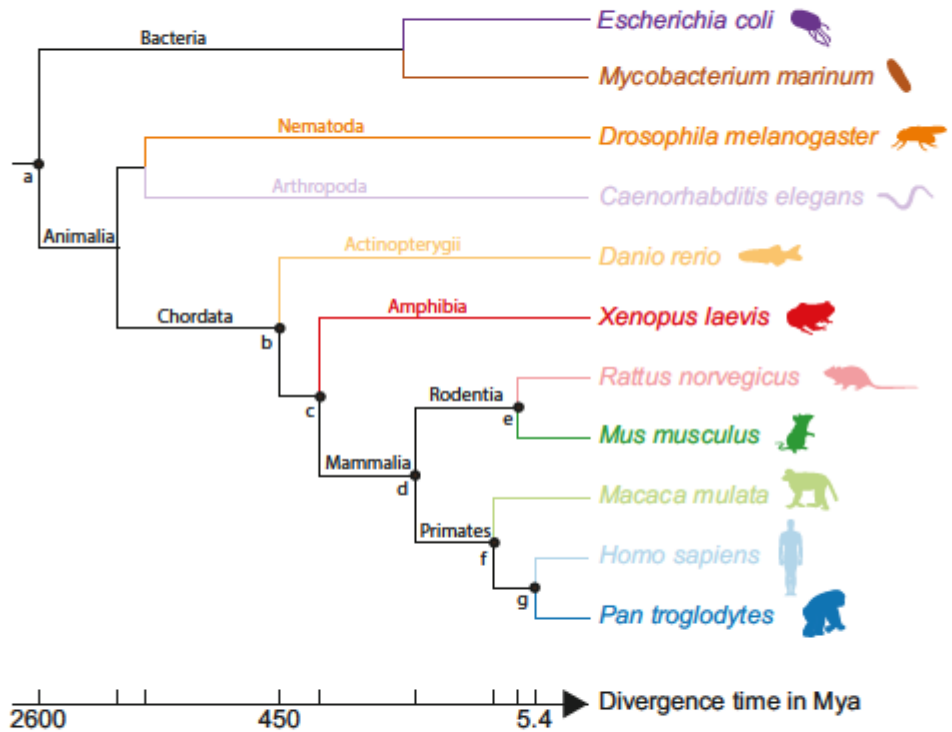
CH
1

The use of *in silico* (computer simulated) experiments can assist in the facilitation of proteomics experiments using mass spectrometry (36). To give an example of an *in silico* based experiment, we theoretically predicted the peptide overlap between human and some commonly used animal models. The proteomes of several animal models, obtained from Uniprot (13), were digested *in silico* based on the cleavage rules of trypsin (the proteinase used in all experiments described in this thesis) to acquire all theoretical possible peptides within each proteome. Then peptides of each proteome were compared to human, to acquire shared peptides between animal model species compared to human. Subsequently, the frequency of all possible shared peptides was plotted against the peptide length with a cut-off of minimally 7 and maximally 78 amino acids, to be able to realistically mimic an *in vivo* experiment (figure 2 B). Of all animal models used to create this figure, the chimpanzee was found to show the highest number of peptides shared with human. In addition this *in silico* proteomics based plot (figure 2 B) shows high similarity compared to a more commonly used ancestral based phylogeny (figure 2 A). Rats and mice for example, which are closely related in phylogenetics, also show high similarity between shared peptides (when compared to human) within proteomics (figure 2 A and B). The shared peptides found with the described *in silico* experiment, can refer to a (highly) conserved region of the genome present in both the animal model and human, or a similar peptide sequence from a random protein that has evolved separately. As mentioned before, within the field of proteomics, it is of importance to explore possible interferences between the frequency of peptides shared between organisms. Here, we attempted to show that *in silico* experiments can be used to predict and explore an experimental design within the field of proteomics when an animal model has to be applied. We here also would like to stress the importance of using phylogeny when applying animal models.

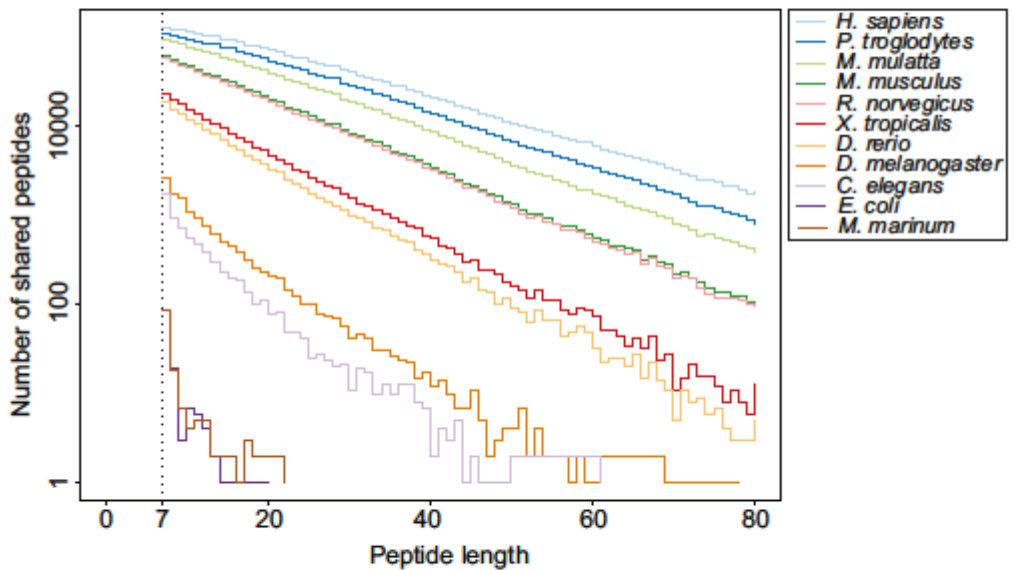
Figure 2: A phylogeny of animal models compared to human. A: The relationships of selected animal models to human, based on genetic studies and divergence times (millions of years ago (Mya)). Branch lengths are not proportional to time. Marked branch nodes refer to a specific divergent times in mya $\pm$ standard error; a: $2,600 \pm 260$, b: 450 ± 36 , c: 179 ± 43 , d: 92 ± 1 , e: 41 ± 1 , f: 23 ± 1.2 , g: 5.4 ± 1.1 . Taxonomic relatedness is generated using iTol and adapted to divergence times reviewed by Hedges (4, 33, 34).

B: *In silico* comparison of selected animal models to human based on tryptic peptides. The frequency of all theoretically possible shared peptides was set against peptide length, and shows the theoretical overlap on peptide level of shared peptides between the different species compared to human. All proteomes were retrieved from UniProt on March 15, 2015 (13), containing roughly 20,000 proteins per proteome. Peptides lengths were generated with minimally 7, and maximally 78 amino acids, and constrained to a maximum of two missed cleavages. The figure was generated using R version 3.1.3 (35).

A



B



THE ZEBRAFISH ANIMAL MODEL

As mentioned before, to study disease mechanisms and pathways within biomedical research, *in vivo* experiments are often required, including the use of animal models. One of these animal models, the zebrafish (*Danio rerio*), a tropical fish species originating from Southeast Asia, was introduced as an animal model in the 1970s (37), and has emerged over time as a powerful animal model for human drug discovery and development (38-42). There are many aspects that make this animal model so successful and popular in research. The zebrafish is easy to handle in a laboratory environment. The model allows applicability in research at different life stages, of which the fast developing early life stages (embryos and larvae) are mainly used. Zebrafish produce approximately 200 eggs per pair per week, which enables the use of high throughput screens (43-49). The natural transparency of the eggs allow the use of fluorescent techniques, assisting in visualizing effects on a cellular level *in vivo* without sacrificing the organism (50, 51). Due to the advantages of transparency, zebrafish strains that remain transparent as adults are now also available (52).

Another aspect that makes the zebrafish a popular animal model is its genetically relatedness to human. Clearly there are major differences between the anatomy of human and zebrafish, reflecting their evolutionary distance. Nevertheless, despite their taxonomic distance, approximately 70% of their protein-coding genes shows strong similarities (27). Several studies indicated the homology of the zebrafish animal model on anatomical, physiologically, and molecular levels, which highlighted the applicability of the zebrafish animal model in biomedical research (41, 53-59).

Its increased use led to further development of laboratory techniques for the zebrafish animal model. Currently, mutant strains and zebrafish cell lines are available (60), many genetic and genomic tools can be applied (61), a system for rapid collection of eggs has been developed (62), and a robotic high-throughput injection and infection screening system is available (43). Many of the described techniques are developed to enable high-throughput screening (HTS) (63). HTS is a powerful tool in biomedical research, especially in drug development, which is applicable when using the embryo/larval stage of this animal model.

TUBERCULOSIS

In this thesis we attempt to investigate the application of proteomics methods on an animal model for infectious disease and its potential use in biomedical research. Tuberculosis (TB) is marked as the number 9 of the most "dangerous" infectious diseases in the world (64). To understand the link to the used animal model in this study (zebrafish) and the infectious disease to be examined, the characteristics of TB disease are described first.

Tuberculosis (TB), caused by the bacillus *Mycobacterium tuberculosis*, is a an epidemic with worldwide 10.4 million cases reported and 1.3 million deaths in 2016 (64). Even

though the TB epidemic is decreasing (with TB mortality falling 3% per year and TB incidence falling 2% per year), the goals to improve this fall to 4-5% per year by 2020 is not reached yet (64). TB is correlated with poor healthcare conditions, and other diseases that negatively affect the immune system (like HIV). In addition, resistance of the bacillus against available drugs keeps expanding (64). There already are five classifications of TB resistance; mono resistance, poly-resistance, multi drug resistant TB (MDR-TB), extensively drug resistant TB (XDR-TB), and rifampicin resistance (RR). Besides the highly infectious active form of TB, latent TB infection (LTBI) is an even bigger health problem affecting one third of the world population (65). In LTBI, infection is controlled but not eliminated by the host, causing an asymptomatic disease. As TB has emerged as a global public health problem, it has become a major challenge to discover new drugs which requires a better understanding of the mechanisms underlying disease development. The wide range of omics research tools recently developed, including proteomics, provides new means to study the molecular details of TB disease.

Understanding the route of infection is highly important in drug discovery and development. *M. tuberculosis* enters the lungs (mainly via human sputum), activating the first line defence (innate immunity) of the immune system, which signals to attract different types of cells including macrophages (MΦ). MΦ are able to kill and degrade bacteria using a variety of mechanisms, but *M. tuberculosis* has evolved mechanisms to withstand the MΦ defences and replicate inside the MΦ (66, 67). Bacteria are phagocytosed by MΦ into phagosomes, a cell membrane-derived vesicle that is formed around the bacteria. In addition, if bacteria escape from a phagosome into the cytoplasm, they can subsequently become entrapped in double-membraned vesicles known as autophagosomes. The phagosome or the autophagosome can fuse with a lysosome (containing hydrolytic enzymes capable of degrading most kind of biomolecules including proteins), into a (auto)phagolysosome. *M. tuberculosis* has developed a diversity of tactics to prevent lysosomal fusion, for example delaying phagosome maturation, blocking acidification, inhibition of phagosome-lysosome fusion, and interference of the host's protein signalling, and so preventing immune recognition and elimination (68-70). Replication of the bacteria, and finally escape from the macrophage through induction of cell death, allows infection of newly attracted MΦ and further spread into the body (71). A major hallmark of TB is the formation of granulomas. Granulomas are well-organized structures and are composed of the following cell types: MΦ (in various forms: apoptotic infected, necrotic infected, foamy, non-infected, epithelioid and apoptotic infected epithelioid), T-cells, B-cells, neutrophils, natural killer cells, dendritic cells and giant cells (72, 73). Granulomas form through the attraction of immune cells by immunological signalling pathway, which is activated by infected macrophages. After replication of the multiple bacteria inside a granuloma, the granuloma will burst open and spread through the body. When this phenomenon occurs in the lungs, it will result in the development of coughing, which promotes the bacteria to the air, causing aerosol spread of the bacteria (72), a circumstance also known as pulmonary disease. Even though TB granulomas are generally known as a symptom of pulmonary disease, the bacterium can reach multiple organs and tissues causing infection in the entire body. Although extensive research has been conducted on mechanisms involved in the

various stages of TB infection, more research on molecular mechanisms and pathways is necessary in order to increase the understanding of the pathways of the disease, and to be able to develop novel treatment strategies (66, 74, 75).

ZEBRAFISH AS AN ANIMAL MODEL TO STUDY TB INFECTION

A TB infection caused by *M. tuberculosis*, is characterized by its complex interaction between the pathogen (tuberculosis) and its host (human), which makes it very difficult to find an ideal model to study mycobacterial pathogenesis (22, 76, 77). TB is an aerosol contagious disease often studied in an animal model using an aerosol experimental design. Even though this method perfectly mimics the natural route of human infection, it carries a great potential safety risk (78). In addition, animal models differ in their anatomy, cellular composition of the tracheobronchial epithelium, local phagocytic and chemical defences, inflammatory and immune response, and therefore these differences need to be taken into account when extrapolating the results of TB infection experiments to humans (22, 76, 79). Many different experimental animal models have been used, but none of them completely reproduces all aspects of TB in humans (22, 75, 80). This does not mean that the diverse animal models available are ineffective, they just have their own advantages and disadvantages, each helpful in unravelling bits of the human TB infection puzzle.

Even though primates seem the most relevant animal model to study TB infection, as discussed earlier, there is a limitation in their use (81). The design of this thesis was focussed on studying immune responses caused by TB infection. There are multiple stages in the human immune response to *M. tuberculosis*, wherefore different animal models have been described that mimic these stages (75, 82). Since we were primarily interested in the first line of defence and early granuloma formation within TB infection, we had to choose an animal model that mimics these two aspects in the best possible way to human. Disadvantages of the mouse model are that *M. tuberculosis* is not the natural pathogen of the mouse and that there are structural differences in granuloma development (77, 81, 83-88). Even though extensively used, animal models like guinea pig and rabbit are limited by the lack of immunological reagents required for the qualitative and quantitative evaluation of immune responses (77, 81). Besides this, all three animal models carry a potential risk for human health due to the use of *M. tuberculosis* and the possibility of aerosol spreading. *Mycobacterium marinum* is a closely related pathogen (89), which is less dangerous to human and can naturally infect cold blooded animals, including frogs and fish (90, 91). One of its hosts is the zebrafish (*Danio rerio*), the model species used in this thesis.

There are clearly major differences between human TB infection, as a pulmonary disease, and TB in zebrafish which do not possess lungs. Nevertheless, several mechanisms involved in infection can be studied also in other tissues, such as the failure of the immune system to eradicate the pathogen, and the formation of granulomas. Therefore, the fitness of this model to study human TB infection, starts with its similarity to humans with regards to the immune system. The immune system of zebrafish and human are

remarkably similar, and both develop innate and acquired immune responses (55, 56, 92), which are involved in protection of the host against TB infection (93–95). The hallmark cells of human immunity including macrophages, granulocytes, T and B lymphocytes, have all been identified in zebrafish (95, 96). Innate immunity provides immediate protection against infection, acting as a first line defence mechanism. Acquired immunity is the second line defence, which is activated by the innate immune system as a result of recognition of pathogen-specific receptors. In zebrafish, innate immunity (functioning from 24 hours post fertilization) and acquired immunity (not active during the first weeks, and mature at 4–6 weeks) can be studied separately, which is useful when studying the role of the innate immunity and especially the role of MΦ cells after TB infection *in vivo* (97–102). In addition, several components of the main immune signalling pathways (e.g. Toll like receptors, and Myeloid differentiation primary response gene 88) are strongly conserved in both fish and mammals (41, 53–56). Toll like receptors (TLRs) are membrane proteins (e.g. on macrophages) that recognize pathogens like mycobacteria, and activate a specific immune response (activation of innate immunity) upon this recognition (103). Myd88 is a universal adapter protein used by TLRs (except TLR3) to activate downstream signalling in the immune response (82, 103). Besides the immune system itself, also the pathological reaction upon TB infection shows similarity between human and zebrafish. Several studies have shown the formation of granulomas in adults and granuloma-like aggregates in embryos as an effect of an infection with *M. marinum*, the pathogen used to model TB in fish (93, 104–106). The granulomas in adult zebrafish undergo caseous necrosis like in human TB infection (91, 93). However, the major strength of the zebrafish model lies in the use of embryos/larvae, since the granuloma-like aggregates formed in this stage of development allow visualization of the earliest stages of granuloma formation, which is not possible in any other animal model (73, 94, 96, 107, 108). Several important discoveries have been reported using the zebrafish TB infection model. One of them is the indication that granuloma formation, rather than having a strictly host protective role, facilitates bacterial proliferation and dissemination during the early stages of infection (109, 110). This idea is supported by the discovery that granuloma formation is dependent on the mycobacterial RD1 virulence locus (encoding the ESX-1 secretion system and secreted virulence factors), which was found to be crucial for the movement of macrophages during granuloma formation in zebrafish (110, 111).

The bacteria *M. marinum* that is used in the zebrafish TB model, is closely related to *M. tuberculosis* (figure 3) (89). Both bacterial strains share 3000 orthologs, including important virulence factors with an average amino acid identity of 85% (112), and both have the ability to replicate inside macrophages and induce the formation of granulomas (66, 91). Experiments in which *M. marinum* infection was studied during early stages of pathogenesis, led to new insights into the function of granulomas, demonstrating the importance of the embryonic zebrafish model, as described above (66, 109, 110, 113–115). Using a model pathogen with its natural host (*M. marinum* in zebrafish) to study human TB (*M. tuberculosis*) infection has several advantages (116). *M. marinum* develops rapidly and can be kept at a lower laboratory biosafety level than *M. tuberculosis*. Furthermore *M. marinum* and zebrafish require a temperature around 28°C to function optimally, while *M. tuberculosis* grows optimally at higher temperatures (89,

90). *M. marinum* can also cause granulomas to form on the human skin, but disseminated infections are rare due to the lower temperature requirements of this pathogen (117). Since zebrafish prefer an ambient temperature of 28°C and are not a natural host for *M. tuberculosis*, studying *M. tuberculosis* in zebrafish would not mimic the bacterial growth like in humans. In summary, to obtain more informative data, and studying co-evolved host-pathogen pairs is probably very efficient, we therefore have chosen to study *M. marinum* infection in zebrafish for later translation to the human situation.

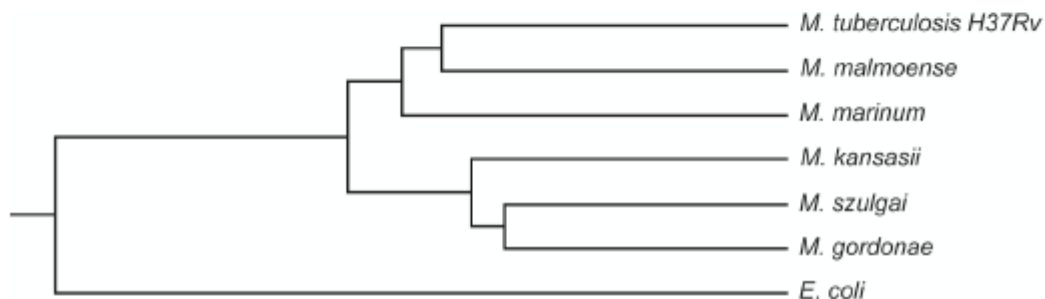


Figure 3: Distance relation clustering of different mycobacteria species. This clustering shows the relation between different mycobacteria species based on LC/MS/MS ion trap data (de Keijzer *et al.*, in prep.). *Mycobacterium tuberculosis* and *Mycobacterium marinum* are clustered closely together, which is comparable to a genetic (16S rRNA sequences) study by Tonjum *et al.* (89).

The application of the zebrafish animal model to study TB infection has led to the development of several techniques. By making use of fluorescently labelled bacteria (*M. marinum*), infection routes can be followed over time (105). Cells of the host (e.g. MΦ or neutrophils) can be fluorescently labelled additionally, to assist the visibility of infection pathways *in vivo* (50, 51). Moreover, several injection techniques are well established for infecting zebrafish with *M. marinum* (105) including the use of a robotic high-throughput injection and infection screening system (43).

Until now, research using the zebrafish TB infection model has mainly focussed on genetics, leading to many important discoveries (118-120). Still, many pathways remain to be discovered to reveal the role of MΦ and granuloma formation in TB infection (96), causing either active or latent TB. Several proteomics studies related to TB, have shown proteomics to add valuable information for unravelling the TB infection and treatment puzzle. A recently published review showed exosomes to be an excellent source of biomarkers and described the latest discoveries of their role in TB with the use of proteomics (121). Another proteomics study revealed the existence of four potential antigen candidates for TB vaccine development using a mouse and guinea pig model (122). An even more recent proteomics publication has shown an important role for proteins secreted by *M. tuberculosis* in facilitating and maintaining infection (123). Already in 2012 the possibility of using proteomics to study phagosome specific proteins and signalling pathways was described, for a better understanding and control of TB (124). The above mentioned studies provide just a glimpse of all literature available on TB studies within the field of proteomics. But notably, few publications were found applying

the zebrafish TB infection model within the field of proteomics. Since proteomics can potentially make an important contribution to increase our understanding of TB pathogenesis, and the zebrafish TB infection model has been shown to be important in unravelling parts of the TB puzzle, we applied the zebrafish TB infection model within a proteomics approach.

SCOPE OF THE THESIS

This thesis project sets out to explore the use of proteomics in research applying the zebrafish as a model, to eventually use it to study infectious disease, exemplified by *Mycobacterium marinum* infection. This thesis brings together three different areas; proteomics, zebrafish as animal model, and tuberculosis. To bridge these topics the thesis consists of proteomics, method development, phylogenetics, and the application of proteomics to the *M. marinum* zebrafish infection model.

Even though the zebrafish is widely used in biomedical research, it is a relatively new model within proteomics. As mentioned in this introduction, mass spectrometry data can be searched in many different ways, with a preference for library searching due to its advantages (higher speed, lower error rates and improved sensitivity) compared to database searches. Many libraries are available for a number of (model) organisms, e.g. human, mouse and *Drosophila*. For zebrafish however, a spectral library was lacking. Therefore we aimed to create a spectral library for this animal model. As building spectral libraries can be time consuming, **chapter 2** describes a workflow for the creation and validation of a zebrafish spectral library. In addition, the created library was made available on NIST, a webpage that hosts additional important spectral libraries.

In planning strategies for an experimental design to test the fitness of an animal model to the experimental hypothesis, shared ancestry and similarities down to molecular level are crucial. As mentioned under “animal models and phylogenetics” of this introduction, exploration of this field in proteomics could be expanded into the field of proteomics. A new method published using a direct comparison between tryptic peptide tandem mass spectra alone, suggests a wide range of practical applications for example for molecular phylogenetics, experimental design, and forensics. **Chapter 3** describes several practical applications of this previously developed software tool.

A disadvantage of the zebrafish model is its small organ size and limited amount of cells. A closely related and extensively studied species, the common carp (*C. carpio*) might be useful to overcome these limitations due to its larger size and relatedness. **Chapter 4** describes the use of a closely related species to fill these gaps.

The zebrafish animal model is marked as a valid model for genomic studies. However, its application within proteomics is found less frequent. **Chapter 5** provides an overview of the possibilities, limitations and practical considerations of the application of the zebrafish animal model within a proteomics design. Also available tools and resources that are suitable for zebrafish proteomics are described. In addition, this chapter

briefly describes the application of the zebrafish model to study TB infection within a proteomics design. Finally, in **chapter 6** the results of this thesis are discussed in relation to the recent developments in the field of proteomics.

REFERENCES

1. Davis, C.D. & Milner, J. Frontiers in nutrigenomics, proteomics, metabolomics and cancer prevention. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 551, 51-64 (2004).
2. Kelkar, D.S. et al. Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis. *Mol Cell Proteomics* (2014).
3. Singh, S.K., Aravamudhan, S., Armant, O., Kruger, M. & Grabher, C. Proteome dynamics in neutrophils of adult zebrafish upon chemically-induced inflammation. *Fish & shellfish immunology* 40, 217-224 (2014).
4. Hedges, S.B. The origin and evolution of model organisms. *Nat Rev Genet* 3, 838-849 (2002).
5. Palmblad, M. et al. Parallel deep transcriptome and proteome analysis of zebrafish larvae. *BMC Res Notes* 6, 428 (2013).
6. Zhao, Y. & Jensen, O.N. Modification-specific proteomics: Strategies for characterization of post-translational modifications using enrichment techniques. *PROTEOMICS* 9, 4632-4641 (2009).
7. Olsen, J.V., Ong, S.E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics* 3, 608-614 (2004).
8. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* 422, 198-207 (2003).
9. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7, 655-667 (2007).
10. Nesvizhskii, A.I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* 73, 2092-2123 (2010).
11. Wang, P. & Wilson, S.R. Mass spectrometry-based protein identification by integrating de novo sequencing with database searching. *BMC Bioinformatics* 14 Suppl 2, S24 (2013).
12. Kapp, E. & Schutz, F. Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.] Chapter 25, Unit 25.22* (2007).
13. The UniProt Consortium UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45, D158-D169 (2017).
14. Stein, S.E. & Scott, D.R. Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom* 5, 859-866 (1994).
15. Lam, H. et al. Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* 5, 873-875 (2008).
16. 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.* 74, 5383-5392 (2002).
17. Deutsch, E.W. et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 10, 1150-1159 (2010).
18. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466-1467 (2004).
19. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 75 (2003).
20. Russell, W.M.S.a.B., R.L. The Principles of Humane Experimental Technique.

(Methuen, London; 1959).

21. Wood, B. & Richmond, B.G. Human evolution: taxonomy and paleobiology. *Journal of Anatomy* 197, 19-60 (2000).
22. Lopez Hernandez, Y., Yero, D., Pinos-Rodriguez, J.M. & Gibert, I. Animals devoid of pulmonary system as infection models in the study of lung bacterial pathogens. *Frontiers in microbiology* 6, 38 (2015).
23. NVWA (Nederlandse Voedsel en Warenautoriteit Proefdieren en Dierproeven, 2012).
24. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-562 (2002).
25. Kehrer-Sawatzki, H. & Cooper, D.N. Structural divergence between the human and chimpanzee genomes. *Human genetics* 120, 759-778 (2007).
26. Kehrer-Sawatzki, H. & Cooper, D.N. Understanding the recent evolution of the human genome: Insights from human-chimpanzee genome comparisons. *Human mutation* 28, 99-130 (2007).
27. Howe, K. et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature advance online publication* (2013).
28. Kaletta, T. & Hengartner, M.O. Finding function in novel targets: *C. elegans* as a model organism. *Nature reviews. Drug discovery* 5, 387-398 (2006).
29. Zuckerkandl, E. The Evolution of Hemoglobin. *Sci Am* 212 (1965).
30. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J Theor Biol* 8 (1965).
31. Welker, F. et al. Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature* 522, 81-84 (2015).
32. Palmblad, M. & Deelder, A.M. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun. Mass Spectrom.* 26, 728-732 (2012).
33. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39, W475-478 (2011).
34. Letunic, I. & Bork, P. Interactive Tree Of Life (ITOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23, 127-128 (2007).
35. R Core Team (R Foundation for Statistical Computing, Vienna, Austria; 2015).
36. Cagney, G., Amiri, S., Premawaradena, T., Lindo, M. & Emili, A. In silico proteome analysis to facilitate proteomics experiments using mass spectrometry. *Proteome science* 1, 5 (2003).
37. Streisinger, G., Walker, C., Dower, N., Knauber, D. & Singer, F. Production of clones of homozygous diploid zebra fish (*Brachydanio rerio*). *Nature* 291, 293-296 (1981).
38. Lieschke, G.J. & Currie, P.D. Animal models of human disease: zebrafish swim into view. *Nat Rev Genet* 8, 353-367 (2007).
39. Brittijn, S.A. et al. Zebrafish development and regeneration: new tools for biomedical research. *International Journal of Developmental Biology* 53, 835-850 (2009).
40. Ali, S., Champagne, D.L., Spaink, H.P. & Richardson, M.K. Zebrafish embryos and larvae: a new generation of disease models and drug screens. *Birth defects research. Part C, Embryo today : reviews* 93, 115-133 (2011).
41. Meijer, A.H. & Spaink, H.P. Host-pathogen interactions made transparent with the zebrafish model. *Curr. Drug Targets.* 12, 1000-1017 (2011).
42. Kari, G., Rodeck, U. & Dicker, A.P. Zebrafish: An Emerging Model System for Human Disease and Drug Discovery. *Clin Pharmacol Ther* 82, 70-80 (2007).

43. Carvalho, R. et al. A high-throughput screen for tuberculosis progression. *PLoS one* 6, e16779 (2011).
44. Veneman, W.J. et al. A zebrafish high throughput screening system used for *Staphylococcus epidermidis* infection marker discovery. *BMC Genomics* 14, 255 (2013).
45. Pardo-Martin, C. et al. High-throughput in vivo vertebrate screening. *Nat Meth* 7, 634-636 (2010).
46. Lessman, C.A. The developing zebrafish (*Danio rerio*): a vertebrate model for high-throughput screening of chemical libraries. *Birth defects research. Part C, Embryo today : reviews* 93, 268-280 (2011).
47. Letamendia, A. et al. Development and validation of an automated high-throughput system for zebrafish in vivo screenings. *PLoS one* 7, e36690 (2012).
48. Spalink, H.P. et al. Robotic injection of zebrafish embryos for high-throughput screening in disease models. *Methods* 62, 246-254 (2013).
49. Miscevic, F., Rotstein, O. & Wen, X.Y. Advances in zebrafish high content and high throughput technologies. *Comb Chem High Throughput Screen* 15, 515-521 (2012).
50. Henry, K.M., Loynes, C.A., Whyte, M.K. & Renshaw, S.A. Zebrafish as a model for the study of neutrophil biology. *Journal of leukocyte biology* 94, 633-642 (2013).
51. Cui, C. et al. Infectious disease modeling and innate immune function in zebrafish embryos. *Methods Cell Biol.* 105, 273-308 (2011).
52. White, R.M. et al. Transparent adult zebrafish as a tool for in vivo transplantation analysis. *Cell stem cell* 2, 183-189 (2008).
53. Meijer, A.H. et al. Expression analysis of the Toll-like receptor and TIR domain adaptor families of zebrafish. *Molecular immunology* 40, 773-783 (2004).
54. Stein, C., Caccamo, M., Laird, G. & Leptin, M. Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biol* 8, R251 (2007).
55. Traver, D. et al. The zebrafish as a model organism to study development of the immune system. *Adv. Immunol.* 81, 253-330 (2003).
56. Trede, N.S., Langenau, D.M., Traver, D., Look, A.T. & Zon, L.I. The use of zebrafish to understand immunity. *Immunity*. 20, 367-379 (2004).
57. Reimers, M.J., Hahn, M.E. & Tanguay, R.L. Two zebrafish alcohol dehydrogenases share common ancestry with mammalian class I, II, IV, and V alcohol dehydrogenase genes but have distinct functional characteristics. *J Biol Chem* 279, 38303-38312 (2004).
58. Renier, C. et al. Genomic and functional conservation of sedative-hypnotic targets in the zebrafish. *Pharmacogenetics and genomics* 17, 237-253 (2007).
59. Li, C., Tan, X.F., Lim, T.K., Lin, Q. & Gong, Z. Comprehensive and quantitative proteomic analyses of zebrafish plasma reveals conserved protein profiles between genders and between zebrafish and human. *Scientific Reports* 6, 24329 (2016).
60. Zebrafish International Resource Center (ZIRC), University of Oregon, USA, (<http://zebrafish.org/home/guide.php>)
61. Rasooly, R.S. et al. Genetic and genomic tools for zebrafish research: the NIH zebrafish initiative. *Developmental dynamics : an official publication of the American Association of Anatomists* 228, 490-496 (2003).
62. Adatto, I., Lawrence, C., Thompson, M. & Zon, L.I. A new system for the rapid collection of large numbers of developmentally staged zebrafish embryos. *PLoS*

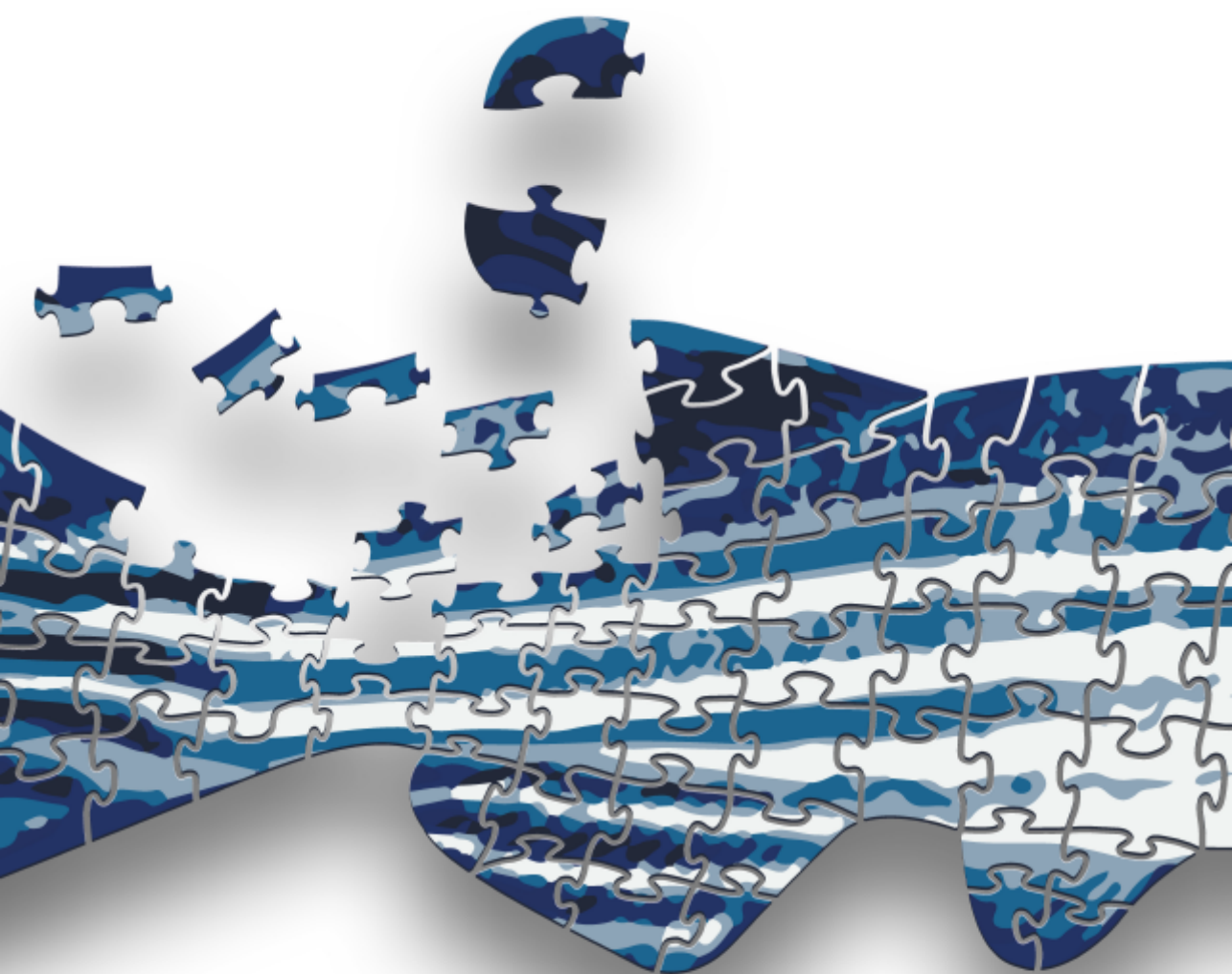
one 6, e21715 (2011).

63. Takaki, K., Cosma, C.L., Troll, M.A. & Ramakrishnan, L. An in vivo platform for rapid high-throughput antitubercular drug discovery. *Cell reports* 2, 175-184 (2012).
64. WHO (2017).
65. Houben, R.M. & Dodd, P.J. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS medicine* 13, e1002152 (2016).
66. van der Vaart, M., Spalink, H.P. & Meijer, A.H. Pathogen recognition and activation of the innate immune response in zebrafish. *Advances in hematology* 2012, 159807 (2012).
67. Hussain Bhat, K. & Mukhopadhyay, S. Macrophage takeover and the host-bacilli interplay during tuberculosis. *Future microbiology* 10, 853-872 (2015).
68. Meena, L.S. & Rajni Survival mechanisms of pathogenic *Mycobacterium tuberculosis* H37Rv. *FEBS Journal* 277, 2416-2427 (2010).
69. Vandal, O.H., Pierini, L.M., Schnappinger, D., Nathan, C.F. & Ehrt, S. A membrane protein preserves intrabacterial pH in intraphagosomal *Mycobacterium tuberculosis*. *Nature medicine* 14, 849-854 (2008).
70. Russell, D.G. *Mycobacterium tuberculosis*: here today, and here tomorrow. *Nature Rev. Mol. Cell Biol.* 2, 569-577 (2001).
71. Duclos, S. & Desjardins, M. Subversion of a young phagosome: the survival strategies of intracellular pathogens. *Cellular microbiology* 2, 365-377 (2000).
72. Russell, D.G. Who puts the tubercle in tuberculosis? *Nat Rev Micro* 5, 39-47 (2007).
73. Ramakrishnan, L. Revisiting the role of the granuloma in tuberculosis. *Nature reviews. Immunology* 12, 352-366 (2012).
74. Flynn, J.L. Lessons from experimental *Mycobacterium tuberculosis* infections. *Microbes and infection / Institut Pasteur* 8, 1179-1188 (2006).
75. Ernst, J.D. The immunological life cycle of tuberculosis. *Nature reviews. Immunology* 12, 581-591 (2012).
76. Vilaplana, C. & Cardona, P.J. The lack of a big picture in tuberculosis: the clinical point of view, the problems of experimental modeling and immunomodulation. The factors we should consider when designing novel treatment strategies. *Frontiers in microbiology* 5, 55 (2014).
77. Dharmadhikari, A.S. & Nardell, E.A. What animal models teach humans about tuberculosis. *American journal of respiratory cell and molecular biology* 39, 503-508 (2008).
78. Chen, B. et al. Einstein Contained Aerosol Pulmonizer (ECAP): Improved Biosafety for Multi-Drug Resistant (MDR) and Extensively Drug Resistant (XDR) *Mycobacterium tuberculosis* Aerosol Infection Studies. *Applied biosafety : Journal of the American Biological Safety Association* 16, 134-138 (2011).
79. Fernandes, C.A. & Vanbever, R. Preclinical models for pulmonary drug delivery. *Expert opinion on drug delivery* 6, 1231-1245 (2009).
80. Myllymaki, H., Niskanen, M., Oksanen, K.E. & Ramet, M. Animal models in tuberculosis research - where is the beef? *Expert opinion on drug discovery* 10, 871-883 (2015).
81. Gupta, U.D. & Katoch, V.M. Animal models of tuberculosis. *Tuberculosis (Edinb)* 85, 277-293 (2005).
82. van de Vosse, E., van Dissel, J.T. & Ottenhoff, T.H. Genetic deficiencies of innate immune signalling in human infectious disease. *The Lancet. Infectious diseases*

- 9, 688-698 (2009).
83. Grumbach, F., Canetti, G., Grosset, J. & le Lirzin, M. Late results of long-term intermittent chemotherapy of advanced, murine tuberculosis: limits of the murine model. *Tubercle* 48, 11-26 (1967).
 84. Karakousis, P.C. et al. Dormancy phenotype displayed by extracellular *Mycobacterium tuberculosis* within artificial granulomas in mice. *J Exp Med* 200, 647-657 (2004).
 85. Mollenkopf, H.J., Kursar, M. & Kaufmann, S.H. Immune response to postprimary tuberculosis in mice: *Mycobacterium tuberculosis* and *Mycobacterium bovis* bacille Calmette-Guerin induce equal protection. *The Journal of infectious diseases* 190, 588-597 (2004).
 86. Hunter, R.L., Jagannath, C. & Actor, J.K. Pathology of postprimary tuberculosis in humans and mice: contradiction of long-held beliefs. *Tuberculosis (Edinb)* 87, 267-278 (2007).
 87. Young, D. Animal models of tuberculosis. *European journal of immunology* 39, 2011-2014 (2009).
 88. Apt, A.S. Are mouse models of human mycobacterial diseases relevant? *Genetics says: 'yes!'*. *Immunology* 134, 109-115 (2011).
 89. Tonjum, T., Welty, D.B., Jantzen, E. & Small, P.L. Differentiation of *Mycobacterium ulcerans*, *M. marinum*, and *M. haemophilum*: mapping of their relationships to *M. tuberculosis* by fatty acid profile analysis, DNA-DNA hybridization, and 16S rRNA gene sequence analysis. *Journal of clinical microbiology* 36, 918-925 (1998).
 90. Clark, H.F. & Shepard, C.C. Effect of environmental temperatures on infection with *mycobacterium marinum* (balnei) of mice and a number of poikilothermic species. *Journal of bacteriology* 86, 1057-1069 (1963).
 91. Tobin, D.M. & Ramakrishnan, L. Comparative pathogenesis of *Mycobacterium marinum* and *Mycobacterium tuberculosis*. *Cellular microbiology* 10, 1027-1039 (2008).
 92. Meijer, A.H. & Spalink, H.P. Host-pathogen interactions made transparent with the zebrafish model. *Current drug targets* 12, 1000-1017 (2011).
 93. Swaim, L.E. et al. *Mycobacterium marinum* infection of adult zebrafish causes caseating granulomatous tuberculosis and is moderated by adaptive immunity. *Infect.Immun.* 74, 6108-6117 (2006).
 94. Davis, J.M. et al. Real-time visualization of mycobacterium-macrophage interactions leading to initiation of granuloma formation in zebrafish embryos. *Immunity*. 17, 693-702 (2002).
 95. Meeker, N.D. & Trede, N.S. Immunology and zebrafish: spawning new models of human disease. *Developmental and comparative immunology* 32, 745-757 (2008).
 96. Lesley, R. & Ramakrishnan, L. Insights into early mycobacterial pathogenesis from the zebrafish. *Current opinion in microbiology* 11, 277-283 (2008).
 97. Davidson, A.J. & Zon, L.I. The 'definitive' (and 'primitive') guide to zebrafish hematopoiesis. *Oncogene* 23, 7233-7246 (2004).
 98. Lam, S.H., Chua, H.L., Gong, Z., Lam, T.J. & Sin, Y.M. Development and maturation of the immune system in zebrafish, *Danio rerio*: a gene expression profiling, in situ hybridization and immunological study. *Dev.Comp Immunol.* 28, 9-28 (2004).
 99. Herbolmel, P., Thisse, B. & Thisse, C. Ontogeny and behaviour of early macrophages

- in the zebrafish embryo. *Development* 126, 3735-3745 (1999).
100. Willett, C.E., Cortes, A., Zuasti, A. & Zapata, A.G. Early hematopoiesis and developing lymphoid organs in the zebrafish. *Developmental dynamics : an official publication of the American Association of Anatomists* 214, 323-336 (1999).
 101. Mathias, J.R. et al. Characterization of zebrafish larval inflammatory macrophages. *Developmental and comparative immunology* 33, 1212-1217 (2009).
 102. Warga, R.M., Kane, D.A. & Ho, R.K. Fate mapping embryonic blood in zebrafish: multi- and unipotential lineages are segregated at gastrulation. *Developmental cell* 16, 744-755 (2009).
 103. Kumar, H., Kawai, T. & Akira, S. Pathogen recognition in the innate immune response. *Biochem J* 420, 1-16 (2009).
 104. van der Sar, A.M. et al. *Mycobacterium marinum* strains can be divided into two distinct types based on genetic diversity and virulence. *Infect.Immun.* 72, 6306-6312 (2004).
 105. Benard, E.L. et al. Infection of zebrafish embryos with intracellular bacterial pathogens. *J.Vis.Exp.* (2012).
 106. Matty, M.A., Roca, F.J., Cronan, M.R. & Tobin, D.M. Adventures within the speckled band: heterogeneity, angiogenesis, and balanced inflammation in the tuberculous granuloma. *Immunological reviews* 264, 276-287 (2015).
 107. Stoop, E.J. et al. Zebrafish embryo screen for mycobacterial genes involved in the initiation of granuloma formation reveals a newly identified ESX-1 component. *Dis.Model.Mech.* (2011).
 108. Meijer, A.H. et al. Identification and real-time imaging of a myc-expressing neutrophil population involved in inflammation and mycobacterial granuloma formation in zebrafish. *Developmental & Comparative Immunology* 32, 36-49 (2008).
 109. Volkman, H.E. et al. Tuberculous granuloma induction via interaction of a bacterial secreted protein with host epithelium. *Science* 327, 466-469 (2010).
 110. Volkman, H.E. et al. Tuberculous granuloma formation is enhanced by a mycobacterium virulence determinant. *PLoS Biol* 2, e367 (2004).
 111. Davis, J.M. & Ramakrishnan, L. The role of the granuloma in expansion and dissemination of early tuberculous infection. *Cell* 136, 37-49 (2009).
 112. Stinear, T.P. et al. Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. *Genome Res* 18, 729-741 (2008).
 113. Tobin, D.M. et al. The *Ita4h* locus modulates susceptibility to mycobacterial infection in zebrafish and humans. *Cell* 140, 717-730 (2010).
 114. Clay, H., Volkman, H.E. & Ramakrishnan, L. Tumor necrosis factor signaling mediates resistance to mycobacteria by inhibiting bacterial growth and macrophage death. *Immunity* 29, 283-294 (2008).
 115. Clay, H. et al. Dichotomous role of the macrophage in early *Mycobacterium marinum* infection of the zebrafish. *Cell host & microbe* 2, 29-39 (2007).
 116. Ramakrishnan, L. Using *Mycobacterium marinum* and its hosts to study tuberculosis. *Current Science* 86, 82-92 (2004).
 117. Huminer, D. et al. Aquarium-borne *Mycobacterium marinum* skin infection. Report of a case and review of the literature. *Archives of dermatology* 122, 698-703 (1986).
 118. Ramakrishnan, L. The zebrafish guide to tuberculosis immunity and treatment.

- Cold Spring Harbor symposia on quantitative biology 78, 179-192 (2013).
119. Cronan, M.R. & Tobin, D.M. Fit for consumption: zebrafish as a model for tuberculosis. *Disease models & mechanisms* 7, 777-784 (2014).
 120. Meijer, A.H. Protection and pathology in TB: learning from the zebrafish model. *Seminars in Immunopathology* (2015).
 121. Kruh-Garcia, N.A., Wolfe, L.M. & Dobos, K.M. Deciphering the role of exosomes in tuberculosis. *Tuberculosis (Edinb)* 95, 26-30 (2015).
 122. Sharma, S. et al. Differential proteomics approach to identify putative protective antigens of *Mycobacterium tuberculosis* presented during early stages of macrophage infection and their evaluation as DNA vaccines. *Indian J Exp Biol* 53, 429-439 (2015).
 123. Vargas-Romero, F. et al. Secretome profile analysis of hypervirulent *Mycobacterium tuberculosis* CPT31 reveals increased production of EsxB and proteins involved in adaptation to intracellular lifestyle. *Pathogens and disease* 74 (2016).
 124. He, Y., Li, W., Liao, G. & Xie, J. *Mycobacterium tuberculosis*-specific phagosome proteome and underlying signaling pathways. *J Proteome Res* 11, 2635-2643 (2012).



Chapter 2

Identifying Proteins in Zebrafish Embryos Using Spectral Libraries Generated from Dissected Adult Organs and Tissues

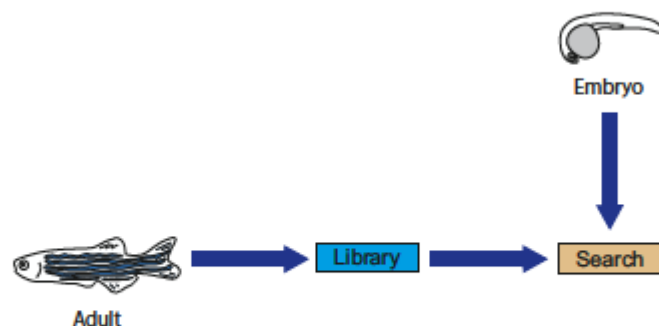
**S.J. van der Plas-Duivesteijn, Y. Mohammed,
H. Dalebout, A.H. Meijer, A. Botermans, J.L. Hoogendijk,
A.A. Henneman, A.M. Deelder, H.P. Spaink and
M. Palmblad**

Journal of Proteome Research 13, 2014: 1537-1544

Reprinted (adapted) with permission, copyright (2014) American Chemical Society

ABSTRACT

Spectral libraries provide a sensitive and accurate method for identifying peptides from tandem mass spectra, complementary to searching genome-derived databases or sequencing *de novo*. Their application requires comprehensive libraries including peptides from low-abundant proteins. Here we describe a method for constructing such libraries using biological differentiation to “fractionate” the proteome by harvesting adult organs and tissues, and build comprehensive libraries for identifying proteins in zebrafish (*Danio rerio*) embryos and larvae (an important and widely used model system). Hierarchical clustering using direct comparison of spectra was used to prioritize organ selection. The resulting and publicly available library covers 14,164 proteins, significantly improved the number peptide-spectrum matches in zebrafish developmental stages and can be used on data from different instruments and laboratories. The library contains information on tissue and organ expression of these proteins and is also applicable for adult experiments. The approach itself is not limited to zebrafish, but would work for any model system.



Abstract figure. Build a spectral library of zebrafish organs and to better identify peptides and proteins in zebrafish embryos.

INTRODUCTION

Proteomics has become a powerful tool for answering biological questions (<http://www.nature.com/reviews/focus/proteomics/index.html>). In a standard bottom-up proteomics experiment, peptides are identified by matching experimental tandem mass spectra with those predicted from hypothetical peptides derived from the genome sequence. Complementarily, tandem mass spectra can also be matched against previously acquired and identified spectra. It has been shown (1, 2) that this is a significantly more sensitive method, given sufficiently comprehensive spectral libraries. Such libraries are already available for multiple organisms including human, yeast and *Escherichia coli*. The key additional information provided by spectral libraries is the relative intensities of fragment ions in the tandem mass spectra. These are not trivial to predict accurately *in silico*. Furthermore, even the largest spectral libraries represent a much smaller search space than all peptides predicted from the genome (even after conservative gene prediction), decreasing the time necessary to search large datasets. Combining the

results of different studies, many of the same peptides are observed again and again for the same protein. Spectral libraries also provide a practical and useful scheme for collecting and organizing such tandem mass spectrometry data, allowing researchers to benefit from the work of others and providing a gold mine for determining good peptides and transitions for selected- or multiple reaction monitoring (3), statistics on protein expression, peptide observability and cataloguing post-translational modifications (4).

Some proteins are ubiquitous and likely to be found in many samples, across cell types and tissues, whereas others are expressed at much lower levels or only under very specific conditions, and thus less likely to be observed in most experiments. There is also a strong bias at the peptide level, with some peptides far more likely than others to be detected, given the chromatographic separation, ionization mechanism and fragmentation method.

Zebrafish embryos and larvae, widely used as model organism in a variety of research areas (e.g. development (5), regeneration (5), immunity (6-8), infectious disease (7, 9), cancer (10) and aging (11)) express many low-abundant proteins due to rapid differentiation. Although plenty of tandem mass spectrometry data from zebrafish is already publicly available in repositories such as PRIDE (12), according to the authors knowledge no spectral library for zebrafish is currently available (January 2014). Building a spectral library from all available zebrafish tandem mass spectrometry data – which in particular consist of zebrafish embryos and larvae – available in the public domain would result in missing low-abundant proteins that might play a key role solving important biological questions. Complementary to embryos and larvae, organs and tissues of adult zebrafish are differentiated and highly specialized, expressing many proteins that probably could not be detected in whole embryos and larvae. We describe a different and complementary approach based on a novel idea: using biological differentiation of adult organs and tissues for creating a spectral library to improve peptide and protein identification in embryos and larvae. The goal was to eventually generate a comprehensive spectral library that include all commonly observed, or observable, peptides from all proteins expressed in 1-6 day old zebrafish embryos and larvae (our model system; later referred to as zebrafish developmental stages). We also discuss how the spectral content of organ-specific tandem mass spectrometry datasets can be directly compared using a hierarchical clustering method developed for molecular phylogenetics (13). This information helps prioritize which sets of organs, tissues or fractions should be selected for in-depth analysis to add the greatest amount of nonredundant information to the spectral library.

MATERIALS AND METHODS

Zebrafish

Zebrafish (*Danio rerio*) were handled and maintained in compliance with the local animal welfare regulations and to standard protocols (14). Dissecting experiments were approved by the local animal welfare committee (DEC) of Leiden University under DEC nr. 11221. Organs of male and female zebrafish of approximately 4.5 months old

(strain AB) were dissected. Before dissection, the zebrafish were maintained at 28°C on a 14-h/10-h light/dark cycle. The day before dissection, the fish were withheld food to reduce the amount of food matter in the intestine.

Organ dissection

The zebrafish were anaesthetized with ice-water. Organ dissection started immediately after sedation (15). Within 15 minutes the following organs were collected: ovaries, testes, muscle (tail region), eyes, brain, intestine, fins, gills, heart, skin, swim bladder, spleen, kidney, and liver, pancreas and gall bladder. Each organ was placed in a Safe-Lock micro centrifuge tube (Eppendorf), snap-frozen in liquid nitrogen, and stored at -80°C. To get a convenient amount of protein to work with, we pooled some of the smaller organs (2 brains, eyes and intestines; 4 fins and livers including pancreas and gall bladders; 5 testes; 6 gills; 7 hearts; kidneys, skins and swim bladders; and 19 spleens).

Zebrafish developmental stages

To evaluate the spectral library, we obtained embryo samples using fertilized eggs of the same AB zebrafish line as used for dissection. Embryos were grown at 28.5 – 30°C in egg water (60 µg/ml Ocean Salts) and sampled at 1–6 dpf in Safe-Lock micro centrifuge tubes (Eppendorf), snap frozen in liquid nitrogen and stored at -80°C. Prior to collection we dechorionated 1 and 2 dpf embryos using 10 mg/mL Pronase (Merck Millipore) (14), and 1–4 dpf developmental stages were deyolked according to a previous described method (16, 17). Twenty of 1 dpf and ten of 2–6 dpf were sampled per developmental stage per eppendorf tube. Six replicates of each developmental stage were sampled.

Protein extraction and digestion

To make a hierarchical cluster comparison between organs and tissues proteomes, we extracted proteins of both male and female organs (gills, kidney, fins, brain, heart, eyes, muscle, skin, swim bladder, intestine, liver, testes, ovaries, and additionally a whole female sample) using urea buffer with 8M urea, 75mM NaCl, 50 mM Tris-HCL pH 8.2, 50 U/ml Benzonase (E1014-5KU, Sigma-Aldrich), 2 mM MgCl₂, and protease inhibitors (cOmplete ULTRA tablets, mini, EDTA-free, Roche) by homogenizing using 0.5 mm zirconium oxide beads and the Bullet Blender (Next Advance) at speed setting “8” for 3 minutes. Samples were placed at 4°C for 30 minutes and the supernatant collected after centrifugation at 16,000 × *g* for another 30 minutes at 4°C. A BCA assay (product #23235, Bio-Rad) was conducted to determine the protein concentration using bovine serum albumin as standard. In-solution digestion was performed as previously described (18), and the peptide digests were stored at -80°C until analysis. For testing the contents of the library, in-solution digests of 1–6 dpf embryos were made as previously described.

Guided by the comparative analysis of the hierarchical clustering, an in-depth library was built (figure 1), using: all female organs (gills, kidney, fins, brain, heart, eyes, muscle, skin, swim bladder, intestine, liver, spleen, and ovaries), including the male testes, liver, pancreas and gall bladder. In addition, both female and male separated gall bladders (five pooled) were added. A double protein extraction was performed on each pooled set of organs, one using 20 mM NaCl, 20 mM Tris-HCl pH 8.2, 50 U/mL benzonase and 2

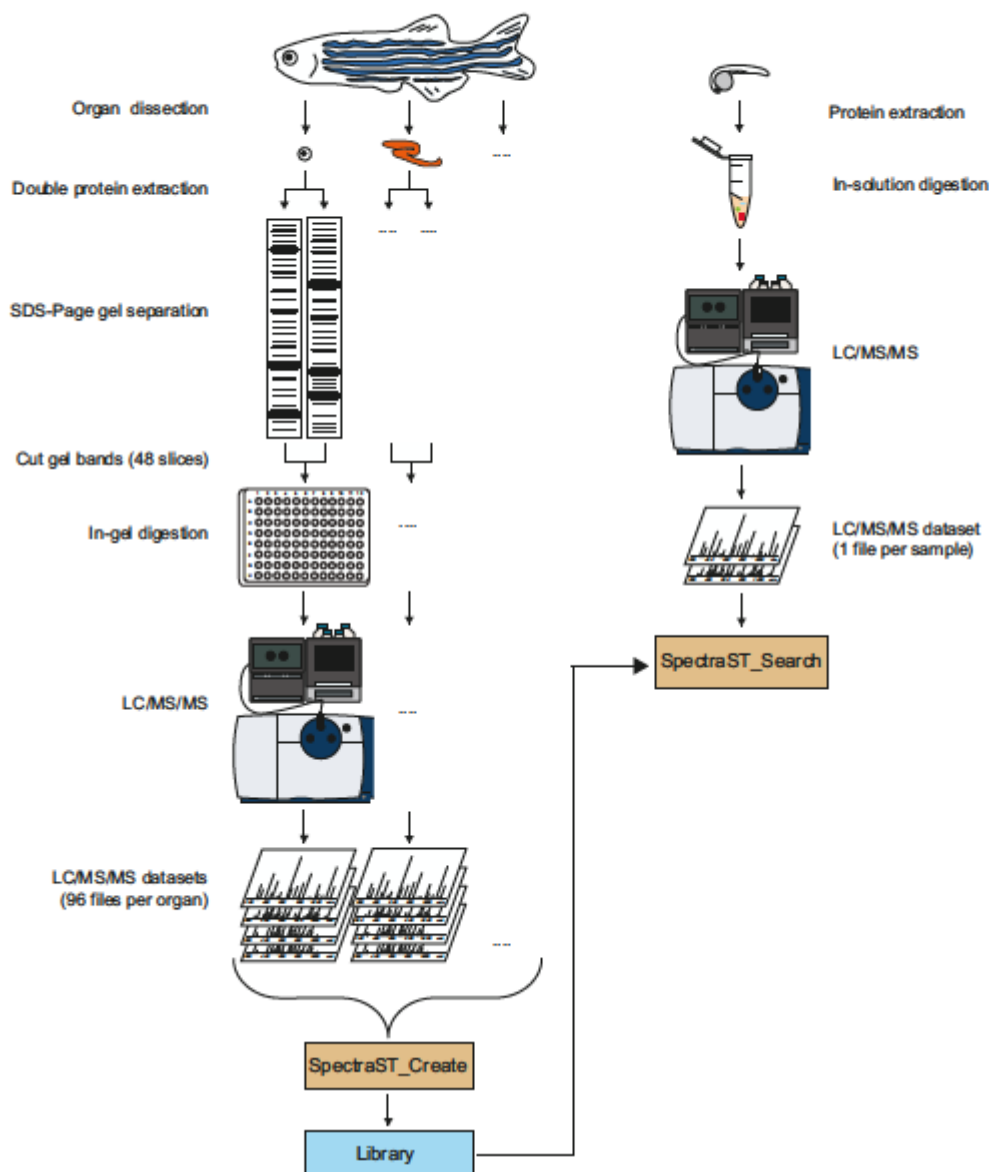


Figure 1. Experimental workflow of building and searching the zebrafish in-depth library. Tissues and organs of adult zebrafish were harvested to generate an in-depth MS/MS analysis (using gel- on top of LC- separation). An organ fractionated spectral library (A-library) was built. Spectra of 1-6 dpf zebrafish embryos were generated using a high-throughput method (in-solution digestion), and searched against the created library.

mM MgCl_2 , and one using 1% SDS, 50 U/mL benzonase, 2 mM MgCl_2 (both containing protease inhibitors). Both extractions were followed by homogenizing, centrifugation, and determining the protein concentration as previously described. Approximately 40 μg of each extraction was fractionated using Novex 4-12% Bis-Tris gels (NuPAGE, Invitrogen) with MOPS running buffer (Invitrogen). The gel was stained overnight (16 h) in colloidal Coomassie blue staining solution (Invitrogen) containing 5% methanol

and washed with milli-Q water afterwards. The gel lanes were cut into 48 1.5 × 5 mm bands with a disposable grid cutter (MEE1.5-5-48, the Gel Company), and transferred to 96-well PCR plates (GBO). In-gel digestion was performed as previously described (18). After pooling the supernatants, the samples were concentrated by speed-vacuum (RVC 2-25 CD plus, Salm en Kipp) and frozen at -35°C until analysis. For the additional AE-library, 1-6 dpf embryos were processed identically (in-gel digestion) to the organs as previously described.

Liquid Chromatography – Tandem Mass Spectrometry

Two µL (in-solution digests) or 10 µL (in-gel digested fractions) of each sample were loaded and desalted on a C18 PepMap 300 µm, 5 mm-i.d., 300 Å precolumn (Thermo Scientific) and separated by reversed-phase liquid chromatography using two identical 150 mm 0.3 mm-i.d. ChromXP C18CL, 120 Å columns (Eksigent) coupled parallel and connected to a split less NanoLC-Ultra 2D plus system (Eksigent) with a linear 90-minute (for in-solution digests) or 45-minute (for in-gel digests) gradient from 4 to 35% acetonitrile in 0.05% formic acid and a constant (4 µL/minute) flow rate. The LC system was coupled to an amaZon speed ETD ion trap (Bruker Daltonics) equipped with an Apollo II ESI source. After each MS scan, up to 10 abundant multiply charged species in m/z 300-1300 were selected for MS/MS and actively excluded for 1 minute after having been selected twice. Each individual scan or tandem mass spectrum was saved to the hard drive. The LC system was controlled by HyStar 3.2 and the ion trap by trapControl 7.1.

Building Spectral Libraries

We developed a scientific workflow using Taverna 2.4 to automatically generate spectral libraries (19) (figure 2). The workflow searches all spectra against the required database, filters the results and saves only PSMs with probabilities above the PeptideProphet default 5%, and then builds the library from all spectra with probabilities of at least 90%. These thresholds can be modified by the user. Taverna offers various types of processors (19). In our implementation we used mainly BeanShell processors, which enable executing small Java code snippets as part of a workflow. Typically these are used for small tasks like simple file and data manipulation, parsing and formatting, saving to a local directory, calling local program, interacting with the user and so on. Each BeanShell processor in our workflow is used to launch software with their correct inputs. Our Taverna workflow consists of four processing steps. First it uses CompassXport 3.0.5 to convert the raw ion trap data (stored in .yep format) to mzXML, which is an XML open standard for storing MSⁿ data (20). A list of peptide-spectrum matches (PSMs) is generated using X!Tandem (21) by searching each spectrum against the *Danio rerio* sequence database. In the third step, the workflow assigns a theoretical probability to each PSM by a mixture model using PeptideProphet (22, 23). In the end SpectraST (2) is used to build a library with the minimum probability provided as an input (in this case 90%). In the workflow execution and X!Tandem search used for this work, we assumed strict tryptic cleavage specificity (C-terminally for R and K, not N-terminally of P), a precursor mass measurement error of -0.5 to 2.5 Da was tolerated, two missed enzymatic cleavages were allowed, and carbamidomethylation of cysteines was considered as the only fixed modification.

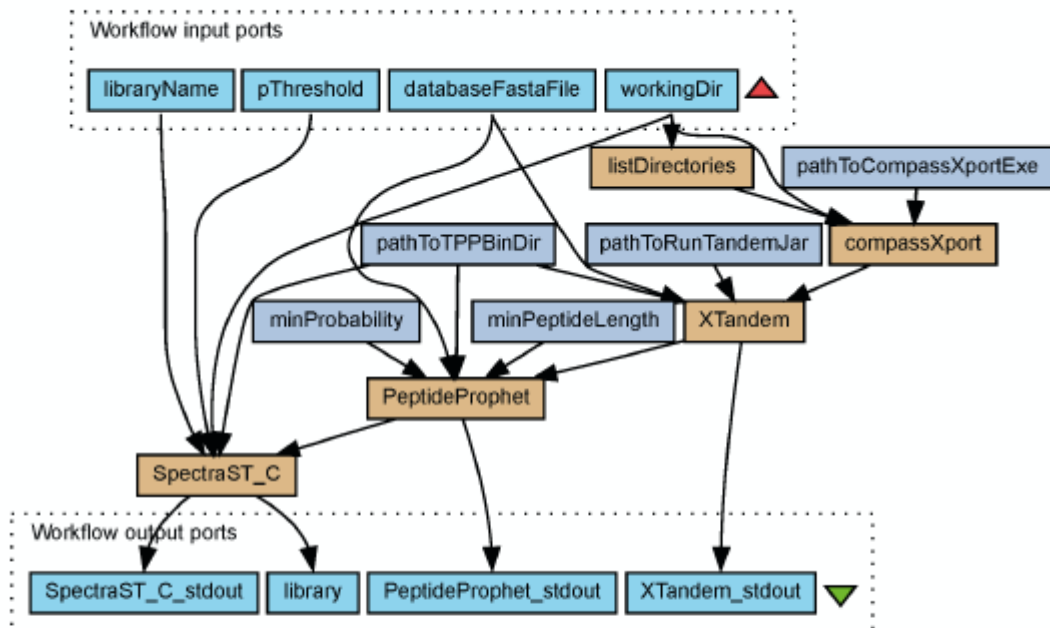


Figure 2. Scientific workflow for creating spectral libraries.

The workflow consists of five Java BeanShell processors (brown color). The `listDirectories` prepares a list of all `.d` directories used as input. The `compassXport` is a format converter and prepares the data into the right format for the `XTandem` search. `XTandem` is a database search engine. `PeptideProphet` performs statistical validation by fitting true and false positive score distributions to the measured discriminant scores and assigns a probability to each PSM. `SpectraST_C` builds the library including all PSMs that pass the provided threshold in `pThreshold`. All processors are Java BeanShells and run locally.

Hierarchical clustering

To help prioritize which set of organs, tissues or fractions should be selected for in-depth analysis, a simple “phylogenetic” tree was generated from raw spectral data as previously described (13). To construct a tree from the different organs and tissues, raw ion trap datasets of whole organ digests were converted to MGF using `DataAnalysis 4.0` (Bruker) and a maximum of 4,000 compounds. These were compared by `compareMS2` (13), creating a NEXUS (24) file, which was converted to the MEGA (25) format using an in-house script (available with the workflow), and read into MEGA version 5.0.5 to construct a UPGMA (26) tree.



Figure 3. Relationship between whole-organ LC/MS/MS datasets from dissected zebrafish.

Hierarchical clustering of LC/MS/MS datasets from dissected adult male and female zebrafish organs and tissues using a previously described method for phylogenetic analysis (13). Most organs show little difference between sexes, with the salient exceptions of the male and female reproductive glands and livers. The liver samples included the pancreas and gall bladder.

RESULTS

Hierarchical clustering of organ/tissue proteomes

Tandem mass spectra of zebrafish whole organ (in-solution) digests from both sexes were used to generate a “phylogenetic” tree (figure 3). Most organs show little difference between sexes (gills, kidney, fins, brain, heart, eyes, muscle, skin, swim bladder and intestine). Organs like the reproductive glands, and liver show stark differences between the sexes. Pairing of the female ovaries with the female liver in the hierarchical clustering can be explained by vitellogenesis, the process in which the protein vitellogenin is synthesized in the liver and transported via the blood stream to the ovaries (27). Vitellogenin is the main protein in the yolk of eggs, which are constantly produced in the ovaries of egg-laying females, including the 4.5 month adults used here. Because of the abundance of vitellogenin in the female liver (>10% of the total extracted protein according to spectral counts in our data) and ovaries, it is easy to see why these organs are grouped in hierarchical clustering. The proteome of the male liver is in this particular sense most similar to those of the male and female hearts, probably due to the large blood content, which were not removed before analysis, e.g. by perfusion. The analysis also includes a whole animal sample. Unsurprisingly, this clustered with the muscle

samples as the zebrafish body mainly consists of muscle tissue. Because muscle proteins are then most abundant in the whole animal, the whole-animal sample is most similar to the muscle samples.

When constructing the library, SDS-PAGE was used to first fractionate the proteins from the selected organs and tissues. This provides deeper coverage of the proteome, but is also more time-consuming. Hierarchical clustering served as a tool to select and prioritize the organ and tissue samples for this in-depth analysis. Organs like the reproductive glands and liver that showed stark differences between sexes were included from both sexes for this analysis.

Organ Differentiated Zebrafish Spectral Library

On the basis of the hierarchical clustering, all female organs, male reproductive glands and liver, and additionally the gall bladder from both sexes were dissected, analyzed and used to generate an adult zebrafish spectral library (figure 1 and 3). A Taverna (19) workflow was used to search all data against genome-derived sequences and combine the results into a spectral library (figure 2). Mass spectrometry data from both the in-depth analysis (SDS-PAGE fractionation and in-gel digestion), and from samples digested in-solution (already used for the hierarchical clustering) were included in the library. Henceforth, we refer to this library as the A-library (Adult library). The A-library contains 60,048 high quality consensus spectra (unique peptides) generated from 1,145,481 out of 16,534,113 (75.05 GB) spectra passing the probability threshold. The A-library was compared with an embryonic spectral library built from high-quality data from 24 to 32 hours post fertilization (hpf) zebrafish embryos previously published by Löffner *et al.*, (28). As small dataset probes, zebrafish developmental stages of 1-6 days post fertilization (dpf) digested in-solution were searched separately against both spectral libraries with SpectraST and the results were compared with a sequence database search using X!Tandem (figure 4). An unpaired *t* test was performed to calculate *p* values for the comparison between the X!Tandem sequence database search and the SpectraST A-library using GraphPad Prism (version 6.02). Searches against the A-Library resulted in more peptide spectrum matches (PSMs) at a 1% false discovery rate (FDR) compared to both the embryonic library and sequence database search. In total, 8,563 spectra were identified from this data by matching against the A-library. SpectraST provides better discrimination between correct and incorrect matches than X!Tandem regardless of the absolute number of correct PSMs. The results of any comparison are therefore dependent on the metric used, such as the FDR cutoff or the absolute area under the ROC curve. The *p* values should be seen as an indication of how these spectral library searches perform compared with X!Tandem sequence database searches.

In a library search, the spectra are compared against a library of previously identified spectra rather than against hypothetical predicted spectra from a sequence database. Libraries are shown to be much faster and capable of identifying low-quality spectra than sequence search engines (2), as they search a smaller space (fewer candidates to choose from) and use real reference spectra with known ion intensities as opposed to simplistically predicted intensities in the sequence search engines (29). The low-quality

spectra of the early developmental stages (1 and 2 dpf) therefore might have been expected to perform so much better when searching against the library than against the sequence database (figure 4). For all developmental stages, larger numbers of spectra could be identified using SpectraST and the A-library than using SpectraST and the embryo-derived library or an X!Tandem sequence database search. Because the embryo library was constructed from early stage (24-32 hpf) embryos, it is not surprising that the search against spectra of 1 and 2 dpf embryos resulted in a larger increase in the number of correct PSMs (155 and 217 compared to 44 and 100 PSMs, respectively) than the later developmental stages (3-6 dpf), showing a smaller difference between the two searches.

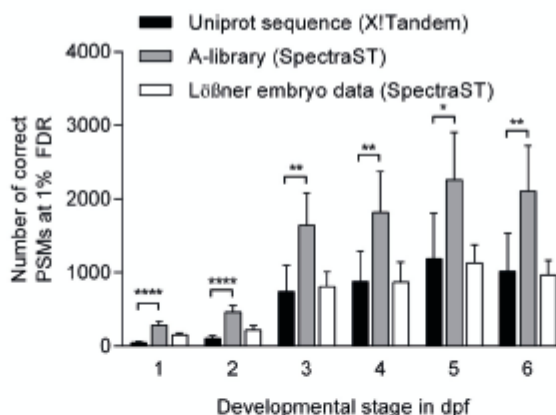


Figure 4. Improvement on peptide spectrum matches in embryos by using spectral libraries from adult organs. Searching data from different developmental stages (1-6 dpf) against the A-library using SpectraST resulted in more correct peptide-spectrum matches (PSMs) compared with an X!Tandem search against a Uniprot sequence database from *Danio rerio*. The SpectraST search against a spectral library from embryos only (compiled from a dataset previously published by Löffner *et al.* (28)), showed comparable results to the X!Tandem search. Each developmental stage was sampled six times and analysed separately by LC/MS/MS on an ion trap. The number of correct peptide spectrum matches at 1% false discovery rate (FDR) were averaged over all six runs per developmental stage defined in days post fertilization (dpf) ($n=6$). Embryo data from the previously published dataset were acquired on LTQ Velos ion trap mass spectrometer (Thermo Fisher). An unpaired *t* test was performed to compare the sequence database with the SpectraST A-library searches. **** p value < 0.0001 , ** p value < 0.01 , * p value < 0.05 .

During zebrafish development, the expression of transcription factors and many other proteins is up- and down regulated (30-33). In adults, some of these proteins are less abundant, making it more difficult to collect high-quality peptide spectra from these particular proteins with the untargeted methods used here. The A-library was built with the purpose of improving peptide and protein identification in studies using embryo model systems. To avoid missing important developmental data in our library, we generated an embryonic library from 1-6 dpf developmental stages by the same method used to construct the A-library. This library is combined with the A-library to form a consensus library of both, which is henceforth referred to as AE-library (Adult and Embryonic-library). The AE-library contained 66,158 high-quality consensus

spectra generated from 1,381,602 out of 22,170,118 spectra (107.68 GB) that passed the probability threshold. These consensus spectra/peptides described 14,164 proteins with a 1.1% protein-level FDR as estimated by ProteinProphet (34).

The goal of creating a spectral library is to use this for peptide and protein identification, possibly in combination with a sequence database search, and with additional statistical analysis using tools such as PeptideProphet and InterProphet. It is important that statistical models that a user would likely include when applying the library, such as weighing PSM probabilities using the number of sibling peptides, are not used more than once. Multiple application of the same statistical model multiplies the penalty (for say, observing few unique peptides for a protein) when in fact repeated observation of the same peptides from the same protein should increase confidence in these PSMs, especially for small proteins and cases where one can explain why only a few peptides are observed. The A- and AE- libraries were therefore constructed without applying further refinement steps such as InterProphet (35). While recommended for peptide/protein identification purposes, such stringent filters reduce the value of the libraries by preventing good PSMs for peptides with no or few siblings from passing the set probability threshold. These filters should be applied at most once, typically *after* using the library. However, we also generated two extra AE-libraries: one with applying a single InterProphet step on PSMs from female, male and developmental stages all together (the single InterProphet AE-library), and a second library by applying individual InterProphet steps on each organ separately, containing the atomic data (the individual InterProphet AE-library). Both the single and individual InterProphet AE-libraries contained in the first step more spectra than our normal AE-library, that is 1,622,083 for the single InterProphet AE-library and 1,579,159 for the individual InterProphet AE-library compared with 1,381,602 for our library with no InterProphet step. However, the consensus library contained 37,128 and 63,557 spectra expressing 5,280 and 8,120 proteins for the single and individual InterProphet libraries respectively, compared with 66,158 expressing 14,164 proteins for our library with no InterProphet step.

To investigate possible bias in using data from the same laboratory to both generate and search the spectral libraries, we used zebrafish datasets available from the Proteomics Identifications Database (PRIDE) (12) to validate the use of both the A- and AE-libraries compiled without an Interprophet step (figure 5). An unpaired *t* test was performed calculating the *p* values between the sequence database search and the SpectraST A- and AE- library searches against zebrafish fin data (36) using GraphPad Prism (version 6.02). Searches against both libraries resulted in significantly more PSMs at a 1% FDR. Overall, similar amounts of PSMs were found between our libraries and a sequence database when searched against different fractions of zebrafish brain (37) and 3 dpf embryos (38) acquired from PRIDE. In two cases, the brain and the 3 dpf samples with PRIDE references 12052 and 2104 respectively, the sequence database search performed slightly better than both libraries. Even though the PRIDE data was analyzed with other types of mass analyzers than the ion traps used to build the libraries, these PRIDE datasets could be searched using our libraries with at least as good or better results than a sequence database search.

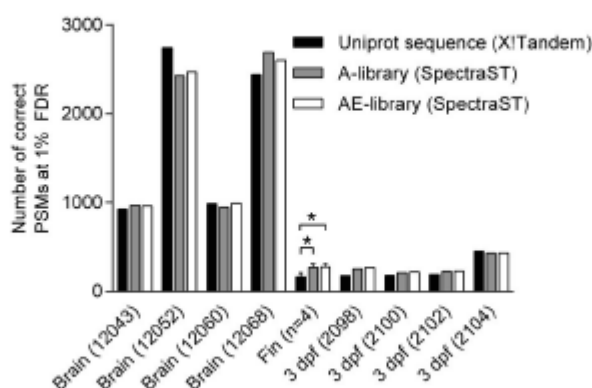


Figure 5. Validation of the A- and AE-libraries against data from other laboratories. Spectra from different datasets available from PRIDE were searched against our A- and AE-libraries using both SpectraST and X!Tandem with the zebrafish UniProt sequence database. Both libraries identified significantly more PSMs compared with the X!Tandem search on all zebrafish fin ($n=4$) data (PRIDE references: 13683, 13686, 13687, 13690). Both libraries also identified a similar amount of correct PSMs compared to the X!Tandem search in PRIDE datasets from zebrafish brain and 3 dpf embryos. The dataset numbers ($n=1$) for brain and 3 dpf embryo data represent PRIDE references. The PRIDE datasets of the brain were acquired on a MALDI TOF/TOF (Applied Biosystems) (37), the fin data with a LTQ ion trap mass spectrometer (Thermo Fisher) (36) and the 3 dpf embryos also on a Thermo Finnigan liquid ion trap mass spectrometer (38). The p values were calculated as in figure 4). * p value < 0.05 .

DISCUSSION

This is the first study in which *adult* organs and tissues were specifically harvested and analyzed to build spectral libraries for peptide identifications from *embryos and larvae*. We have demonstrated that biological differentiation (by using adults instead of developmental stages) is an extremely powerful “protein fractionation” method for constructing spectral libraries. All bioinformatics tools, from the optimization of the experimental design (hierarchical clustering) to the data analysis (X!Tandem and SpectraST) and scientific workflows (Taverna) for remote high-performance computing, were already described in the literature and are all available as open source. This work therefore also illustrates the power and efficiency of combining available tools using scientific workflow managers.

Hierarchical clustering based on shared spectral content may be a generally useful tool to prioritize which fractions or samples to analyze in-depth to construct large, comprehensive and non redundant spectral libraries. Here we used the results from hierarchical clustering to restrict the analyses from 26 (all organs of both sexes) to 14 (all female organs plus male liver and testes). Although it has been shown that compareMS2 analysis can be used for molecular phylogenetics *across* species (13), it also can measure similarity between organs and tissues *within* a species (current study).

Reducing protein or peptide complexity by multiple fractionation techniques (e.g.

by SDS-PAGE and/or LC separation) is an efficient tool to increase the number of identifications (39) and to learn more about the proteins themselves. In this work, we built rich spectral libraries from fractionated organs and tissues. This idea can be pursued further, for instance with fractionation on the cellular or even subcellular level, using fluorescent activated cell sorting (FACS) or organelle fractionation techniques. This can be done using cells from adults as well as from zebrafish developmental stages, and would be expected to produce even richer spectral libraries as well as catalogue more information about the expression and localization of the proteins themselves. All such experiments can be combined with additional enrichment and analysis strategies for post-translational modifications of interest, for example phosphorylation, glycosylation.

The spectral libraries work best when applied to data from similar instruments in the same laboratory, but also works well when used with data from different instruments generated in other laboratories. Search results using this library will only improve by expanding the library with data from other laboratories. We here compared spectral library and sequence database searches directly. Libraries can naturally also be used in combination with sequence databases, for example using InterProphet. As shown with the PRIDE brain dataset, incomplete library coverage is found to be the key issue that needs to be addressed before the method can completely replace sequence searching methods when the discovery of new peptides or proteins remains a major experimental goal (2). We therefore recommend using both the combined AE-library and a sequence database search to identify peptides and proteins from zebrafish embryos and larvae. All spectral libraries are publicly available (in SpectraST .splib format) and can be used for either protein identification from embryonic samples, exploration of the zebrafish proteome (33), or as a foundation for building larger zebrafish spectral libraries.

CONCLUSION

In this study, we have demonstrated the use of one biological system, adult zebrafish, to generate high-quality tandem mass spectra and build a spectral library to use for the identification of peptides and proteins in a different system, zebrafish embryos and larvae. The method essentially uses biological differentiation or development to fractionate the proteome for spectral library building. As part of this work, we generated a Taverna scientific workflow for creating spectral libraries from millions of tandem mass spectra. This workflow is made freely available to the research community. We have also demonstrated the use of a hierarchical clustering method, originally developed for molecular phylogenetics, to prioritize experiments for spectral library building independent of any sequence database search. The zebrafish libraries built in this study are the first spectral libraries for zebrafish, an important model organism, and are also made freely available. In addition, the library provides information on organ-specific protein expression in zebrafish. All methods described here are not limited to zebrafish, but can be used for any biological system.

SUPPORTING INFORMATION

The Taverna workflow and both libraries (A and AE) are available free of charge via the Internet at: <https://www.ms-utils.org/zebrafish>

ACKNOWLEDGMENTS

We thank Ulricke Nehrdich and Davy de Witt for fish care. Support was obtained from The Netherlands Organization for Scientific Research (NWO) Vidi grant 917.11.398 (M.P.).

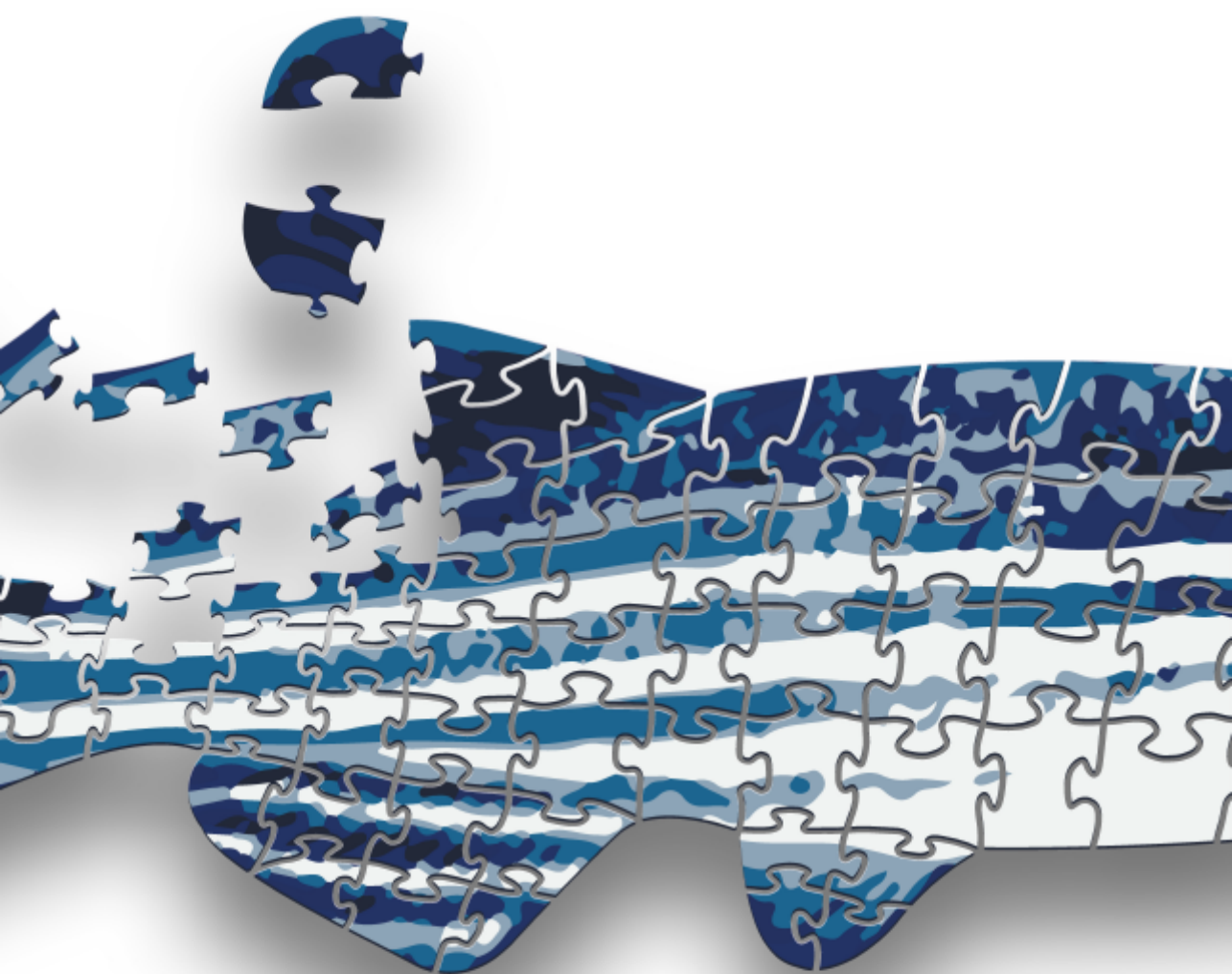
REFERENCES

1. Lam, H. et al. Building consensus spectral libraries for peptide identification in proteomics. *Nat.Methods* 5, 873-875 (2008).
2. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7, 655-667 (2007).
3. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular systems biology* 4, 222 (2008).
4. Deutsch, E.W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* 9, 429-434 (2008).
5. Brittijn, S.A. et al. Zebrafish development and regeneration: new tools for biomedical research. *International Journal of Developmental Biology* 53, 835-850 (2009).
6. Traver, D. et al. The zebrafish as a model organism to study development of the immune system. *Adv.Immunol.* 81, 253-330 (2003).
7. Phelps, H.A. & Neely, M.N. Evolution of the zebrafish model: from development to immunity and infectious disease. *Zebrafish* 2, 87-103 (2005).
8. Trede, N.S., Langenau, D.M., Traver, D., Look, A.T. & Zon, L.I. The use of zebrafish to understand immunity. *Immunity* 20, 367-379 (2004).
9. Cui, C. Infectious disease modeling and innate immune function in zebrafish embryos. *Methods in cell biology* 105, 273 (2011).
10. Mione, M., Meljer, A.H., Snaar-Jagalska, B.E., Spalink, H.P. & Trede, N.S. Disease modeling in zebrafish: cancer and immune responses—a report on a workshop held in Spoleto, Italy, July 20-22, 2009. *Zebrafish* 6, 445-451 (2009).
11. Gerhard, G.S. Comparative aspects of zebrafish (*Danio rerio*) as a model for aging research. *Exp.Gerontol.* 38, 1333-1341 (2003).
12. Vizcaíno, J.A. et al. The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research* 41 (2013).
13. Palmblad, M. & Deelder, A.M. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun.Mass Spectrom.* 26, 728-732 (2012).
14. Westerfield, M. The zebrafish book. A guide for the laboratory use of zebrafish (*Danio rerio*). Edn. 4. (Univ. of Oregon Press, Eugene; 2000).
15. Gupta, T. & Mullins, M.C. Dissection of organs from the adult zebrafish. *J.Vis.Exp.* (2010).
16. Link, V., Shevchenko, A. & Heisenberg, C.P. Proteomics of early zebrafish embryos. *BMC.Dev.Biol.* 6, 1 (2006).
17. Lin, Y., Chen, Y., Yang, X., Xu, D. & Liang, S. Proteome analysis of a single zebrafish embryo using three different digestion strategies coupled with liquid

- chromatography-tandem mass spectrometry. *Anal.Biochem.* 394, 177-185 (2009).
18. Mostovenko, E., Deelder, A.M. & Palmblad, M. Protein expression dynamics during *Escherichia coli* glucose-lactose diauxie. *BMC.Microbiol.* 11, 126 (2011).
 19. Wolstencroft, K. et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research* (2013).
 20. Pedrioli, P.G. et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nat.Biotechnol.* 22, 1459-1466 (2004).
 21. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics.* 20, 1466-1467 (2004).
 22. 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.* 74, 5383-5392 (2002).
 23. Ma, K., Vitek, O. & Nesvizhskii, A. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* 13, S1 (2012).
 24. Maddison, D.R., Swofford, D.L. & Maddison, W.P. NEXUS: an extensible file format for systematic information. *Syst Biol* 46 (1997).
 25. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol.Biol.Evol.* 24, 1596-1599 (2007).
 26. Sokal, R. & Michener, C. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38 (1958).
 27. Arukwe, A. & Goksoyr, A. Eggshell and egg yolk proteins in fish: hepatic proteins for the next generation: oogenetic, population, and evolutionary implications of endocrine disruption. *Comp Hepatol.* 2, 4 (2003).
 28. Lossner, C. et al. Expanding the zebrafish embryo proteome using multiple fractionation approaches and tandem mass spectrometry. *Proteomics.* 12, 1879-1882 (2012).
 29. Deutsch, E.W. et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 10, 1150-1159 (2010).
 30. Jessen, J.R., Willett, C.E. & Lin, S. Artificial chromosome transgenesis reveals long-distance negative regulation of *rag1* in zebrafish. *Nat Genet* 23, 15-16 (1999).
 31. Grandel, H. et al. Retinoic acid signalling in the zebrafish embryo is necessary during pre-segmentation stages to pattern the anterior-posterior axis of the CNS and to induce a pectoral fin bud. *Development* 129, 2851-2865 (2002).
 32. Qi, H.H. et al. Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. *Nature* 466, 503-507 (2010).
 33. Abramsson, A. et al. Proteomics profiling of single organs from individual adult zebrafish. *Zebrafish.* 7, 161-168 (2010).
 34. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal.Chem.* 75 (2003).
 35. Shteynberg, D. et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* 10 (2011).
 36. Singh, S.K., Lakshmi, M.G., Saxena, S., Swamy, C.V. & Idris, M.M. Proteome profile of zebrafish caudal fin based on one-dimensional gel electrophoresis LCMS/MS and two-dimensional gel electrophoresis MALDI MS/MS analysis. *J.Sep.Sci.* 34,

225-232 (2011).

37. Singh, S.K., Rakesh, K.S., Ramamoorthy, K., Pardha Saradhi, A.V. & Idris, M.M. Proteome Profile of Zebrafish Brain Based on Gel LC-ESI MS/MS Analysis. *J Proteomics Bioinform* 3, 135-142 (2010).
38. Lucitt, M.B. et al. Analysis of the zebrafish proteome during embryonic development. *Mol.Cell Proteomics*. 7, 981-994 (2008).
39. Antberg, L, Cifani, P, Sandin, M., Levander, F. & James, P. Critical Comparison of Multidimensional Separation Methods for Increasing Protein Expression Coverage. *Journal of Proteome Research* 11, 2644-2652 (2012).



Chapter 3

Differentiating Samples and Experimental Protocols by Direct Comparison of Tandem Mass Spectra

S.J. van der Plas-Duivesteijn, T. Wulff, O. Klychnikov, D. Ohana, H. Dalebout, P.A. van Veelen, J. de Keijzer, M.A. Nessen, Y.E.M. van der Burgt, A.M. Deelder and M. Palmblad

Rapid Communications in Mass Spectrometry 30,
2016: 731-738

ABSTRACT

CH
3

Peptide tandem mass spectra can be analyzed by a number of means. They can be compared against predicted spectra of peptides derived from genome sequences, compared against previously acquired and identified spectra, or – sometimes – sequenced *de novo*. We recently introduced another method which compares spectra between liquid chromatography/tandem mass spectrometry (LC/MS/MS) datasets to determine the shared spectral content, and demonstrated how this can be applied in a molecular phylogenetic study using sera from human and non-human primates. We will here explore if such a method have other, serendipitous uses. We used the existing compareMS2 algorithm without modification on a diverse set of experiments. First we conducted a small phylogenetic study, using (mammalian) bone samples to study old material, and human pathogens aiming to distinguish clinically important strains. Although not as straightforward as primate sera analysis, the method shows significant promise for all these applications. We also used the algorithm to compare 24 different protocols for extraction of proteins from muscle tissue. The results provided useful information in comparing protocols. Finally, we applied compareMS2 aiming for quality control of two traceable protein reference standards (troponin) used in clinical chemistry assays, by analysing the effect of storage conditions. The results illustrate a broad applicability of the metric based on shared tandem mass spectra between LC/MS/MS datasets for analysing protein digests in different types of experiments. There is no reason to assume that our instance of this method is optimal in any of these situations, as it makes limited or no use of accurate mass and chromatographic retention time. We propose that with further improvement and refinement, this type of analysis can be applied as a simple but informative first step in many pipelines for bottom-up tandem mass spectrometry data analysis in proteomics and other fields, comparing or analysing large numbers of samples or datasets.

INTRODUCTION

Peptide tandem mass spectra can be analysed in several different ways, either individually or collectively. They can be compared against spectra predicted from a sequence database, compared against previously acquired, identified and collected spectra in spectral libraries, or, if the quality is sufficient, be read as an amino acid sequence *de novo*. We recently introduced a different method (1), in which spectra are compared between liquid chromatography/tandem mass spectrometry (LC/MS/MS) datasets to determine the shared spectral content and thereby infer something on the relationship between these datasets. In celebration of the 50th anniversary of the seminal paper on molecular phylogenetics by Zuckerkandl, Jones and Pauling (2), we applied this new method in a partial reproduction of the original study of the evolutionary relationship between human and non-human primates using patterns of tryptic peptides – replacing paper chromatography by tandem mass spectrometry. In planning the acquisition of tandem mass spectrometry data for a zebrafish spectral library, we used the software developed for the primate study, compareMS2, to measure the overlap between different tissue preparations of adult zebrafish and prioritize their analysis (3). The resulting spectral

library is now available from NIST (4). A similar method, collecting reference (but unidentified) spectra in such spectral libraries, was independently developed by Önder and co-workers for the identification of tick blood meals (5). Lam and co workers have also applied spectral “libraries” of small molecules for chemical fingerprinting and species identification of marine bacteria (6). Simultaneously, Wulff *et al.* (7) showed how an almost identical approach could be taken to the identification of fish from (prepared) food products. This work has later been extended to mammal, bird and mollusc food species (8) and examined in detail in an effort to discriminate between closely related flatfish (9). Of more clinical relevance, we have also applied this spectral library method to identification of mycobacterium species and different strains of *M. tuberculosis* and *Staphylococcus aureus* (10).

The idea of comparing biological samples by directly comparing large numbers of tandem mass spectra, acquired by established bottom-up proteomics methodology using LC/MS/MS has practical applications ranging from molecular phylogenetics to experiment design, forensics and investigation of entomological feeding habits. How widely then is this idea (practically) applicable? From a philosophical point of view, we should want to avoid reinventing any proverbial wheels and “not invented here” principles. At the same time, we do not wish to overstretch the application of a favorite tool instead of finding and learning how to use an existing tool, which may be fundamentally more appropriate for the task at hand. Nevertheless, we will here explore if methods based on this idea have other, perhaps serendipitous, uses, no matter how particular or peculiar, in the hope that it may stimulate thinking about different ways of designing experiments and analyzing mass spectrometry data. We will not, however argue that the methods would be superior to others in any particular field or a panacea for all mass spectrometry data analysis problems.

MATERIALS AND METHODS

Samples for molecular phylogenetics

To extend the application of the method for molecular phylogenetics to different and more difficult types of samples, frozen femoral bone from chicken (*Gallus gallus*), cow (*Bos taurus*), fallow deer (*Dama dama*), moose (*Alces alces*) and wild boar (*Sus scrofa*) were cut into 10 × 2 mm pieces, washed and extracted at 4°C following a similar protocol to that described by Jiang *et al.* (11). The bone pieces were first washed overnight (O/N) in phosphate buffered saline (PBS). After removing the PBS, the bone samples were incubated for 24 hours in 1.2 M HCl. The remaining bone samples were washed twice in milli-Q water and incubated in 100 mM Tris, 6 M guanidine-HCl at pH 7.4 with the addition of protease inhibitor 1:5 (cOmplete ULTRA Tablets, Mini, EDTA-free, EASYpack, Roche Diagnostics) for 72 hours. The residues were then incubated a final time with 100 mM Tris, 6 M guanidine-HCl and 0.5 M tetrasodium EDTA (again with the protease inhibitor) at pH 7.4, for 72 hours. After centrifugation for 10 minutes at 16,000 × *g* and 4°C, the final supernatants were collected in fresh Eppendorf tubes and stored at -80 °C until further processing.

After thawing the protein extracts at room temperature, the proteins were precipitated by adding four volumes of ice-cold acetone and placed at -20°C O/N. After incubation and centrifugation for 15 minutes at $16,000 \times g$ and 4°C , the supernatants were discarded and the pellets freeze-dried (Christ, Salm en Kipp), and redissolved in 50 mM ammonium bicarbonate. Protein concentration was determined by BCA (#23235, Bio-Rad) using the supplied bovine serum albumin as reference. The proteins were then digested in solution, starting with the addition of 10 mM Dithiothreitol (DTT) for 45 minutes at 56°C , followed by adding 25 mM iodoacetamide (both Sigma Aldrich), and left for one hour at room temperature. The proteins were digested using trypsin (sequencing grade modified, Promega) 1:100 and incubate overnight at 37°C . To quench the digestion, 10% TFA was added to the sample and centrifuged at $2,500 g$ for 10 minutes at room temperature. The resulting peptide mixtures were transferred into a fresh Eppendorf tube and stored at -80°C until analysis.

Peptide separation and MS/MS measurement have been described previously (3). A volume of $2 \mu\text{L}$ of each bone digest was loaded and desalted on a C18 PepMap precolumn ($300 \mu\text{m}$, 5 mm-i.d., $300 \text{ \AA}$; Thermo Scientific) and separated by reversed-phase liquid chromatography using two identical ChromXP C18CL columns (150 mm 0.3 mm -i.d. $120 \text{ \AA}$; Eksigent) coupled parallel and connected to a split less NanoLC-Ultra 2D plus system (Eksigent) with a linear 45-minute gradient from 4 to 35% acetonitrile in 0.05% formic acid and a constant ($4 \mu\text{L}/\text{minute}$) flow rate. The LC system was coupled to an amaZon speed ETD ion trap (Bruker Daltonics) equipped with an Apollo II electrospray ionization (ESI) source. After each MS scan, up to 10 abundant multiply charged species in the range m/z 300-1300 were selected for MS/MS using collisional induced dissociation (CID) and actively excluded for 1 minute after having been selected twice. Each individual scan or tandem mass spectrum was saved to the hard drive. The LC system was controlled by HyStar 3.2 and the ion trap by trapControl 7.1. Two technical replicates were acquired.

Similarly, two isolates of three strains of *Mycobacterium tuberculosis* (EAI and modern and ancient Beijing) were analysed in duplicate by LC/MS/MS using an LTQ-FT Ultra mass spectrometer (Thermo Scientific), at the LUMC. The analysis was performed using on-line, reverse phase nano-HPLC with a model 1100 gradient HPLC system (Agilent, Germany). A volume of $2 \mu\text{L}$ of sample was injected onto a home-made pre-column ($100 \mu\text{m}$ i.d. $\times$ 15 mm) packed with Reprosil-Pur C18-AQ ($3 \mu\text{m}$; Dr Maisch, Germany) and eluted via an similarly home-made analytical nano-HPLC column ($150 \text{ mm} \times 50 \mu\text{m}$ i.d.). The gradient was run from 0% to 30% solvent B (10/90/0.1 water/acetonitrile/formic acid) in 10-155 minutes. A tip of $5 \mu\text{m}$ was drawn to the end of the nano-HPLC column that acted as electrospray needle. Full scan mass spectra were acquired in the fourier transform ion cyclotron resonance (FTICR) with a resolution of 25,000 at a target value of 5×10^6 . The five most intense ions were selected and fragmented in the linear ion trap using CID at a target value of 10,000.

Comparison of protein extraction protocols

To test if compareMS2 can be used to investigate which parameters have the largest influence on protein extraction, we applied 24 different protocols for protein extraction

to 24 highly similar ~20 mg samples of muscle tissue from rainbow trout (*Oncorhynchus mykiss*). In the protocols, temperature, time, bead material and size, detergent, buffers and solvents were varied systematically (described in Figure 4). All extracts were analyzed by LC/MS/MS using the same method and equipment as for the bone samples described above.

Quality control/stability of troponin standards

Human cardiac troponin complex (consisting of three subunits troponin T, troponin I and troponin C) was prepared by HyTest Ltd. (Turku, Finland) from human heart and obtained from the National Institute of Standards and Technology (NIST) as Standard Reference Material 2921 (SRM2921) in a buffer composed of 150 mM sodium chloride, 5 mM calcium chloride and 20 mM TRIS (pH 7.5). The reference material was shipped on dry ice in vials with approximately 115 μ L certified concentrations of 31.2 ± 1.4 mg/L for the hcTnI, 36.9 ± 3.8 mg/L for the hcTnT and 24.2 ± 1.3 mg/L for the hcTnC and stored at -80°C after arrival. Calbiochem 648480 purified hcTnI was purified from human heart and purchased from Merck (Darmstadt, Germany) as 100 μ g lyophilized powder. After arrival, the subunit TnI was dissolved in an identical buffer used for the SRM2921 to a final concentration of 100 μ g/mL, separated into 10- μ L aliquots and stored in a -80°C freezer. Previously, the temperature-dependent stability of these two reference standards was evaluated via detailed characterisation of specific peptides obtained from cleavage sites (12). In the current study stability and degradation of NIST SRM2921 and the Calbiochem hcTnI were investigated in several ways. First, aliquots of the NIST and Calbiochem standards were incubated for 48 h at 21°C (room temperature) or 48 h at 37°C . Other aliquots were allowed to degrade for 7 days at the same temperatures. To simulate oxidation, both standards were treated with 2% hydrogen peroxide in 50 mM ammonium bicarbonate (ABC) for 5 minutes.

To reduce disulfide bonds, both NIST SRM2921 and Calbiochem hcTnI were incubated for 30 minutes at 50°C with 10 μ L 50mM ABC and 1.5 μ L 100 mM DTT. After reduction, 3 μ L 100 mM iodoacetamide was added to each sample and the samples were kept in dark for 30 minutes to alkylate the cysteines. To simulate hydrolysis or proteolysis, six aliquots of each troponin standard were partially digested by chymotrypsin using 0.01 μ g enzyme (Promega) per aliquot and allowing the digestion to proceed for 2 hours at 25°C . All samples except three of each standard partially digested with chymotrypsin were digested by trypsin (Promega) by adding 0.1 μ g trypsin to each aliquot followed by overnight digestion at 37°C . All experiments (conditions) were performed in triplicate, except the untreated control which were analysed fivefold.

compareMS2 and MEGA

All LC/MS/MS datasets were analyzed using compareMS2 as previously described (1). In brief, Bruker DataAnalysis is used to extract 2,000 tandem mass spectra exactly selected on the highest signal intensities, and exported into MGF files, not making use of the retention time constraints. For the troponin experiment, only 200 spectra were used. The precursor tolerance is a variable parameter which for this work was set at 1.7 m/z units to cover the "isotope error" and the mass measurement error in the used instruments. The

compareMS2 output, or fraction of shared tandem mass spectra between each pair of datasets, was combined into a distance matrix in the MEGA format and clustered using a UPGMA tree generated by MEGA 5.05 (13) with default settings. No parameters were changed or tuned between analyses as the purpose was to show applicability rather than optimality. As an alternative to generating phylogenetic trees or dendrograms in MEGA, we can read the lower diagonal distance matrix from MEGA file into R, make it symmetric, and use the heatmap() R function to visualize the distance matrix, using UPGMA as above.

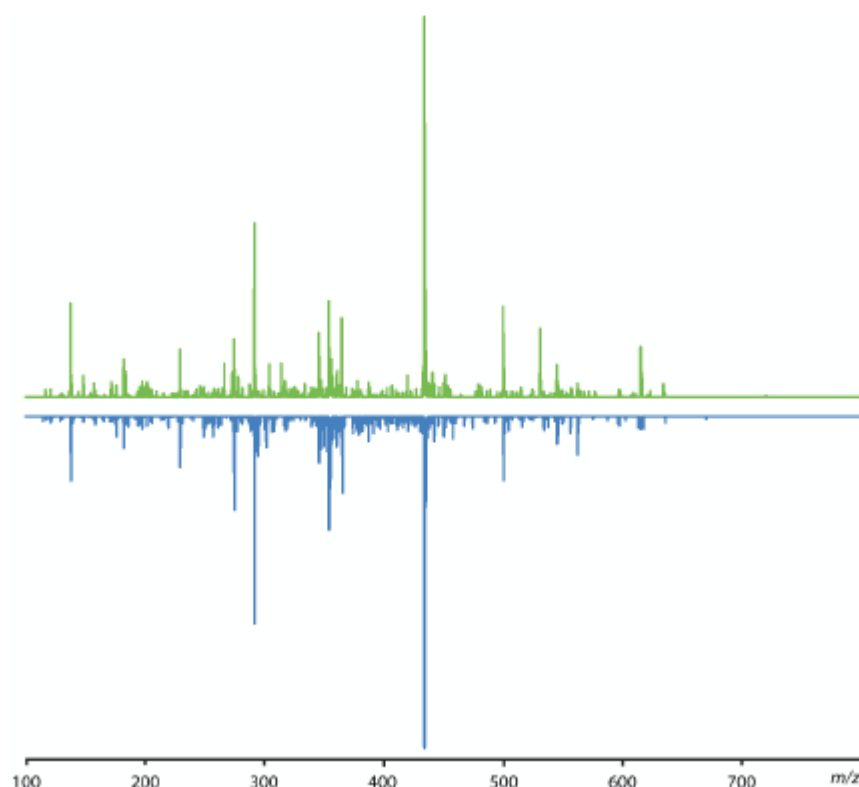


Figure 1. Two shared amaZon ion trap tandem mass spectra from *Mycobacterium szulgai* (top) and *Mycobacterium avium* (bottom, inverted). The peptide is not identified in the analysis, but PeptideProphet was applied on the dot products calculated by SpectraST for these two spectra, after placing one spectrum in a SpectraST spectral library and using the other as the query. The confidence in the peptide assignment in the library was set to 1 (certain), as we are here interested only in the probability of the two spectra deriving from the same peptide (a non-random event), not the actual identity of the peptide. In this way PeptideProphet can still calculate a probability for these two spectra deriving from the same peptide (or strictly speaking, being a non-random match). For these two spectra, that probability is >0.99995 (represented as 1.0000 by PeptideProphet). Such a match represents one “count” and the smallest unit of information used by compareMS2.

RESULTS AND DISCUSSION

All data presented here is based on spectral counting, separating shared and unique (or non-shared) spectra. As it happens, most of these spectra correspond to peptides, or more precisely, are dominated by fragments from a single peptide. However, this is never assumed in the compareMS2 analysis. Although we compare thousands of tandem mass spectra, each count is a pair of similar, shared spectra, illustrated by those in figure 1. By similar and shared we here mean a compareMS2 dot (scalar) product of the two spectra above 0.8. This is similar to the scoring in spectral library searching, e.g. using SpectraST (14, 15), except we compare two unknowns instead of searching a consensus library of confidently identified spectra. As for SpectraST, PeptideProphet (16, 17) can estimate the probability that two spectra derive from the same peptide, or having a non-random degree of similarity. This is a separate probability from that of the match of the spectrum to the amino acid sequence stored in the SpectraST libraries. In a spectral library search, the confidence of the assignment of a query spectrum to a peptide sequence is the product of these two probabilities. In this work, we only care about the probability that two spectra derive from the same peptide. Even though compareMS2 in principle would work equally well on chimeric spectra, its reproducibly requires extremely stable chromatographic conditions. However, most spectra are dominated by one peptide species, and two chimeric spectra with the same dominating peptide would still be considered similar by compareMS2.

Molecular phylogenetics

Towards the end of our original publication on compareMS2 (1), we suggested that “the applicability of the method to other types of samples, such as animal hair, skin, fresh or fossilized bone, remains to be investigated”. In analysis of old (centuries or more) bone material in other studies, highly modified (hydroxyproline) collagen peptides were primarily identified (18, 19). This is obviously a challenge for compareMS2, although a 200-year old pig bone was still identifiable as porcine by compareMS2 (data not shown). Collagen dominated also in the extracts of fresh bone (10% of the spectra from pig, counting the *unmodified* peptides only). The other proteins on the “top ten” list for fresh bone in these extracts were biglycan (7.0%), serum albumin (5.8%), vimentin (3.4%), fibromodulin (2.6%), aggrecan core protein (2.4%), fibrinogen (2.3%), lamin-A/C (1.8%), actin (1.8%) and decorin (1.7%). However, and it should be stressed, no proteins are actually identified by compareMS2. These proteins were identified in a search against a porcine sequence database using X!Tandem and PeptideProphet, as before (3), to characterize the input data. Figure 2 (A) shows the overlay of two replicate phylogenetic trees for the fallow deer, moose, cow, pig and chicken. This tree agrees with what is known to be the correct evolutionary relationship between these organisms (20–24). The very minor (barely noticeable) difference between the two replicates suggest some measure of robustness of the method, at least for data of the quality generated for these samples and on the same instrumentation. The phylogenetic tree for the three *Mycobacterium tuberculosis* strains (figure 2 B) is also correct (25), albeit less reproducible than the tree generated from bone material. This is most likely due to the use of a less robust chromatography and electrospray method (lower flow rate, nanoelectrospray).

We know from all other work that the setup with the Eksigent chromatography systems, Bruker standard ESI source and amaZon ion traps yields highly reproducible spectra (1, 3, 7-10). Unknown bacterial isolates can be putatively identified by comparing their distance in the compareMS2 metric based on shared spectral content from a number of reference datasets from well-characterized isolates of known species or lineage (10). We are aware of instrumentation like Biotyper (Bruker) which can easily analyze unknown colonies, however the used methods here (bottom-up, 45-minute gradient, cap-LC) generate considerably more information especially when combined with the compareMS2 metric.

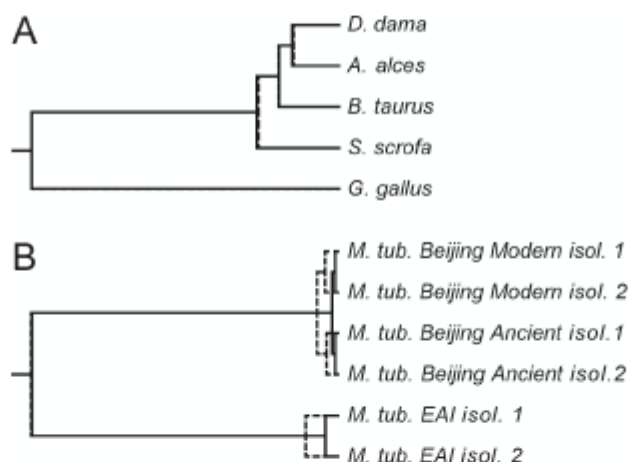


Figure 2. Molecular phylogenetics from bone (A) and mycobacterium colonies (B), superimposing two phylogenetic trees (dashed and solid lines) generated from the two sets of technical replicates.

Figure 3 shows the distributions of the dot product calculated by compareMS2 for all included tandem mass spectra for the bone and *Mycobacterium tuberculosis* phylogenetic experiments. Although the samples were different, it is clear that if the more accurate precursor mass information from the LTQ-FT is used, the majority of the matched spectra are very similar. This effect is smaller but still present for organisms of larger genomes, encoding more proteins with more peptides of very similar mass. The precursor mass information can be omitted in the compareMS2 analysis, in which case the distributions look more similar (the LTQ and amaZon ion traps have similar mass measurement accuracy and resolving power).

For full disclosure, we also applied the compareMS2 method on a number of very similar *Clostridium difficile* isolates (26), with limited success (Supplemental Figure 1). The results suggested that the correct isolate could be identified in perhaps 60% of the cases (data not shown). Further work is needed for this pathogen. The reason for why the method performs so well in some cases (*M. tuberculosis*, *S. aureus*) and less well in others (*C. difficile*) may also have something to do with how the clinically relevant isolates are defined and how different they actually are under standard culture conditions. But there may be several other reasons for why these proteomics analyses “failed” to form

a similar phylogenetic tree as the PCR-based ribotyping, for example, Knetch discusses that plausibly many studies fail to distinguish PCR RT020 from PCR RT014, due to the difference in typing results (26).

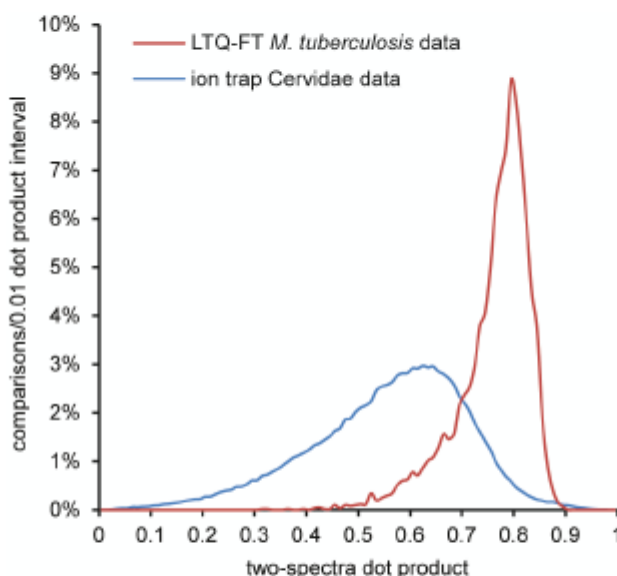


Figure 3. Two-spectra dot product distributions for bone spectra acquired on the Bruker ion trap (blue) and mycobacterium spectra on the LTQ-FT (red).

Method development and quality control

At the time of the experiment, rainbow trout (*Oncorhynchus mykiss*) was not fully sequenced and only few sequences were available in UniProt. Now (August 2015), a reference proteome with 49,986 sequences is available in UniProt). However for the majority of species, there is of course still very little sequence data available. In the case of rainbow trout, even a search against the close relative salmon (*Salmo salar*), will fail to match for a number of high-quality spectra. compareMS2 can use all spectra in the analysis, assessing the quality and information content of the datasets from any species, guiding the further analyses, for instance using methods such as QualScore (27) and *de novo* sequencing for identifying unidentified high-quality spectra. This experiment illustrates how compareMS2 can be applied in any situation when DNA- or RNA-Seq data is unavailable, including of course while such data is being acquired, and one would like quick feedback to the quality of the proteomics data to optimize experimental protocols.

Figure 4 shows a hierarchal clustering of the rainbow trout muscle tissue extracts generated by R and comparing a number of different protocols. Technically, this is not molecular phylogenetics, but we re-appropriate the method for hierarchal clustering and comparison of these datasets. The horizontal axes, or distances, in these trees have been omitted here. The numbers from the UPGMA analysis are not very meaningful in this context and vary from experiment to experiment. The numbers are also influenced by the quality of the data, particularly the contribution of noise. But it is the topology that

matters. In the phylogenetic analyses, the numbers give the evolutionary relationship between species and lineages, enabling the identification of unknown samples. Here, the tree topology allow us to rank the relative importance of experimental factors such as time, temperature or which beads were used for homogenization of the tissue, all valuable information when designing proteomics experiments.

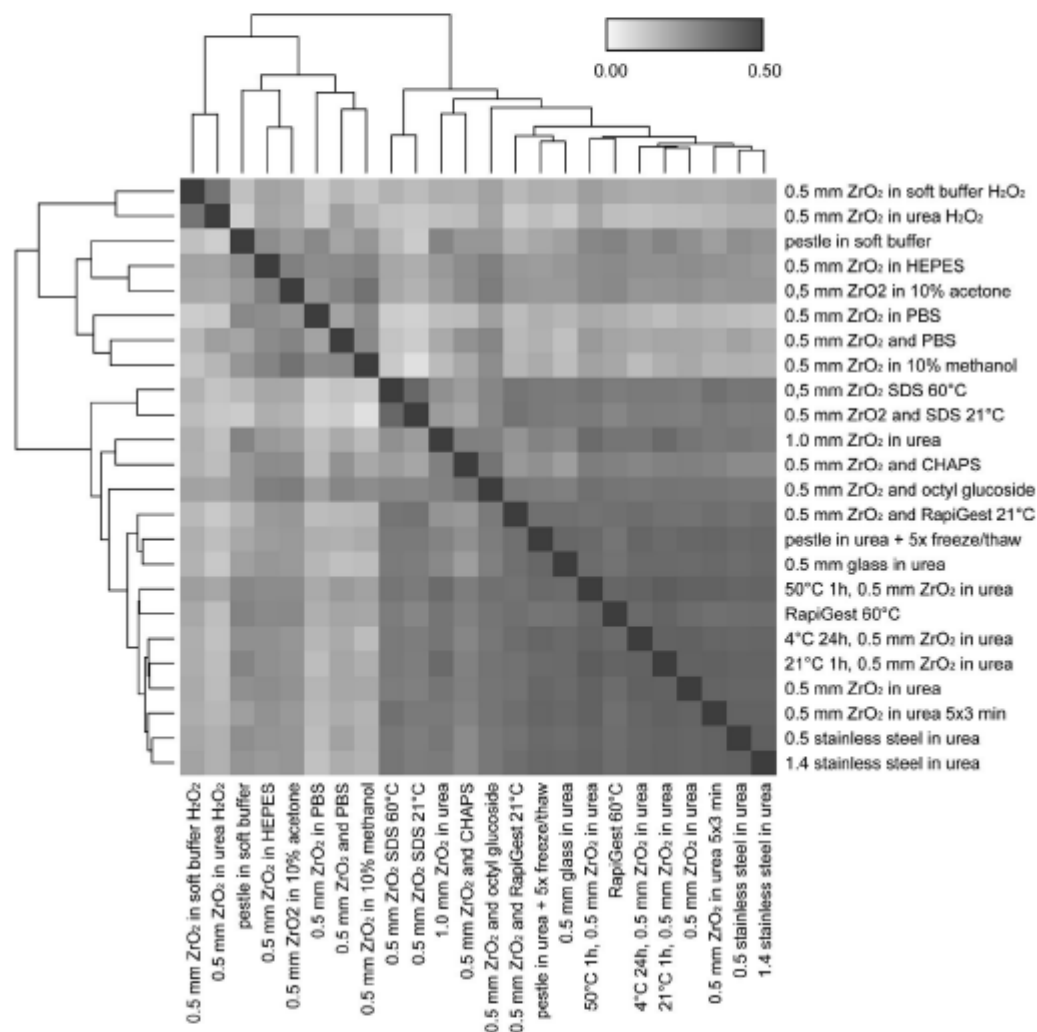


Figure 4. Heatmap comparing extraction protocols for rainbow trout muscle tissue. This analysis can be performed on samples of any organism, independent on availability of sequence information, and may be used to guide further experiments and analyses, for example while waiting for access to sequences in concurrent next-generation sequencing and proteomics proteogenomic studies. Standard extraction is in 100 µL buffer with 3 minutes in bullet blender on level 8 using 0.5 mm zirconium oxide beads, followed by 27 minutes storage (at 4°C unless otherwise indicated) and centrifugation 30 minutes at 16,000 x g. The "soft" extraction buffer is 20 mM Tris-HCl, 20 mM NaCl, 40 mM MgCl, 50 U/ml Benzonase and protease inhibitor.

Similarly for the quality control of the Calbiochem and NIST troponin standards, their stability can be visualized for various storage conditions in one tree generated by exactly the same compareMS2/R analysis used in the phylogenetic studies. The resulting tree is shown in figure 5. There is a clear separation between the two standards independent of storage conditions, as long as these are kept within reason. This is obviously due to the protein content actually being different, with the NIST standard containing the entire complex and the Calbiochem only one of the three subunits. Incubation with hydrogen peroxide oxidizes most peptides, and the two standards treated in this way show high similarity. An even larger difference results from using another enzyme than trypsin, simulating protease activity. The analysis suggests that the three subunit NIST standard is more sensitive to temperature degradation than the single subunit Calbiochem standard, as the samples stored at 37°C cluster apart from the samples stored at 21°C. This is in excellent agreement with our previous temperature-dependent stability study on the same materials (12).

Oxidation with hydrogen peroxide changes the masses of several amino acid side chains, predominantly methionine and cysteine, which means that the masses of any peptide containing one or more of these residues also change. Since we are using a 1.7 m/z -unit window for the spectral comparisons, the spectra of the native and oxidized version of the same peptide will not be compared (even if they may have a high dot-product if some fragment ions overlap). Since many peptides are mutated by this extreme treatment, it is not surprising that these samples stand out. In practice, one may use the overlap with, or similarity to oxidized references such as the ones generated here to detect and quantify the degree of oxidation of a standard, without prior knowledge of which peptides and residues are most susceptible to oxidation.

Spectral matching against a reference library was used as a data quality metric already in 2010 by Rudnick *et al.* (28). Analogously, the similarity with a reference dataset (or a spectral library) built from the unadulterated standards is a measure of *sample* quality. However, comparing not only with the control, but also reference data from deliberately maltreated (oxidized, degraded, contaminated) samples may also indicate how a reference material or sample has been handled. Again, this analysis is performed prior to any peptide identification and can take all data into account without knowing what to look for in a database search. The same data can be further analyzed by other methods, incorporating more information. The software we have developed, compareMS2, was *not tweaked in any way* to suit these very different applications and data from different instrumentation. This suggests it was almost certainly applied in a suboptimal way. However, our main objective here was to demonstrate the general utility and inherent robustness of the method, a goal which is best served by applying it without adjustment or modification. The ability to obtain meaningful and interpretable results in such different situations points to a general applicability and simplicity of the concept of measuring similarities and differences between samples by their shared tandem mass spectra.

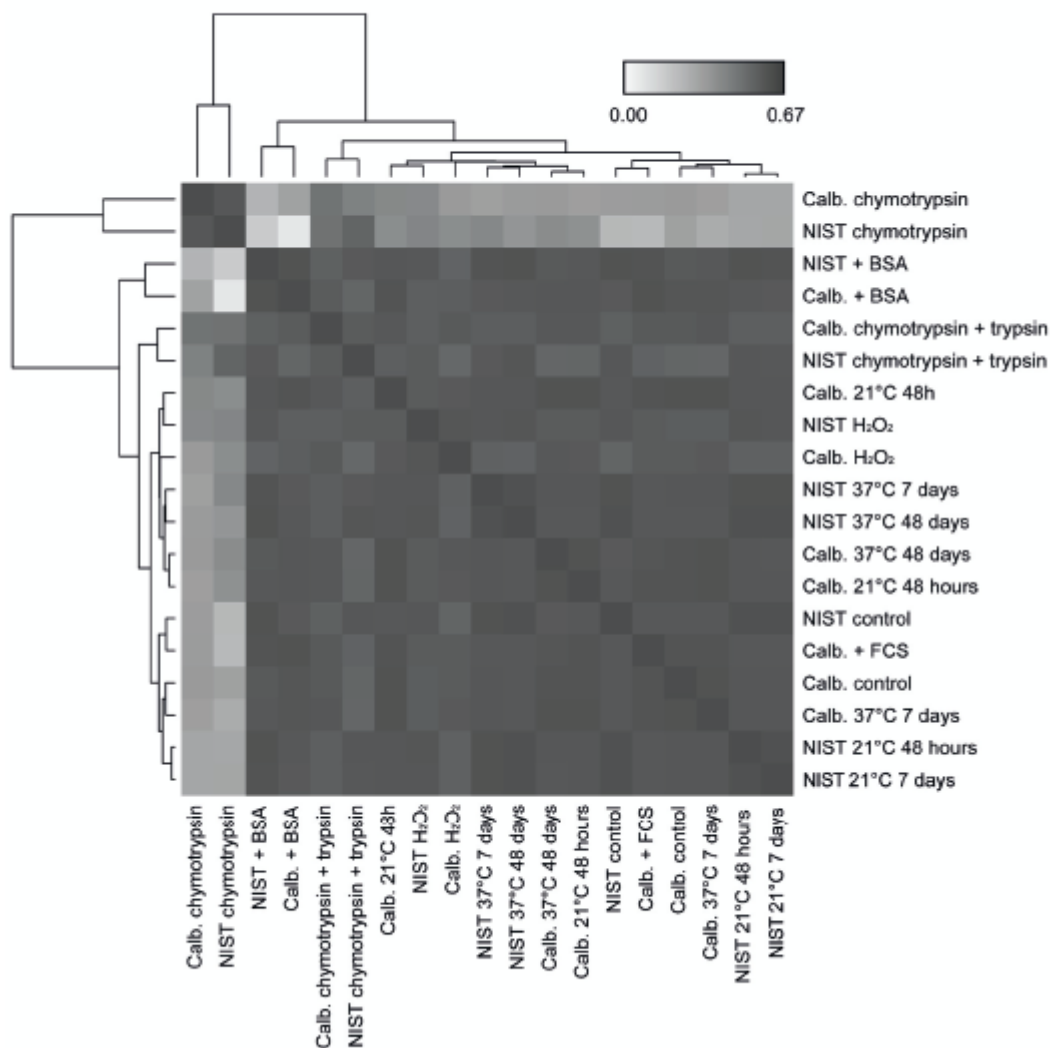


Figure 5. Quality control and simultaneous investigation of several parameters potentially affecting the quality of troponin standards currently used in clinical proteomics. Knowing the range of allowable storage and handling conditions for such standards is important in routine clinical proteomics, as absolute quantitation will only be, at best, as good as the reference.

CONCLUSIONS

The compareMS2 software is open source and can be downloaded from www.ms-utils.org/compareMS2.html. Although we do not suggest a direct and uncritical application of the methods explored here, or the compareMS2 software, we argue that the diverse set of examples shown here does suggest a very broad applicability of this and any similar tool that compares tandem mass spectra *en masse* between LC/MS/MS datasets or groups of such sets. compareMS2 can rank importance of arbitrary experimental factors and be applied in analyses where very little is known about the samples (or biological

species). Although the analysis may seem complicated at a first glance, the method is actually extremely simple and uses only a minimum of information: the tandem mass spectra themselves. No knowledge on amino acids (or chemistry of any kind), sample composition or analysis method (e.g. chromatography or fragmentation technique) is used by compareMS2. This does *not* exclude the use of such information. On the contrary, the addition of such information will provide an orthogonal component to the results of compareMS2. Due to the minimal assumptions made, compareMS2 could be expected to “play well” with other tools. We suggest that compareMS2 can be used to quickly compare a large number (hundreds) of datasets to find possible outliers or low-quality datasets in large-scale proteomics studies before feeding them to a proteomics data analysis pipeline together with the other datasets. But more generally, we feel that only our imagination limits the application of generic tools such as that exemplified by compareMS2.

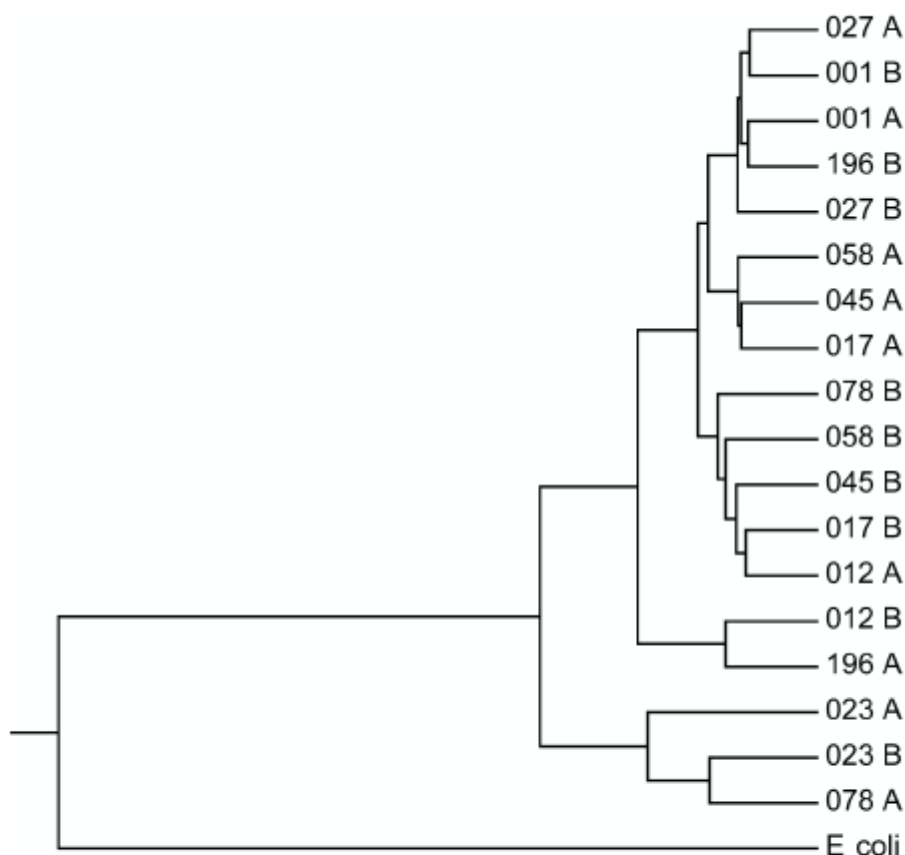
ACKNOWLEDGEMENTS

MP, SJPD and DO are supported by the NWO VIDI grant 917.11.398

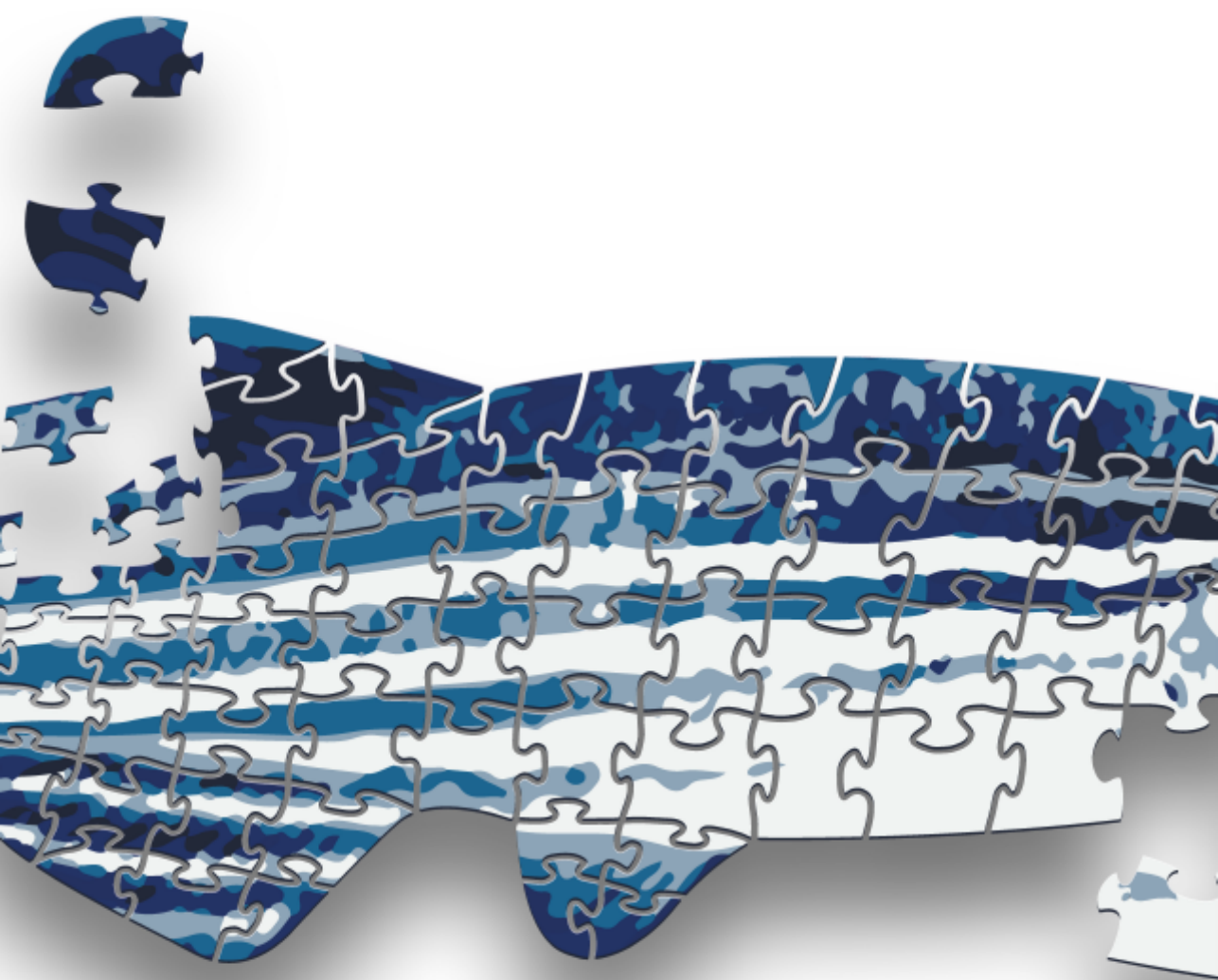
REFERENCES

1. Palmblad, M. & Deelder, A.M. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun. Mass Spectrom.* 26, 728-732 (2012).
2. Zuckerkandl, E., Jones, R.T. & Pauling, L. A Comparison of Animal Hemoglobins by Tryptic Peptide Pattern Analysis. *Proc Natl Acad Sci U S A* 46, 1349-1360 (1960).
3. van der Plas-Duivesteyn, S.J. et al. Identifying proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *Journal of proteome research* 13, 1537-1544 (2014).
4. NIST <http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:download>.
5. Onder, O., Shao, W., Kamps, B.D., Lam, H. & Brisson, D. Identifying sources of tick blood meals using unidentified tandem mass spectral libraries. *Nat Commun* 4, 1746 (2013).
6. Lu, L. et al. In American Society for Mass Spectrometry Annual Conference (ASMS, Minneapolis, MN; 2013).
7. Wulff, T., Nielsen, M.E., Deelder, A.M., Jessen, F. & Palmblad, M. Authentication of fish products by large-scale comparison of tandem mass spectra. *Journal of proteome research* 12, 5253-5259 (2013).
8. Palmblad, M. In 63rd American Society for Mass Spectrometry Annual Conference (ASMS, St. Louis, MO; 2015).
9. Nessen, M. In 63rd American Society for Mass Spectrometry Annual Conference (ASMS, St. Louis, MO; 2015).
10. Ohana, D. In 63rd American Society for Mass Spectrometry Annual Meeting (ASMS, St. Louis, MO; 2015).
11. Jiang, X. et al. Method development of efficient protein extraction in bone tissue for proteome analysis. *J Proteome Res* 6, 2287-2294 (2007).
12. van der Burgt, Y.E., Cobbaert, C.M., Dalebout, H., Smit, N. & Deelder, A.M. Temperature-dependent instability of the cTnI subunit in NIST SRM2921 characterized by tryptic peptide mapping. *Journal of chromatography. B*,

- Analytical technologies in the biomedical and life sciences 902, 147-150 (2012).
13. Tamura, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28, 2731-2739 (2011).
 14. Lam, H. & Aebersold, R. Spectral library searching for peptide identification via tandem MS. *Methods Mol Biol* 604, 95-103 (2010).
 15. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7, 655-667 (2007).
 16. Keller, A., Eng, J., Zhang, N., Li, X.J. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* 1, 2005 0017 (2005).
 17. 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.* 74, 5383-5392 (2002).
 18. Marom, A., McCullagh, J.S., Higham, T.F., Sinitsyn, A.A. & Hedges, R.E. Single amino acid radiocarbon dating of Upper Paleolithic modern humans. *Proc Natl Acad Sci U S A* 109, 6878-6881 (2012).
 19. Schweitzer, M.H., Schroeter, E.R. & Goshe, M.B. Protein molecular data from ancient (>1 million years old) fossil material: pitfalls, possibilities and grand challenges. *Anal Chem* 86, 6731-6740 (2014).
 20. Zhou, H., Chen, S., Wang, M. & Cheng, A. Interferons and their receptors in birds: a comparison of gene structure, phylogenetic analysis, and cross modulation. *International journal of molecular sciences* 15, 21045-21068 (2014).
 21. Finarelli, J.A. Estimating endocranial volume from the outside of the skull in Artiodactyla. *Journal of Mammalogy* 92, 200-212 (2011).
 22. Pitra, C., Fickel, J., Meijaard, E. & Groves, P.C. Evolution and phylogeny of old world deer. *Molecular phylogenetics and evolution* 33, 880-895 (2004).
 23. Dekel, Y. et al. Dispersal of an ancient retroposon in the TP53 promoter of Bovidae: phylogeny, novel mechanisms, and potential implications for cow milk persistency. *BMC Genomics* 16, 53 (2015).
 24. Immel, A. et al. Mitochondrial Genomes of Giant Deers Suggest their Late Survival in Central Europe. *Scientific Reports* 5, 10853 (2015).
 25. Casali, N. et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 46, 279-286 (2014).
 26. Knetsch, C.W. et al. Comparative analysis of an expanded *Clostridium difficile* reference strain collection reveals genetic diversity and evolution through six lineages. *Infection, Genetics and Evolution* 12, 1577-1585 (2012).
 27. Nesvizhskii, A.I. et al. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics* 5, 652-670 (2006).
 28. Rudnick, P.A. et al. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol Cell Proteomics* 9, 225-241 (2010).



Supplemental figure 1. Dendrogram generated by compareMS2 for *Clostridium difficile* ribotypes. Although the *Clostridium* strains are easily distinguished from *E. coli*, replicates of the same ribotype (A and B) did not always cluster, suggesting the proteomic differences measured in this experiment were too small to be resolved given the experimental variation and data quality, although ribotype 023 (lineage 3) does split from the other samples and both replicates of the hypervirulent ribotype 027 (lineage 1) are found in the same branch. The *C. diff.* strains and ribotypes are described by Knetsch *et al* (26).



Chapter 4

A Full-Body Transcriptome and Proteome Resource for the European Common Carp

I.C.R.M. Kolder, S.J. van der Plas-Duivesteijn, G. Tan, G.F. Wiegertjes, M. Forlenza, A.T. Guler, D.Y. Travin, M. Nakao, T. Moritomo, I. Irnazarow, J.T. den Dunnen, S.Y. Anvar, H.J. Jansen, R.P. Dirks, N.M. Palmblad, B. Lenhard, C.V. Henkel, H.P. Spaink

BMC genomics 17, 2016, 701

ABSTRACT

The common carp (*Cyprinus carpio*) is the oldest, most domesticated and one of the most cultured fish species for food consumption. Besides its economic importance, the common carp is also highly suitable for comparative physiological and disease studies in combination with the animal model zebrafish (*Danio rerio*). They are genetically closely related but offer complementary benefits for fundamental research, with the large body mass of common carp presenting possibilities for obtaining sufficient cell material for advanced transcriptome and proteome studies.

Here we have used 19 different tissues from an F1 hybrid strain of the common carp to perform transcriptome analyses using RNA-Seq. For a subset of the tissues we also have performed deep proteomic studies. As a reference, we updated the European common carp genome assembly using low coverage Pacific Biosciences sequencing to permit high-quality gene annotation. These annotated gene lists were linked to zebrafish homologs, enabling direct comparisons with published datasets. Using clustering, we have identified sets of genes that are potential selective markers for various types of tissues. In addition, we provide a script for schematic anatomical viewer for visualizing organ-specific expression data.

The identified transcriptome and proteome data for carp tissues represent a useful resource for further translational studies of tissue-specific markers for this economically-important fish species that can lead to new markers for organ development. The similarity to zebrafish expression patterns confirms the value of common carp as a resource for studying tissue-specific expression in cyprinid fish. The availability of the annotated gene set of common carp will enable further research with both applied and fundamental research purposes.

INTRODUCTION

The common carp (*Cyprinus carpio*) is the oldest, most domesticated and one of the most cultured fish species for food consumption, angling purposes and expensive species as ornamental fish. Geographical and reproductive isolation resulted in the formation of two subspecies; European common carp (*Cyprinus carpio carpio*) in the West and East-Asian common carp (*Cyprinus carpio hematopterus*) in the East (1). It is closely related to zebrafish (*Danio rerio*), a commonly used animal model to study human disease (2-4), with both lineages diverging approximately 11-21 million years ago (5, 6). Because of its large body size, a single animal can yield sufficient amounts of tissue from its organs and blood for extensive genomic, proteomic and metabolic studies without compromising to possible contaminations with neighboring tissues. Especially for small organs the carp also offers far better possibilities than zebrafish for obtaining tissues that are well-separated from other tissues. Furthermore, carp itself, besides being an important edible species and offering complementary benefits to the closely-related zebrafish, is emerging as a highly useful animal model on its own, providing valuable information for questions of physiology, genetics, immunology and disease (1). For instance, common

immune genes that were notably lacking in zebrafish were shown to be present in the carp genome (7).

Recently, several full genome assemblies of carp species have become available (8, 9). These studies confirm the duplicate nature of the carp genome with respect to zebrafish and several other teleosts, with the carp lineage having experienced a recent (8 Mya) allotetraploidization event (10, 11). The carp genome consists of 50 chromosome pairs, in contrast to 25 for zebrafish (9). Between carp and zebrafish, extensive conservation of synteny remains (8, 9). Curiously, the carp genome is of approximately equal size to the zebrafish genome, the result of less intra- and intergenic repetitive content (8). Further comparative genomics and transcriptomics studies using Hungarian, North American and Chinese carp strains have provided insights into the species structure, and identified genes associated with skin color and scale phenotypes (9). However, further analyses of carp transcriptomes are not yet abundant, with an emphasis on embryonic samples and mixed tissues (8, 9).

A major drawback of the genomic information on carp is the lack of a genome-based annotation of all gene predictions in public databases. As a result, the accessibility of carp data has been limited to specialist analyses. Here, we describe the generation of an annotated gene set of *C. carpio* based on published and new genomic data, generated by the sequencing of long DNA fragments using Pacific Biosciences technology. We use this annotated set to investigate tissue-specific gene expression, based on RNA-Seq data generated for dozens of distinct tissue samples. In addition, we have performed deep proteomic analyses of several tissues for comparison with the transcriptome data.

In order to assess the common carp as more than only an important edible species, we aimed to validate carp as an animal model complementary to the zebrafish used for human disease studies, and investigate the relationship between these two closely-related cyprinid species by comparing the expression of tissue-specific genes in carp and predicted tissue-specific expression in zebrafish. Together, access to a well-annotated genome and the whole animal approach to tissue gene expression provide new resources which will substantiate research on common carp.

MATERIALS AND METHODS

Carp genome

Carp fish line

R3 x R8 are the hybrid offspring of a cross between carp of Hungarian origin (R8 strain) and of Polish origin (R3 strain) (12), each of which are kept purebred by single brother-sister matings (13-15). A cross between single R3 and R8 individuals purebred for 5 generations led to a base population for artificial reproduction by induced gynogenesis and subsequent production of a homozygous all-female clonal line, which was sampled for genomic DNA to construct genomic libraries (12). A cross between single R3 and R8 individuals purebred for 11 generations led to a base population for sampling of organs

for RNA sequencing. Genomic DNA was used for construction of genomic libraries as described (8). The breeding of adult fish was approved by the local animal welfare committee (DEC) of Wageningen University. All protocols adhered to the international guidelines specified by the EU Animal Protection Directive 2010/63/EU

CLIP-PE sequencing

CLIP-PE Libraries were made mostly according to the protocol published by Peng *et al.* (16). DNA was sheared using a Covaris g-TUBE by centrifuging for 2×1 minute at 8000 rpm in an Eppendorf 5414R centrifuge. The DNA was subsequently run on a 0.6% agarose gel in TAE buffer to isolate the desired range of fragments. The Qiagen QIAexII kit was used to isolate the DNA from the gel. To repair the DNA damage a PreCR incubation was done according to the manufacturers description (NEB, Ipswich MA, USA). After Ampure XP purification the DNA fragments were end repaired and A-tailed using the NEBNext DNA library prep reagent set (NEB). Adapters containing loxP sites were ligated to the DNA using the Quick ligase kit (NEB). Adapter ligated fragments were then circularized using Cre recombinase (Life technologies). After circularization the linear DNA fragments were digested using Plasmid safe ATP-dependent DNase (Epicentre). Circularized DNA fragments were digested using *NlaIII* (NEB). After digestion the reaction volume was increased to promote self-ligation and DNA fragments were again circularized using T4 ligase. Linear DNA fragments were removed using Plasmid safe ATP-dependent DNase. A PCR with Illumina PE PCR primers was performed to amplify the fragments and to add Illumina flow cell compatible ends to the DNA fragments. The libraries were paired end sequenced using Illumina HiSeq 2500 technology with a read length of 51 nt.

Long read sequencing

Genomic DNA for Pacific Biosciences sequencing was isolated from nucleated red blood cells, from a single adult female carp from a homozygous clonal common carp line (R3R8 69-45) described by Henkel *et al.* (8). DNA was isolated using the Qiagen Blood and Tissue DNeasy kit according to manufacturer's manual (Qiagen, Hilden, Germany).

The isolated DNA was fragmented with gTUBEs (Covaris) and end-repaired. SMRTbell DNA template libraries (insert size of 20Kb) were prepared according to the manufacturer's instruction. Sequencing of ten SMRT cells was performed on the Pacific Biosciences RSII platform using the Magbead loading protocol and P5-C3 chemistry. Sequencing reads had an N50 of 11.9 kbp. The total number of bases produced was 6.8 Gbp, which is the equivalent of $4.5 \times$ coverage of the genome.

Assembly

The genome was assembled by integrating the contigs generated previously (8) with the PacBio sequencing reads and the three CLIP-PE libraries with insert sizes of 5, 6.5, and 7.5 kbp. The paired-end data was used to construct scaffolds using SSPACE (17). Subsequently, PBjelly (18) was used to scaffold fragmented contigs and fill the gaps using long and single-molecule PacBio reads. Gaps that were covered by at least 5 PacBio reads were filled by the consensus sequence. Otherwise, Ns were used to fill gaps between contigs that were linked by less than 5 PacBio reads.

Annotation

As the first step of genome annotation, we applied repeat analysis with RepeatMasker (v4.0.5) (19) and the RMBlast engine (19). In addition to Repbase (v20140131) (20), we constructed a the *de novo* transposable element library for the carp genome from RepeatModeler (v1.0.8). The genome was then annotated using the MAKER pipeline (v2.31.8) (21). The construction of gene models was based on three sources of information: *ab initio* gene prediction, homologs-based evidence and expression evidence from a *de novo* transcriptome assembly. The *ab initio* gene prediction software AUGUSTUS (v3.1.0) (22), with trained gene model parameters, was used to predict genes from the repeat-masked genome. For homologs-based prediction, the zebrafish protein sequences from Ensembl version 77 were collected to build a protein database for BLASTX (v2.2.27) to align genome sequences to the corresponding protein sequences. To annotate the putative gene functions, we conducted a BLASTP (v2.2.27) search for the identified protein sequences against the Swiss-Prot database with the E-value of $1e-5$. The protein domain information and GO category classification were produced from InterProScan (v5.11-51.0) (23).

RNA-sequencing

Sample collection

From common carp with a comparable genetic background (after 11 generations of brother-sister matings) as the carp used for the carp genome (8), we used four healthy adult F1 hybrids. Carp specimens (1 male, 3 females) which were dissected into 89 samples from 19 different tissues (Additional file 1: Table S1). Additionally, two pools of embryos at 5 days post fertilization (dpf) were used. An overview of tissues collected is depicted in figure 1. Tissue samples were stored in RNAlater (Life Technologies, Carlsbad, USA), after which RNA was isolated using the Qiagen Blood and Tissue DNeasy kit according to manufacturer's manual (Qiagen, Hilden, Germany).

Illumina sequencing and processing

A library was made with the TruSeq Stranded total RNA library prep kit according to manufacturer's description (Illumina Inc, San Diego, USA). Both paired end and single libraries were sequenced using an Illumina HiSeq 2500 according to the manufacturer's description. Illumina software (HCS) was used for basecalling. Tophat version 2.0.5 (24) was used to align the reads to the reference genome. For each read pair, secondary alignments (which meet alignment criteria but are less likely to be correct) were filtered out using samtools version 0.1.18 (25). For each predicted gene, read counts were obtained from the alignment file using HTSeq-count version 0.5.3.p9 (26) using the 'intersection-strict' setting to ignore reads not aligning to annotated exons.

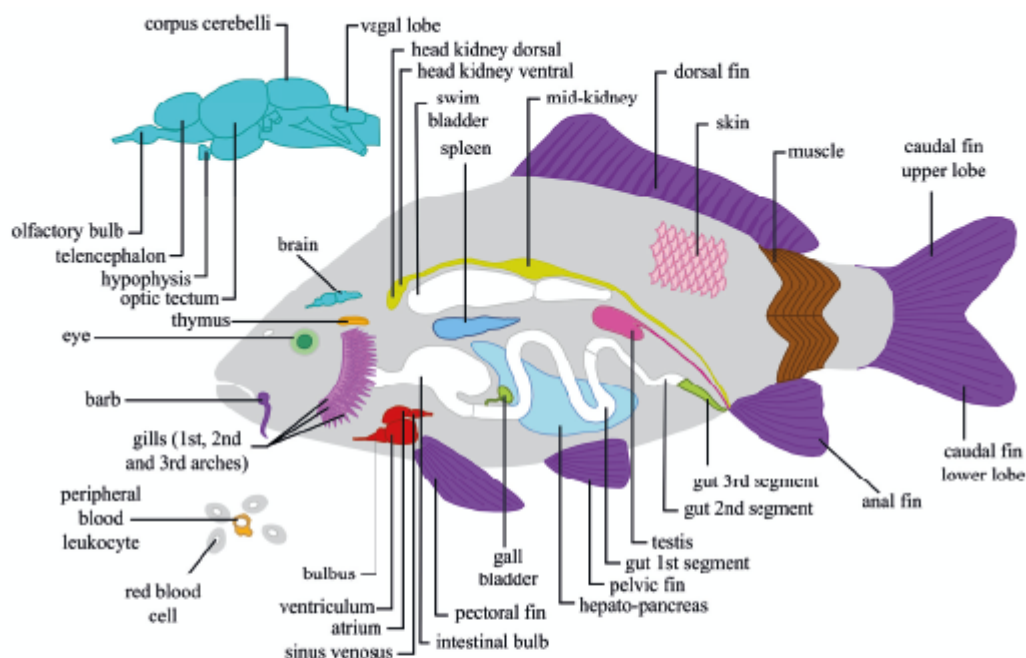


Figure 1. A cross-section of the carp organs sampled for this study. The different tissues were colored according to 16 groups to match the coloring of the following figures. A full list is available in Additional file 1: table S1.

Data analysis

Data quality was assessed using the statistical package R (27). Most importantly, raw RNA-Seq counts need to be normalized to correct for sequencing depth and (optionally) transcript length. Dividing counts by simple sequencing depth (the total number of reads) can lead to inaccurate results (28), as it is strongly influenced by the most abundantly expressed genes. For example, extreme expression of a single gene (e.g. hormone-encoding genes (29)) will strongly affect the expression measures of all other genes if naïve CPM (counts per million) or RPKM (reads per million per kilobase) normalization is used. Alternatively, empirical estimates of sequencing depth (available in the edgeR package, v. 3.12.0 (30)) are commonly used. However, like CPM and RPKM, these assume samples to be similar and perform poorly on our multi-tissue data (Additional file 2: figure S1A-B, log₁₀ scale). We therefore ranked the count data and normalized between samples by considering CPM- or RPKM-based expression based on the ranks instead of actual counts. Using this procedure, expression in all samples can be made comparable (Additional file 2: figure S1C-D, log₁₀ scale). A correlation matrix was calculated on the RPKM data with a Pearson correlation test (Additional file 3: figure S2). After removal of outliers a correlation matrix was calculated on the rank-normalized data of the remaining samples with a Spearman correlation test (figure 2). The correlation matrices were plotted using the corplot package (v0.73) (31). Tissue samples were clustered according to the correlation plots (Additional file 4: table S2). For each of these tissue groups the mean of the samples in that group were taken for

each gene. On these means, gene clusters were calculated using k-means clustering using the MultiExperiment Viewer (MeV version 4.9) (32). The genes in the resulting gene clusters were ranked on the mean of the overall expression. The top 10 expressed genes were plotted in a heatmap using the gplots(v2.17) (33). The software program Last (34) was used to align Chinese Songpu strain to our assembly at DNA level. Post-processing of Last results produces the actual whole-genome alignments. In this first step, original alignment blocks were chained together if their genomic locations in both genomes are close enough. Then netting process chose the best match in each region for the reference species. The post-processing was done by UCSC Kent utilities. InParanoid (35) was used to generate the orthologs mapping between carp and zebrafish.

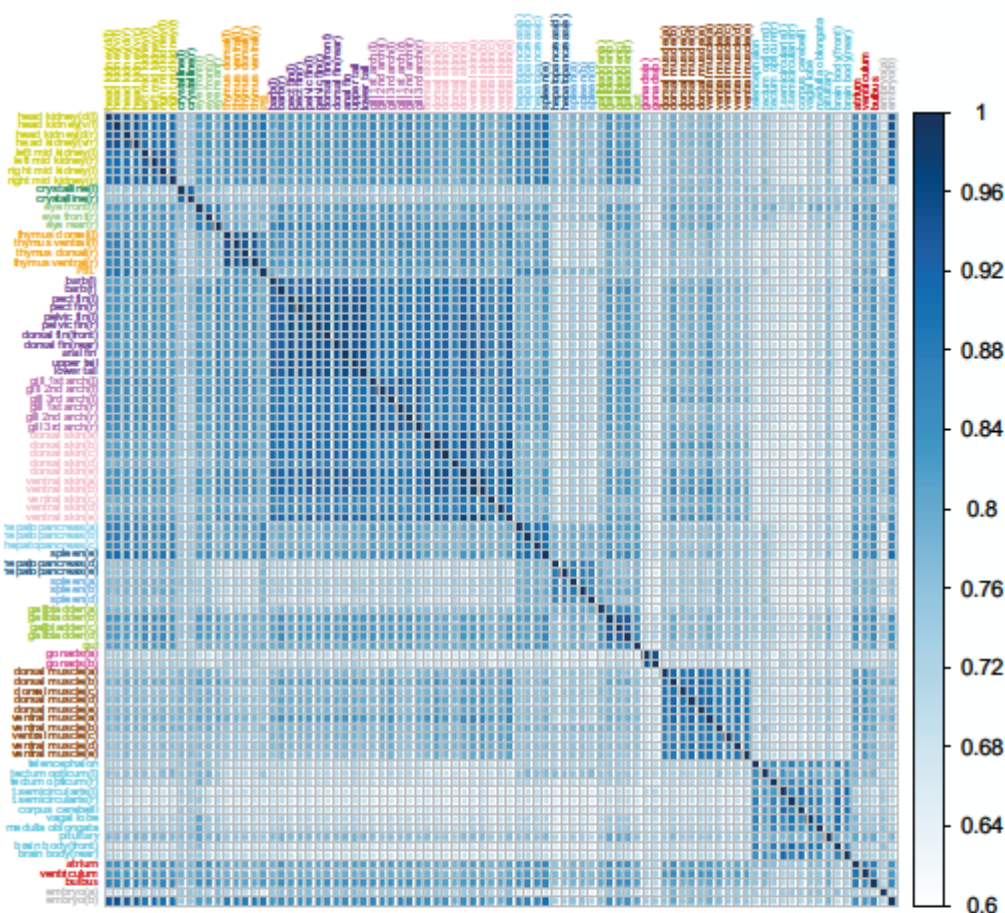


Figure 2. Correlation plot using the Spearman correlation test on the RPKM rank-normalized data (n=87) after mixed/outlier samples were removed. The different tissues were colored according to the 16 groups.

Proteomics

Protein Isolation

Proteins were extracted from heart, hepatopancreas and kidney tissue using 20 mM NaCl, 20 mM Tris-HCl pH 8.2, 50 U/mL Benzonase, 2 mM MgCl₂, and protease inhibitors (Roche), followed by homogenization with zirconium oxide (0.5 mm) beads using a Bullet Blender (Next Advance) at speed 8 for 3 minutes. The samples were placed at 95°C for 5 minutes, and centrifuged for 30 minutes at 4°C and 16,000 × *g*. The supernatant was transferred into a fresh tube and the pellet used to perform a second extraction by adding 1% SDS, 50 U/mL Benzonase, 2 mM MgCl₂, and protease inhibitors, followed by the same extraction procedure. Protein concentration of both extracts were determined using a BCA assay (#23235, Bio-Rad). Approximately 40 µg of both extracts was loaded on Novex 4-12% Bis-Tris gels (NuPAGE, Invitrogen) after adding 5 µL NuPAGE LDS Sample buffer (Invitrogen), and ran at 200 volt using MOPS SDS running buffer (Invitrogen) and an XCell Sure Lock Mini-Cell (Invitrogen) to fractionate proteins from each tissue solution. The gels were stained overnight in colloidal blue staining (LC6025 kit, Invitrogen) solution (55 mL deionized water, 20 mL methanol, 20 mL Invitrogen stain A, and 5 mL Invitrogen stain B) on a shaker at room temperature. Gels were washed twice in deionized water and scanned on an OptiGo UV Imager (Isogen Life Science). The gel lanes were cut in 48 identical pieces (Lane Picker, The Gel Company) and transferred to a 96-well plate using tweezers. Preparation for in gel digestion was performed at room temperature unless mentioned differently, and all washed with 100 µL solution (36).

LC/MS/MS

Peptide separation and MS/MS measurement were performed as described previously (37). In summary, to separate all peptides from each individual gel lane, ten µL of each sample was loaded and desalted on C18 PepMap 300 Å precolumn (Thermo Scientific) and separated by reversed-phase liquid chromatography using two identical 150 mm 0.3 mm-i.d. ChromXP C18CL, 120 Å columns (Eksigent) which were coupled parallel and connected to a split less NanoLC-Ultra 2D plus system (Eksigent) with a linear 90-minute gradient from 4 to 35% acetonitrile in 0.05% formic acid and a constant (4 µL/minute) flow rate. The separated peptides were then inserted into an amaZon speed ETD ion trap (Bruker Daltonics) with an Apollo II ESI source to which the LC was coupled. Up to 10 abundant multiply charged species in *m/z* 300-1,300 were selected for collisionally induced dissociation MS/MS after each MS scan, and excluded for one minute after having been selected twice. The systems were controlled by HyStar 3.2 and trapControl 7.1, for the LC and ion trap respectively.

Data analysis

CompassXport 3.0.5 was used to convert raw ion trap data into mzXML format (38). Spectra were matched to peptides using the program X!Tandem (39) of the Trans Proteomic Pipeline, searching each spectrum against our carp database. These peptide spectrum matches (PSMs) were validated by Peptide Prophet (40, 41) resulting in protein identifications exported with a minimum probability of 0.95 and a global false-discovery rate (FDR) of 1%. Proteomics data was plotted against the RNA-Seq data on a log₁₀ scale

in R (ggplot2) (42). Correlations were calculated with the Spearman correlation method in R.

Organ data viewer

The top 10 most highly expressed genes from cluster 6, 11, 12 and 15 were selected. The expression levels of these genes in different organs and tissues were visualized using a script written in R (27), using the maps (43), maptools (44), sp, lattice, latticeExtra (45) and colorspace (46) packages as in a previous presentation of geographic data (47). We created cartoon-like carp shapefiles in QGIS (48) to visualize the carp organs in a spatial context. Each organ or tissue was defined as a polygon and linked with the expression data. Tissues that had many different randomly dissected parts, skin and muscle, were represented with the mean value of all the different sections. For each gene, RNA-Seq expression levels were linked to the indices locating the respective organ or tissue. Finally, this table was used to generate a heat map in a red color scale to represent the levels of expression on a logarithmic scale.

RESULTS AND DISCUSSION

Carp genome assembly

In order to enable isogenic mapping of RNA-Seq data, we have improved the genomic sequence of carp (8) using long-range sequencing information. Using the same DNA sample employed for the initial assembly (8), we have prepared three paired-end libraries using the Cre-LoxP Inverse PCR (CLIP) approach (16) with target insert sizes of 5, 6.5, and 7.5 kbp. For each library 4.5 million read pairs were obtained. In addition, we have sequenced 6.8 Gbp of DNA isolated from nucleated red blood cells using Pacific Biosciences long read technology. All sequencing data was included in an updated *de novo* genome assembly. The new assembly is 1.38 Gbp in length, and has a scaffold N50 of 66.7 kbp. This is considerably less than the 1.71 Gbp genome reported for mirror carp strain Songpu (9). In table 1 we have compared several quantitative features of our genome assembly and that of the Songpu strain. The software program Last (34) was used to align the Chinese Songpu strain to our assembly at DNA level. The results show that the assemblies preserve syntenic relationships for 1.13Gbp, which constitutes 82% and 67% of our assembly and the published assembly of Songpu, respectively.

At present, it is not possible to determine the exact cause of this discrepancy. One possible explanation is that our genome is based on homozygous double-haploid carp, whereas the Songpu genome is based on a heterozygous individual. Variations can be expected since high phenotypic and high genomic variation are compatible with ongoing re-diploidization following the carp-specific genome duplication. The recently published genome of the Atlantic Salmon suggests re-diploidization following genome duplication is accompanied by substantial genomic instability (49). As the carp-specific duplication is much more recent than the salmonid-specific one (~80 Mya), carp genomes might still be experiencing high levels of genomic rearrangements.

However, the discrepancy between the genome assemblies of common carp strains

should be attributed at least in part to the sequencing technologies employed. Both assemblies are primarily based on short-read assemblies, which are not optimally suitable for obtaining precise long-range genomic distance information. Here, we have already demonstrated the usefulness of low-coverage Pacific Biosciences sequencing for augmenting genome assemblies. It will therefore be interesting to further enhance and verify both genome assemblies using emerging single-molecule technologies, such as Oxford Nanopore long read sequencing, or BioNano Genomics optical mapping (50).

Table 1 Comparison of features of the common carp genomic assembly used in this paper and that of common carp strain Songpu (9)

Data set	Genome size	Scaffolds	N50	GC content	Predicted genes	Median gene span	Predicted exons (total/per gene)	Repeats
This work	1.38Gb	80273	67kb	37%	50527	8316 bp	387245/7.7	32%
Xu <i>et al.</i> , 2014 (9)	1.69Gb	2503	1 Mb	36.30%	52610	11980 bp	390620/7.5	31%

Tissue-specific carp transcriptomes

We have sequenced cDNA obtained from 89 tissues of carp and two embryos, with the aim of generating a comprehensive atlas of the carp transcriptome. These data were used for supporting gene prediction and quantification of gene expression.

Genes were predicted both *de novo* and using RNA-Seq data, as described in the material and methods section. We could predict the presence of 50,527 genes, which is similar to the 52,610 genes predicted for strain Songpu (9). The predicted genes are all made available in the NCBI database, thereby enormously expanding the previously few thousand annotated gene set of common carp.

Subsequently, we quantified the expression of all genes in all 91 samples. For initial data exploration, we used these counts to generate a correlation matrix showing the similarity of transcriptomes between all samples across all genes (Additional file 3; figure S2). This analysis revealed four outlier samples (left eye rear, optical nerve, gallbladder e and swim bladder rear), for which the global expression pattern did not match that of replicates but instead appeared intermediate between tissues. In all cases, this could be explained by slight contamination with neighboring tissues during dissection. This is a known issue when dissecting organ tissue in fish (51, 52).

As expected, the initial analysis also revealed that specialized tissues vary widely in gene expression patterns. This poses a problem, as nearly all RNA-Seq analyses implicitly or explicitly assume that samples are comparable (e.g. treated and non-treated). Therefore,

for clustering purposes, we normalized between samples using expression ranks rather than values (see Materials and Methods). This approach, routinely used in microarray analyses, can be employed to make intrinsically dissimilar samples comparable (53). However, when comparing across samples, the resulting expression can only be interpreted qualitatively rather than quantitatively (e.g. 'high in liver/low in eye' instead of 'higher in liver than in eye'). For display purposes, we therefore used RPKM-expression values, which allow for within-sample comparison of expression levels.

Subsequently, we clustered the ranked data for the remaining 87 samples in a correlation matrix as before (figure 2), confirming that the remaining tissues cluster with their own tissue type. To investigate tissue specific RNA expression, all the samples were grouped by tissue based on the correlation between the samples. We found two samples of hepatopancreas (d,e) that clustered more like spleen tissue and one spleen(e) sample that clustered more like hepatopancreas tissue. Due to their close proximity we assume this is due to the dissection of the tissues. We grouped these 3 samples as a mixed hepatopancreas/spleen group.

Based on the correlation plot (figure 2), the tissues were categorized into 16 groups. For each gene the mean of all the samples in one group was taken. In these 16 tissue groups, we investigated possible gene clusters specific for one type of tissue. Using the k-means clustering function in Mev (32, 54), all genes were divided into 16 clusters on ranked data. Within these 16 gene clusters the genes were ranked based on the sum of the overall expression in the samples. We used the top 10 genes to create a heatmap (Additional file 5: figure S3). We highlight four of these clusters (cluster 6, 11, 12 and 15) in figure 3. Gene cluster 6 exhibits highly specific expression in lens tissue only. Gene clusters 11, 12 and 15 show a tissue-specific expression in muscle, heart and gallbladder, respectively. For each cluster, the most specific genes are functionally closely related. For example, the most abundant transcripts in cluster 6 exclusively encode crystallins, which are water soluble structural proteins present in the eye lens and cornea (55). The beta and gamma crystallins are known to have very high homology between carp and zebrafish (56) and differ in their expression levels over the life span (57). Similarly, the muscle-specific cluster (11) predominantly contains the myosin motor proteins essential for muscle contraction. Interestingly, several of these proteins seem to be skeletal muscle (*myhz1.1*) or heart specific (*myl7*). The latter is widely used as a heart tissue specific marker in zebrafish. Chymotrypsinogen 2, specific for the gall bladder and spleen (cluster 15) is known to be excreted in the gastrointestinal track in zebrafish, and in situ hybridization of its homologue *ctrb1*, expression shows it to be expressed in both the pancreas and gastrointestinal track in zebrafish (58).

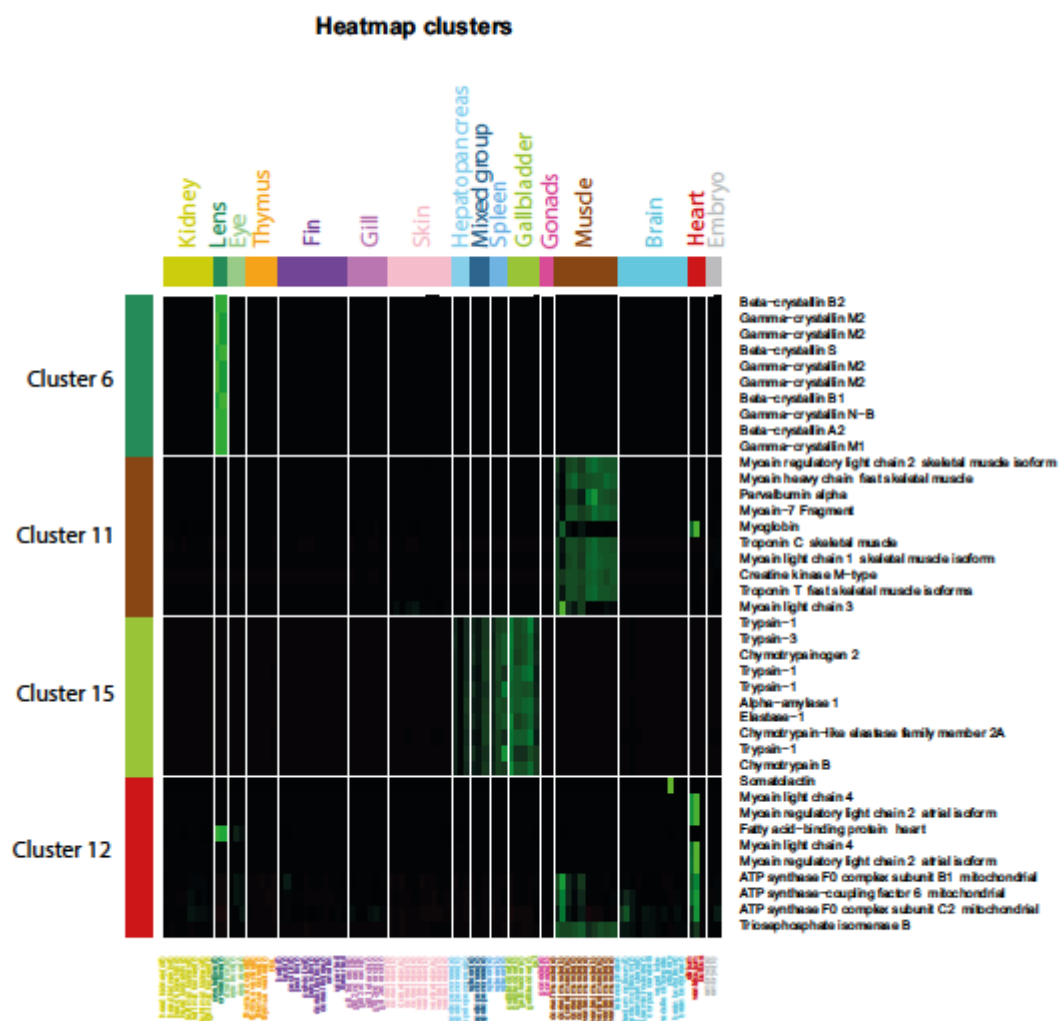


Figure 3. Heatmap on the RPKM-normalized data. Four of the gene clusters (cluster 6-lens, 11-muscle, 12-heart and 15-gallbladder) are visualized, depicting the top 10 highest expressed genes in that respective cluster. The different tissues were colored according to the 16 groups defined in figure 1 (The other gene expression clusters are depicted in Additional file 5: figure S3).

Considering the close genetic relationship between the carp and zebrafish, we tested if the same genes expressed in different zebrafish tissues were also expressed in the carp. From ZFIN we collected lists of genes expressed in different zebrafish tissues, and removed the genes that were not found in our data. In zebrafish, the expression measured for liver and pancreas was combined to compare to our hepatopancreas sample of the carp. For skin we combined the genes found in zebrafish fin and tail tissue. For each tissue, the genes from ZFIN were compared to the genes expressed in the carp.

For the majority of tissues we find that 89.6-98.9% of the genes expressed in zebrafish are also expressed in the same tissue in carp (table 2). The hepatopancreas and spleen

tissues have a lower match 85,2% and 73,6% respectively. As these tissues are known to be very variable between samples this lower match is not surprising and by removing the three potential mixed samples we have still managed to get a very high match with zebrafish expression.

As the carp genome has experienced a recent duplication event since the split with zebrafish, a majority of genes have a duplicate in common carp (8). It is of interest to study whether such ohnologs might have diversified in function from their ancestors as might be indicated by difference in expression patterns. InParanoid (35) was used to generate an orthology mapping between carp and zebrafish. This resulted in 18,241 ortholog group that are presented in Additional file 6. From this orthology mapping we selected 3,549 groups for 1 to 2 mapping between zebrafish and carp. This group probably represents a majority of putative ohnologs that are the results of the carp-specific duplication. This group of putative ohnolog gene pairs has been linked to the expression data of 15 different tissues (Additional file 7: table S3). In Additional file 8: figure S4, a numerical overview is given of differential expression values at an absolute fold change of two and four for this set of 3,549 gene pairs. The results show that in all tissues analyzed over 30% of the gene pairs show an absolute FC change differences of more than 2. Even at fold change 4 cut off many pairs are also differentially expressed in a majority of all organs, e.g. four gene pairs are even differentially expressed at FC 4 in all 15 tissues under consideration (Additional file 8: figure S4). These data suggest that many putative ohnologs have undergone an evolutionary adaptation either towards being less dominant or alternative functions in various tissues.

Table 2. Gene expression comparison between carp and zebrafish tissues.

Organ	Carp only	ZFIN only	Carp & ZFIN (% of ZFIN)
Kidney	17458	57 (7,1%)	741 (92,9%)
Lens	14555	38 (6,2%)	578 (93,8%)
Eye	15041	116 (3,5%)	3181 (96,5%)
Thymus	18673	3 (3,8%)	75 (96,2%)
Fin	16868	39 (1,8%)	2072 (98,2%)
Gill	16964	80 (5,5%)	1371 (94,5%)
Skin	17931	26 (2,7%)	947 (97,3%)
Hepatopancreas	13826	253 (14,8%)	1459 (85,2%)
Spleen	14180	87 (26,4%)	243 (73,6%)
Gallbladder	18011	1 (3,0%)	32 (97,0%)
Gonads	17975	51 (6,1%)	783 (93,9%)
Muscle	16427	93 (5,7%)	1535 (94,3%)
Brain	15219	68 (1,7%)	3871 (98,3%)
Heart	15433	169 (10,4%)	1457 (89,6%)
Embryo	18431	11 (1,1%)	1031 (98,9%)

We compared the list of genes found to be expressed (per tissue) from the ZFIN database to the list of genes found to be expressed in the carp tissue of this study. The numbers given refer to the gene count found exclusively in carp (carp only), ZFIN (ZFIN only) or present in both carp and ZFIN. The overlap in percentage of the total number of ZFIN genes linked to these tissue is also shown.

Integrative data viewer

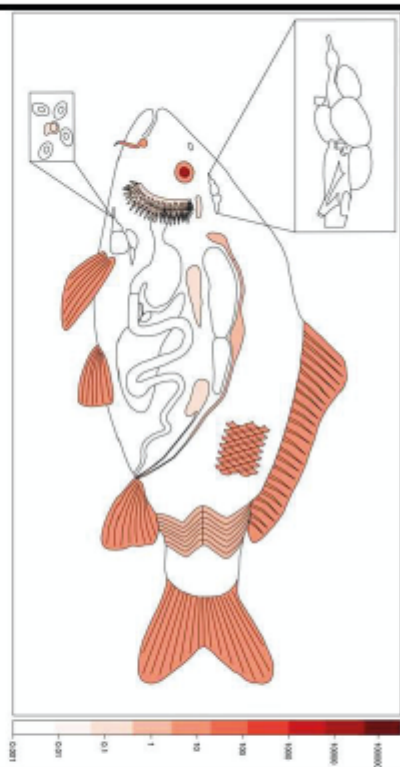
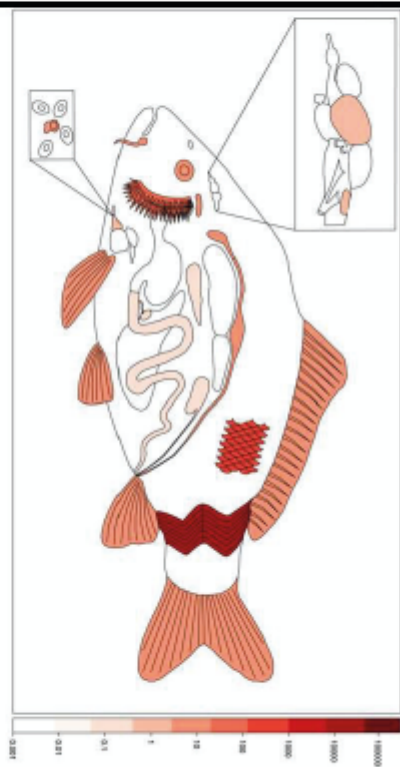
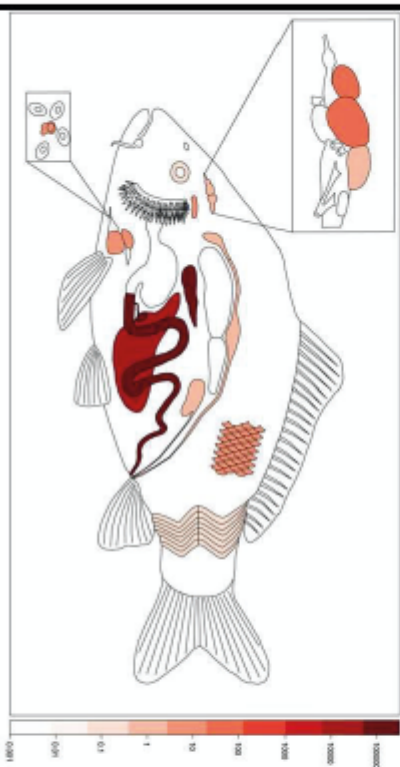
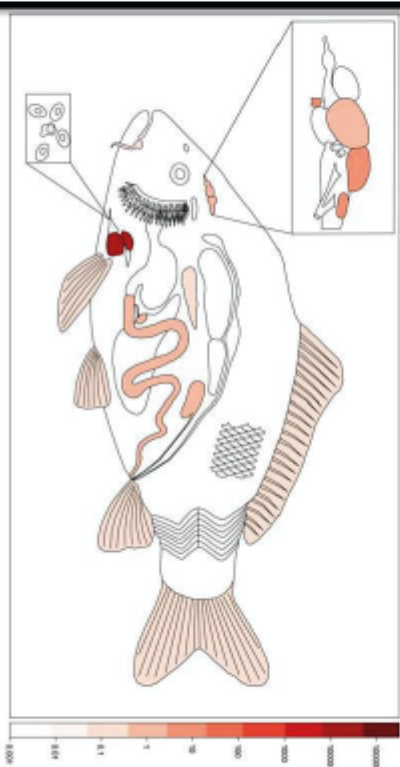
We selected one gene from the top 10 genes from cluster 6, 11, 12 and 15. We used these genes to visualize our transcriptome data in cartoon representation of the carp body plan using R (figure 4). The R script maps the gene and protein expression to individual organs of the carp in a lateral view to obtain a quantitative overview of the organ/organ subpart/tissue when data from an organ pair or from repeated dissections are present. In this manuscript several examples are given, the others can be generated by using the supplementary R script (Additional file 10) and data files (Additional file 9). We are currently working on a web site that will use this script to make it possible to analyze our data in an interactive viewer that will be accessible via the following web address: <http://ms-utils.org/comics/>.

Proteome analysis

For three tissues types (heart, kidney, and hepatopancreas), a proteomics experiment resulted in the identification of 1912 peptides which included 936 uniquely mapped peptides, which were correlated to previously determined RNA levels of these tissues. In total 896 of these peptides were annotated, including 11 histone and 17 ribosomal genes, and a few unknown (figure 5). The correlation between the uniquely mapped peptides and RNA-Seq of each tissue individually were calculated with the Spearman correlation coefficient (ρ). This resulted in a moderate correlation between the two datasets for each organ. The correlations are plotted per tissue type on a log₂ scale (figure 5). The most challenging part was that the majority of proteins were supported by more than one unique peptide. So by selecting only the uniquely mapped peptides for the correlation we lost some power in our analysis, but our results are very similar as found in zebrafish larvae (36).

These data confirm an earlier report on a comparative analysis of transcriptome and proteome analysis of a parallel analysis of embryonic zebrafish samples (8). In this study embryonic gene expression was found to correlate with the abundance of proteins with the exception of some classes of proteins such as exceptions include ribosomal proteins, histones and vitellogenins. Since in several carp organs the expressions of these classes of proteins are variable, the correlation between the transcriptome and proteome data is also variable. However, in general, the measured correlations provide confidence in the analyzed genes and give a good prediction whether they can be used as markers genes at the RNA and/or protein levels in future detailed analyses. More detailed analysis of mRNA-Protein results using approaches such as described by Koussounadis *et al.* (59) and Cheng *et al.* (60) would be desirable in future studies of our data sets.

Figure 4. Tissue specific gene expression visualization in the carp. As an example of this visualization tool, a gene was selected from all four clusters (cluster 6-lens, 11-muscle, 12-heart, and 15-gallbladder) depicted in figure 3. For each gene the RPKM-normalized data (log₁₀) was used. **A:** depicts the carp gene gamma-crystallin M1 (cypCar_00043246) from cluster 6. Gamma-crystallin M1 is one of the three genes in the carp gamma crystallin gene family. **B:** depicts the carp gene myosin heavy chain fast skeletal muscle (cypCar_00029041) from cluster 1. **C:** depicts the carp gene chymotrypsinogen 2 (cypCar_00044522) from cluster 15. **D:** depicts the carp gene myosin regulatory light chain 2 atrial isoform (cypCar_00044421) from cluster 12. The red color scale represents the levels of expression (on a logarithmic scale).

Expression of gene *cypCar_00043246* RAExpression of gene *cypCar_00029041* RAExpression of gene *cypCar_00044522* RAExpression of gene *cypCar_00044421* RA

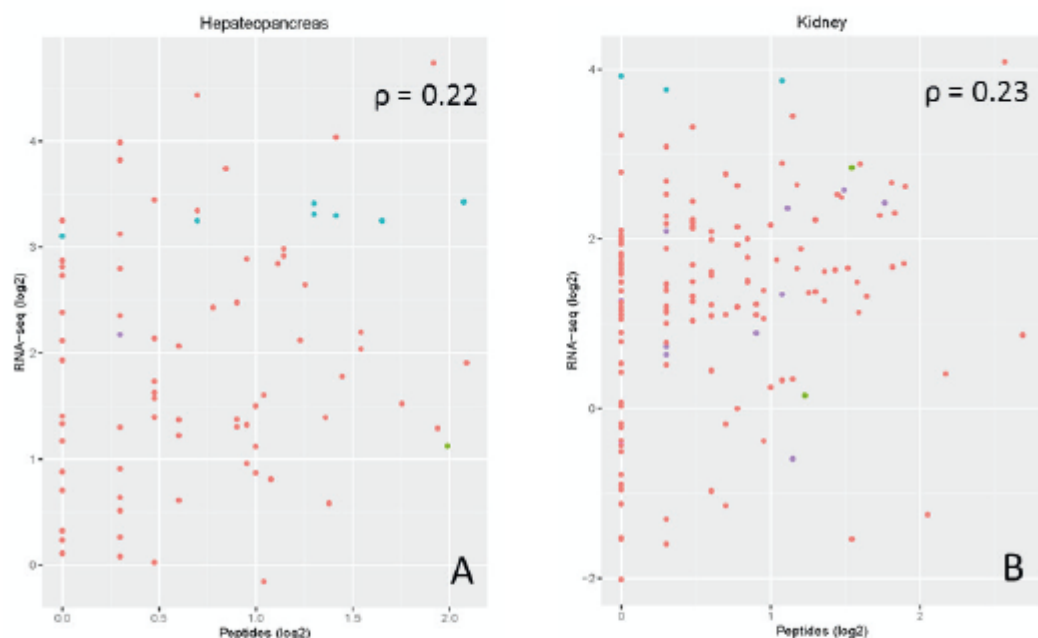


Figure 5. Correlation plots of the RPKM-normalized RNA-Seq and proteins detected data in different carp organs (log2). A: depicts hepatopancreas tissue. B: depicts Kidney tissue. C: depicts heart tissue (see next page). The correlation is calculated by using a Spearman correlation test (correlation coefficient ρ). The genes are colored according to four categories: annotated genes (orange), histones (green), ribosomal (blue) and unknown/unannotated genes (purple).

CONCLUSIONS

In the present study, we have compiled an extensive dataset on tissue-specific expression and translation of genes in the European common carp. These data provide a detailed, sensitive and replicable account of gene activity across tissues. We expect these data to be of use for both carp and zebrafish research, as the vast majority of tissue-specific genes are common to both species. Therefore, the carp data will enable the development of gene expression markers specific for selected organs. Our simple expression viewer will support rapid scanning for such markers. As it can be quickly modified to accommodate additional data, the application can be applied to gene comparison with other cyprinid species, which all share a similar body plan. As an increasing number of cyprinid fish genomes have now been sequenced, this opens up promising opportunities for comparative genomic analyses of evolutionary and developmental biology. In summary, therefore, our combined genome, transcriptome and proteome data of one and the same common carp strain will be a valuable resource for future comparative genomics. Furthermore, these data sets and annotations will further enhance the usefulness of common carp for functional genomic studies.

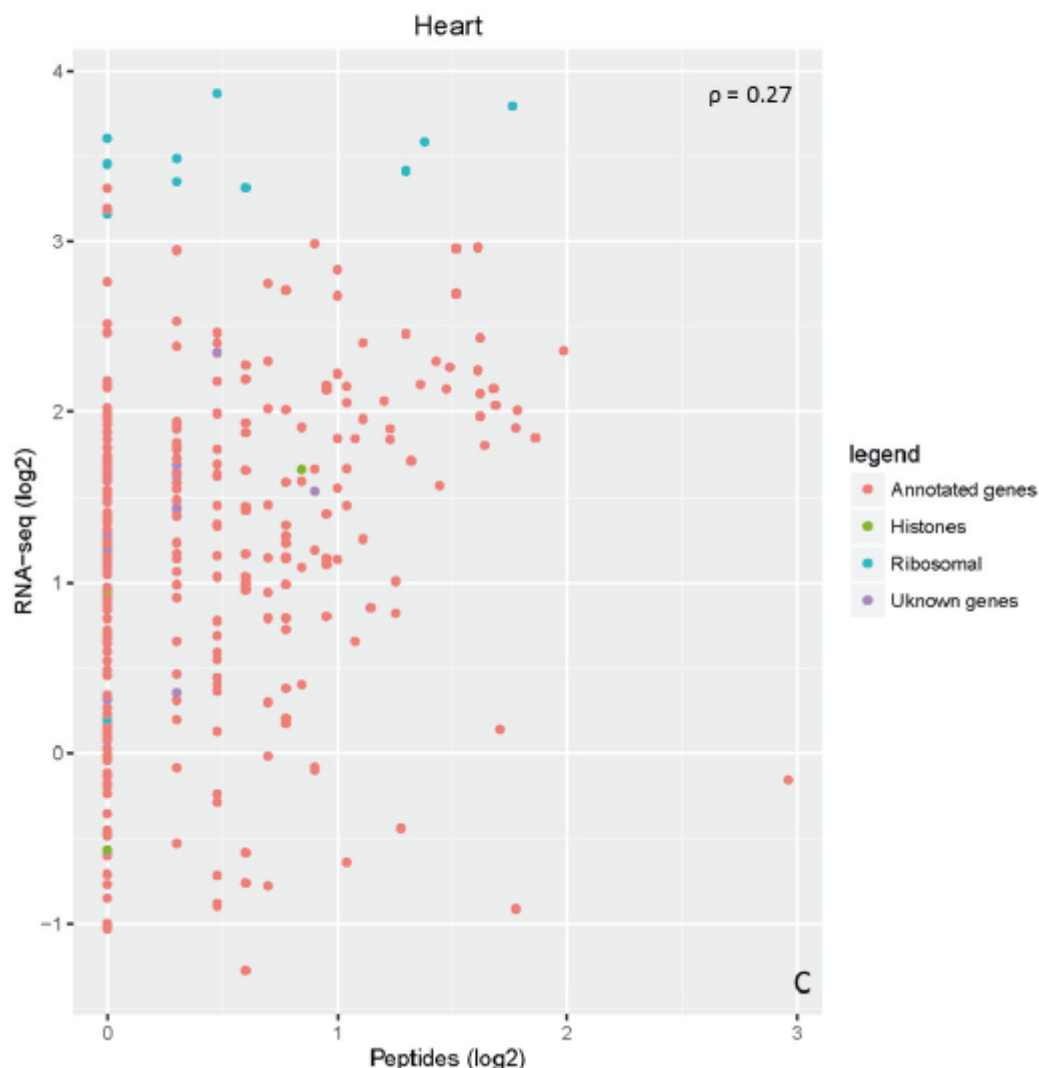


Figure 5 C: (continued from the previous page) Correlation plot of the RPKM-normalized RNA-Seq and proteins detected data in carp heart. The correlation is calculated by using a Spearman correlation test (correlation coefficient ρ). The genes are colored according four categories: annotated genes (*orange*), histones (*green*), ribosomal (*blue*) and unknown/unannotated genes (*purple*).

AVAILABILITY OF DATA AND MATERIALS

All the data (Genome assembly with annotation, RNA-Seq and processed count data) is available online [Accession: PRJNA73579]. The genome assembly and associated annotations can be browsed (jbrowse.genereg.net/index.html?data=data/mpirunAugustus/). The R script and files used to create the organ viewer are available as Additional files 6, 9, and 10. A web portal implementing this viewer for the carp data is under development and will be accessible from: <http://ms-utils.org/comics/>

ADDITIONAL FILES

All additional files can be viewed when scanning the following QR-code and go to "Declarations":



Additional file 1: Table S1. List of the 89 dissected tissues from 19 organs from adult Carps (1 male, 3 female) and their read counts. Total read count was 512.134.436 averaging ~5.754.320 reads per sample. Additionally two 5 dpf embryos were sampled. (XLSX 12 kb)

Additional file 2: Figure S1. Description of data: Density plots of the normalized RNA-Seq count data (log10). **A:** depicts the data normalized using CPM. **B:** depicts the data normalized using RPKM. **C:** depicts the rank data normalized using CPM. **D:** depicts the rank data normalized using RPKM. (PNG 147 kb)

Additional file 3: Figure S2. Correlation plot using the Pearson correlation test on the RPKM normalized data ($n = 91$). The different tissues were colored according to the 16 groups. (PDF 198 kb)

Additional file 4: Table S2. Description of data: Overview of the division of the tissues ($n = 87$) in the 16 clusters. Clusters are based on the correlation between the samples (figure 2). They consist out of biological replicates or closely related tissues. (XLSX 11 kb)

Additional file 5: Figure S3. Heatmap on the RPKM-normalized data for all the 16 tissue clusters. All the clusters are visualized, depicting the top 10 highest expressed genes in that respective cluster. The different tissues were colored according to the 16 groups. The clusters were ordered to match the order of the tissue groups depicted in the columns. (PDF 982 kb)

Additional file 6: FASTA file containing a set of 18,241 ortholog groups that were obtained using InParanoid (35) for orthologs mapping between carp and zebrafish. (FA 1253 kb)

Additional file 7: Table S3. Carp putative ohnolog groups linked to expression data in 15 carp tissues. From 18,241 carp ortholog groups mapped to zebrafish that are presented in the Additional file 6. we selected 3,549 groups for 1 to 2 mapping between zebrafish and carp. This group of putative ohnolog gene pairs have been linked to the expression data of 15 different tissues. A numerical overview of this dataset is presented in Additional file 8: figure S4. (XLSX 2042 kb)

Additional file 8: Figure S4. Numeric overview of differential expression levels of possible paralog pairs. **(A)** Number of paralog pairs that are differentially expressed with a FC of larger than 2 in 15 separate tissues. **(B)** Number of possible paralog pairs filtered on the minimum number of tissues in which a differential expression of larger than FC 2 is observed. E.g. in the bottom of the scale 27 possible paralog pairs have a FC of larger than 2 in all 15 tissue types listed in panel A. **(C and D)** The same analysis as in panels A and B respectively with a FC of 4 as cut-off value. With these criteria there are still 4 possible paralog pairs that have a FC larger than 4 in all 15 tissues. (JPG 130 kb)

Additional file 9: Visualisations of carp organ data. Visualisations that can be used with the supplied R-script. (ZIP 2148 kb)

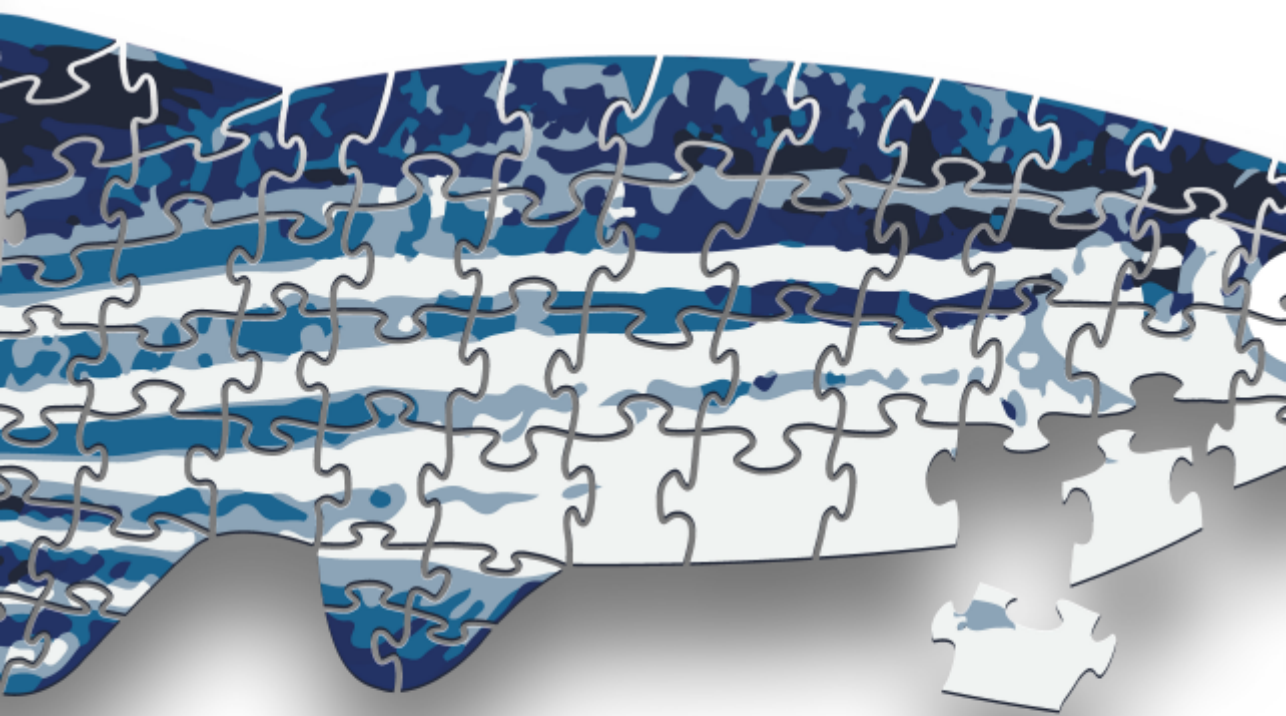
Additional file 10: Carp transcriptome expression visualization tool containing the R-script and input files needed to create these visualization. This includes the carp shape file and RPKM normalized(log10) data for all the proteins measured. (ZIP 7041 kb)

REFERENCES

1. Pietsch C, Hirsch P. Biology and ecology of carp. London: CRC Press; 2015.
2. Dooley, K. & Zon, L.I. Zebrafish: a model system for the study of human disease. *Curr Opin Genet Dev* 10 (2000).
3. Lieschke, G.J. & Currie, P.D. Animal models of human disease: zebrafish swim into view. *Nat Rev Genet* 8, 353-367 (2007).
4. Meijer, A.H. & Spaink, H.P. Host-pathogen interactions made transparent with the zebrafish model. *Curr Drug Targets*. 12, 1000-1017 (2011).
5. Larhammar, D. & Risinger, C. Molecular genetic aspects of tetraploidy in the common carp *Cyprinus carpio*. *Molecular phylogenetics and evolution* 3 (1994).
6. David, L., Blum, S., Feldman, M.W., Lavi, U. & Hillel, J. Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. *Mol Biol Evol* 20 (2003).
7. Zhang, Y. et al. Identification of common carp innate immune genes with whole-genome sequencing and RNA-Seq data. *J Integr Bioinform* 8 (2011).
8. Henkel, C.V. et al. Comparison of the exomes of common carp (*Cyprinus carpio*) and zebrafish (*Danio rerio*). *Zebrafish* 9 (2012).
9. Xu, P. et al. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nat Genet* 46 (2014).
10. Wang, J.T., Li, J.T., Zhang, X.F. & Sun, X.W. Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (*Cyprinus carpio*). *BMC Genomics* 13 (2012).
11. Ji, P. et al. Characterization of common carp transcriptome: sequencing, de novo assembly, annotation and comparative genomics. *PloS one* 7 (2012).
12. Imnazarow I. Genetic variability of Polish and Hungarian carp lines. *Aquaculture*. 1995;129:215-219.
13. Komen J, Bongers AB, Richter CJJ, van Muiswinkel WB, Huisman EA. Gynogenesis in common carp (*Cyprinus carpio* L.): II. The production of homozygous gynogenetic clones and F1 hybrids. *Aquaculture*. 1991;92:127-142.
14. Wiegertjes GF, Stet RJ, Van Muiswinkel WB. Divergent selection for antibody production to produce standard carp (*Cyprinus carpio* L.) lines for the study of disease resistance in fish. *Aquaculture*. 1995;137(1-4):257-262.
15. Wiegertjes, G.F. et al. Characterization of isogenic carp (*Cyprinus carpio* L.) lines with a genetically determined high or low antibody production. *Anim Genet* 27 (1996).
16. Peng, Z. et al. Generation of long insert pairs using a Cre-LoxP Inverse PCR approach. *PloS one* 7 (2012).
17. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27 (2011).
18. English, A.C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one* 7 (2012).

19. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker open-4.0. (2013).
20. Jurka, J. et al. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110 (2005).
21. Cantarel, B.L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18 (2008).
22. Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinf* 7 (2006).
23. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30 (2014).
24. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14 (2013).
25. Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* 25 (2009).
26. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics* 31 (2015).
27. R Core Team (R Foundation for Statistical Computing, Vienna, Austria; 2015).
28. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinf* 11 (2010).
29. Ager-Wick, E. et al. The pituitary gland of the European eel reveals massive expression of genes involved in the melanocortin system. *PloS one* 8 (2013).
30. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (2010).
31. Wei, T. corrplot: visualization of a correlation matrix. (2013).
32. Saeed, A.I. et al. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34 (2003).
33. Warnes, G.R. et al. gplots: various R programming tools for plotting data. (2015).
34. Kielbasa, S.M., Wan, R., Sato, K., Horton, P. & Frith, M.C. Adaptive seeds tame genomic sequence comparison. *Genome Res* 21 (2011).
35. Sonnhammer, E.L. & Ostlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* 43 (2015).
36. Palmblad, M. et al. Parallel deep transcriptome and proteome analysis of zebrafish larvae. *BMC Res Notes* 6, 428 (2013).
37. Plas-Duivesteljn, S.J. et al. Identifying proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *J Proteome Res* 13 (2014).
38. Pedrioli, P.G. et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* 22, 1459-1466 (2004).
39. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*. 20, 1466-1467 (2004).
40. 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.* 74, 5383-5392 (2002).
41. Ma, K., Vitek, O. & Nesvizhskii, A. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* 13, S1 (2012).
42. Wickham, H. ggplot2: elegant graphics for data analysis. (Springer, New York;

- 2009).
43. Becker, R.A., Wilks, A.R., Brownrigg, R., Minka, T.P. & Deckmyn, A. maps: draw geographical maps. (2015).
 44. Bivand, R. & Lewin-Koh, N. maptools: tools for reading and handling spatial objects. (2015).
 45. Pebesma EJ, Bivand RS. Classes and methods for spatial data in R. R News. 2005;5(2). <http://cran.rproject.org/doc/Rnews>.
 46. Ihaka, R., Murrell, P., Hornik, K., Fisher, J.C. & Zeileis, A. colorspace: color space manipulation. (2015).
 47. Lamigueir OP. Maps with R (III). online; 2012.
 48. QGIS_DevelopmentTeam. Geographic Information System. Geospatial Foundation Project. 2009. <http://qgis.osgeo.org>.
 49. Lien, S. et al. The Atlantic salmon genome provides insights into rediploidization. Nature 533 (2016).
 50. Spalink, H.P., Jansen, H.J. & Dirks, R.P. Advances in genomics of bony fish. Brief Funct Genomics 13 (2014).
 51. Rombout, J.H. & Taverne-Thiele, J.J. An immunocytochemical and electron-microscopical study of endocrine cells in the gut and pancreas of a stomachless teleost fish, *Barbus conchonus* (Cyprinidae). Cell Tissue Res 227 (1982).
 52. Rombout, J.H., Grinten, C.P., Binkhorst, F.M., Taverne-Thiele, J.J. & Schooneveld, H. Immunocytochemical identification and localization of peptide hormones in the gastro-entero-pancreatic (GEP) endocrine system of the mouse and a stomachless fish, *Barbus conchonus*. Histochemistry 84 (1986).
 53. Ager-Wick, E., Henkel, C.V., Haug, T.M. & Weltzien, F.A. Using normalization to resolve RNA-Seq biases caused by amplification from minimal input. Physiol Genomics 46 (2014).
 54. Howe, E.A., Sinha, R., Schlauch, D. & Quackenbush, J. RNA-Seq analysis in MeV. Bioinformatics 27 (2011).
 55. Jester, J.V. Corneal crystallins and the development of cellular transparency. Semin Cell Dev Biol 19 (2008).
 56. Wistow, G. et al. gammaN-crystallin and the evolution of the betagamma-crystallin superfamily in vertebrates. FEBS J 272 (2005).
 57. Grelling, T.M., Houck, S.A. & Clark, J.I. The zebrafish lens proteome during development and aging. Mol Vis 15 (2009).
 58. Zhang, S. et al. Heme acts through the Bach1b/Nrf2a-MafK pathway to regulate exocrine peptidase precursor genes in porphyric zebrafish. Disease models & mechanisms 7 (2014).
 59. Koussounadis, A., Langdon, S.P., Um, I.H., Harrison, D.J. & Smith, V.A. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. Sci Rep 5 (2015).
 60. Cheng, Z. et al. Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. Molecular systems biology 12 (2016).



Chapter 5

Zebrafsh Models in Proteomics

**S.J. van der Plas-Duivesteijn, E.L. Benard, I.C.R.M.
Kolder, K.A. Stam, A.H. Meijer, M. Palmblad**

Ready for submission

ABSTRACT

Zebrafish models have been shown to be powerful tools in genomics studies for drug discovery and drug development and their application in biomedical research is increasing. The use of this animal model in the field of proteomics is also accumulating, although it is still less frequently used within this omics field than in the field of genomics. In this review we provide an overview of the possibilities, limitations and practical considerations particular to this animal model in proteomics, and evaluate available tools and resources suitable for zebrafish proteomics.

INTRODUCTION

The zebrafish (*Danio rerio*) is a tropical fish originating from Southeast Asia, where it lives in brackish waters, and is a member of the Cyprinidae family. It was introduced as an animal model in the 1970s (1). The use of the zebrafish as an animal model in biomedical research has emerged over time (2-6). Currently, mutant strains and zebrafish cell lines are available (7). Clearly there are major differences between the anatomy of human and zebrafish, reflecting their evolutionary distance. Nevertheless, despite their taxonomic distance, approximately 70% of their protein-coding genes shows strong similarities (8). In addition, amino acid sequences of proteins are evolutionary conserved between mammals and zebrafish (9-11). Also, several components of the main immune signalling pathways (e.g. Toll like receptors, and Myeloid differentiation primary response gene 88) are strongly conserved in both fish and mammals (5, 12-15). Summarized, these studies indicate the homology of the zebrafish animal model to mammals on anatomical, physiological and molecular levels, and highlight the applicability of the zebrafish animal model in biomedical research.

Many studies that make use of the zebrafish animal model are focussed on genomics and transcriptomics. Gaining a more complete insight on e.g. immune responses to infection and drug treatment, requires a multiple omics approach (including genomics, transcriptomics, proteomics, and metabolomics) (16-18). Proteomics, the study of all proteins of an organism or part of an organism, is a growing field of research. Proteomics is important because it describes direct bio-functional signalling, reflecting the operating machinery for nearly all biological functions. Mapping proteins and protein interaction, will therefore contribute to the clarification of biological pathways followed by the understanding of diseases and development of possible treatments.

A quick bibliometric search of zebrafish proteomics clearly shows that the model system is emerging in this field, with the number of publications rising from 9 in 2001 to 245 in 2017 (figure 1). Launching the first zebrafish spectral library (2014), was an important addition toward the maturation of this model species within proteomics (19). The experimental knowledge on the application of the zebrafish animal model in proteomics, is summarized here by reviewing the possibilities, limitations and practical considerations of the model system. In addition, we evaluate available tools and resources suitable for zebrafish proteomics.

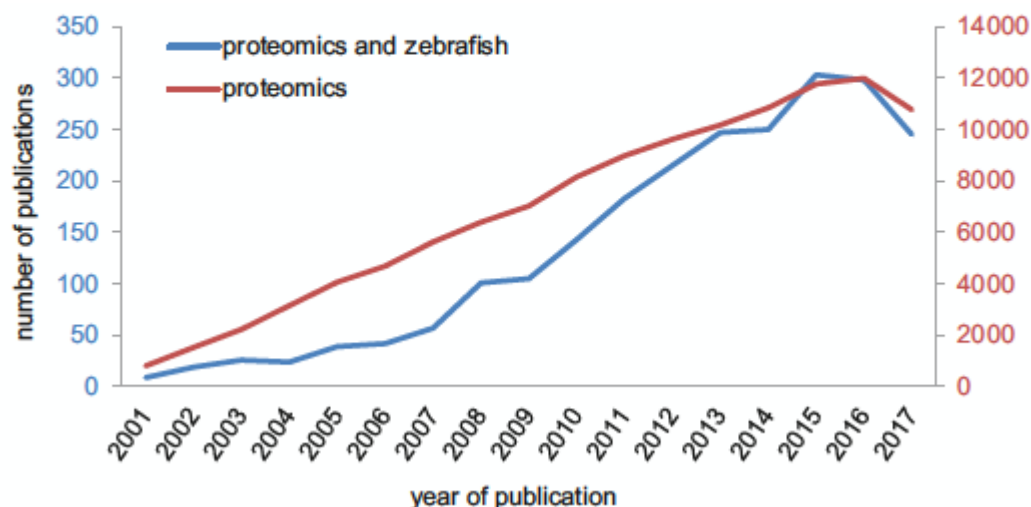


Figure 1: Plot indicating the increase in the number of publications of proteomics studies applying the zebrafish animal model (left axis, blue), compared to the increase of proteomics studies in general (right axis, red) for terms "proteomics AND zebrafish" and "proteomics", respectively. The dip in 2017 is caused by a delay between publication and full indexing and availability in PubMed. Literature search using: <https://europepmc.org>, on April 9, 2018, using the search between publication and full indexing and availability in PubMed.

THE ZEBRAFISH AS AN ANIMAL MODEL

Multiple animal models are available, each with their own advantages and disadvantages (2). The zebrafish animal model is widely used and accepted within human disease studies, including drug discovery and development. Even though anatomically different to human, the zebrafish animal model shows a close genetic relation to human (often represented in phylogenetics) (8) with similarities to several components of immune signalling pathways (5, 12-15), demonstrating its applicability to such studies. As in mouse studies, many strains have been developed for the zebrafish animal model (7, 20). Like the human population, wild type zebrafish strains contain high numbers of polymorphisms, which make the zebrafish animal model extremely relevant to apply in human health studies (21, 22).

An experimental design to study disease mostly includes the early life stages (embryos and larvae) of the zebrafish, which enable the application of high throughput screening (HTS). The natural transparency of the embryos allows the use of fluorescence techniques to follow-up on infection *in vivo* (23-30). In addition, innate immunity (functioning from 24 hours post fertilization) and acquired immunity (not active during the first weeks, and matures at 4-6 weeks) can be studied separately (31), which is a powerful tool to study immune reactions in drug discovery and development. Due to the advantages of transparency, zebrafish strains that remain transparent as adults are now also available (32). In order to meet biomedical study conditions, various mutant strains and zebrafish cell lines have been created (7, 33).

The use of animal models in a laboratory setting is ethically restricted to certain rules. The reflection on reduction (minimise animal use), replacement (methods, strategies and approaches not making use of animal models), and refinement (improvement of procedures and welfare to minimise pain, suffering, distress or lasting harm) should be applied to each study using animal models (34). This widely accepted reflection is meant to protect animal health and welfare, and should exactly describe sedation of and humane endpoints for the animals. Unfortunately directives and policy on the protection of animals used for scientific purposes show large variations between national and regional authorities with respect to application and endpoint (35). According to the EU Directive 2010/63/EU, early life-stages of zebrafish are not protected as animals until the stage where it is capable to feed independently (for zebrafish at 5 days/120 hours post fertilization) (36). This means ethical approval is not required for working with zebrafish embryos and larvae until 5 dpf in scientific studies (another practical advantage when applying embryos/larvae compared to the use of adult stages within a study). The defined stage of independent feeding is influenced by temperature, local conditions, and genotype, causing variation in the defined stage from 5–6 dpf. These factors should be considered to meet the ethical rules that protect this animal and proper staging of larvae 5 dpf should follow the morphological criteria by Kimmel *et al.* (37). Directives and policy are constantly moving, and should therefore be taken into account as well when applying animal models in scientific studies.

PRACTICAL CONSIDERATIONS IN A PROTEOMICS EXPERIMENTAL DESIGN

As mentioned before the zebrafish can be divided in two main categories as an animal model: adult and developmental stages (embryo/early larvae). Adults are preferably kept in groups, where embryos can be separated, e.g. in 96 well plates. In addition, zebrafish adults produce approximately 200 transparent eggs per pair per week, and development progresses fast, while breeding fish until adult stage takes months. This indicates that embryos are more suitable for high throughput processing and screening than adults. High-throughput screening (HTS) is an important method used in human drug discovery and development. The increased use of zebrafish as an animal model and its applicability for HTS has led to the development of several laboratory techniques. For example, a spawning system, named Ispawn, was developed to scale up zebrafish breeding and easily collect eggs (38). A robot injector automated infections of zebrafish embryos, and the additional developed COPAS XL (Union Biometrica) automatically sorts diseased embryos/larvae based on the illuminance of fluorescent signals that can be labelled to the source of infection (e.g. fluorescent bacteria or cancer cells) (23, 28). In addition, several routes of injections are well established for zebrafish (39). Moreover, techniques using fluorescently labelled cells (e.g. macrophages or neutrophils) assist in visualizing the effect on a cellular level *in vivo* without sacrificing the organism. (30, 40). Furthermore, genomic and transcriptomic screens can be applied to the model, providing an overview of most genes active or inactive at that moment, often visualized in heat maps. In combination these techniques make the zebrafish animal model suitable for HTS (41). Within proteomics, a total protein map, visualized in

something like a proteomics heat map, would be desirable to level with genomic and transcriptomic studies. The above mentioned techniques highlight the possibility to apply the zebrafish animal model in HTS, but what about its application to proteomics at present? We therefore review the practical considerations of the zebrafish animal model and its use in proteomics step by step.

Sample preparation

Sample preparation is a first step in proteomics. Where adult zebrafish can be cleared from water, larvae and embryos cannot be processed without a little excess fluid which can interfere in the proteomics extraction process. In addition, zebrafish embryos require additional dechorionation and deyolking steps prior to protein extraction.

The embryonic chorion can easily be digested using pronase. A solution of 2 mg/ml pronase (Roche Applied Science) in egg water (42, 43) can be applied to large badges of embryos at once in a petri dish. Embryos younger than the tailbud stage require coated dishes, since they stick to the plastic of the petri dish. A disadvantage of pronase is its protease activity, which is not desired when applying proteomics, and therefore requires several washing steps to remove (43). Even though robotic chorion removal is available (44), a technique that makes use of 96 well plates, used to expose and remove fluids in HTS techniques like the fish and chips microfluidic perfusion platform (45, 46), would be more desirable for HTS.

For nutrition, zebrafish embryos and early larvae rely on a yolk sac throughout development until exogenous feeding begins (5-6 days post fertilization, depending on water temperature, local conditions, and genotyping). The high abundance of yolk proteins (mainly consisting of vitellogenins) seems a major obstacle when applying proteomics, due to its interference with any proteomic application (43, 47). To enable the detection of less abundant proteins, it was suggested to apply deyolking, a procedure to manually separate and remove the yolk from the body, prior to protein extraction (43, 48). A later study on the impact of deyolking, described that deyolking depleted molecular pathways associated with calcium signaling and immune response signaling (49). It was therefore advised to use both yolk-intact and deyolked embryos to maximize protein coverage for shotgun proteomics studies (49). Furthermore, deyolking becomes more difficult throughout development, since the surrounding tissues are becoming more rigid (S.J. van der Plas - Duivesteljn, personal communication). In addition, deyolking does not result in complete yolk removal, leaving the remaining abundant yolk proteins to interfere the detection of less abundant proteins. According to our own experience, we mark deyolking as a time consuming procedure which has a low accuracy. With the development of multiple omics and drug screens in mind, which require HTS, we would advise to invest on developing automated and faster deyolking methods. For example developing antibody coated LC-columns that extract vitellogenins proteins in an automated way e.g. like IgY RuBisCO coated columns to remove Ribulose-1,5-bisphosphate carboxylase/oxygenase, the most abundant protein in plants (50).

When applying zebrafish embryos into an experimental design, their natural

development should be considered carefully. Within 5 days a fertilized zebrafish embryo develops from a fertilized egg into a free swimming larva (37). During this process, developmental variability can occur (51), which can be caused by several parameters (52). When separating proteins from different embryonic/larval stages using SDS-PAGE gel, the magnitude of changes between the proteomes becomes clear (figure 2), indicating the importance of developmental synchronization, which was also addressed earlier (4). Even though the use of embryos from the same lay of a single pair of adults will be synchronous enough for most studies, with fast developing inventions at present, precise staging of embryos might become a robotic procedure (using automated image recognition) in the near future to serve advanced HTS.

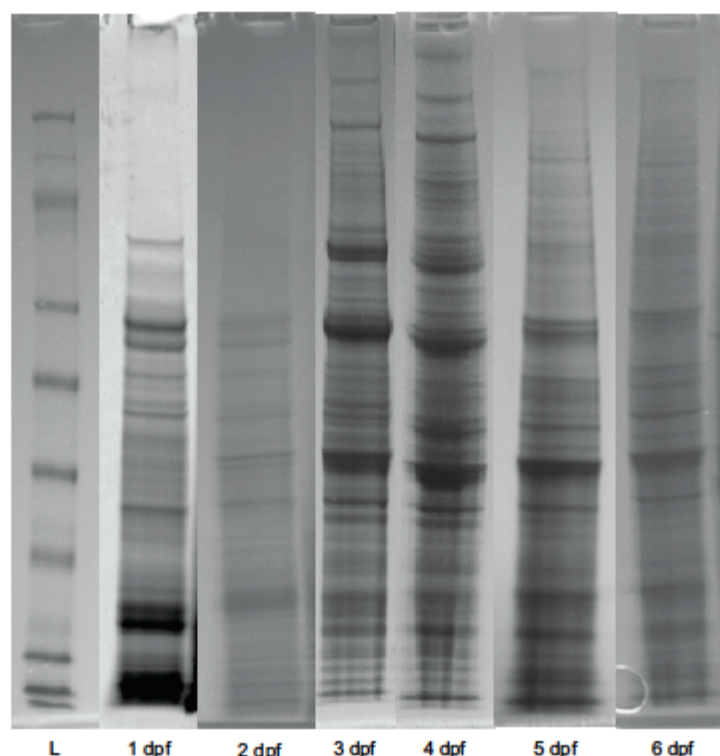


Figure 2: Zebrafish embryo/larva proteome at different developmental stages (1-6 dpf) visualized using SDS-PAGE gel. Figure produced from SDS-PAGE separation used to build the zebrafish spectral library (19).

Protein extraction

A second step in proteomics is protein extraction. Even though essential, a (commercial available) protein extraction kit, validated for the zebrafish animal model, is still lacking. Currently, protein extraction of zebrafish is based on methods described in literature, which have not been validated in a comparative experiment. We here review the different methods found in literature. For homogenization, an essential step in protein

extraction, adult zebrafish can be grinded to powder by using liquid nitrogen. Methods applied to homogenize embryos are sonication (48), freeze thaw cycles (49) or making use of the bullet blender (19). Contrarily it was mentioned that early zebrafish embryos will dissociate rapidly in protein extraction buffer (43).

One advantage of using the bullet blender to homogenize is that it does not require prior dechoriation and it can be applied to adult tissues (more rigid than embryonic tissue) as well (19, 53). In addition, the bullet blender is able to process 24 eppendorf tubes in one automated run, and is a closed system, preventing human keratin contamination. Unfortunately the bullet blender cannot process entire zebrafish adults. Contrary, the use of liquid nitrogen and micropestles increase the risk of keratin contamination to the samples and increases the risk of cold burn to the scientist, and moreover, is a time consuming and manual procedure (which often result in lower accuracy compared to automated procedures).

Different protein extraction buffers used to extract proteins from homogenized samples are described in literature. The use of different protein extraction buffers and methods can lead to a variety in study results. Where one method shows abundance in cell membrane proteins, the other will show a higher abundance in nuclear proteins (48). Therefore sequential fractionation is advised (54). Nevertheless this method does not emphasize HTS.

Another consideration is the addition of protease inhibitors to the protein extraction buffer (55). Even though not explicitly described in literature, adding protease inhibitors to the protein extraction buffer has a positive effect when applied to study zebrafish proteomics (S.J. van der Plas – Duivesteljn, personal communication, (19, 43)).

When applying a shotgun bottom-up approach within proteomics, another challenge is the concentration of proteins that can be extracted from a sample. The experimental design for the creation of a zebrafish spectral library to cover deeper protein coverage, described the use of more than one organ to reach sufficient protein concentration levels (19), while for other experimental designs a single embryo is sufficient (48). In addition, the amount of embryos required for cell dissociation, to study protein effects *in vivo* on cellular level, limits its use to be applied to shotgun bottom-up approach studies (S.J. van der Plas – Duivesteljn, personal communication).

In summary, a validated commercially available protein extraction kit for the zebrafish animal model will hopefully soon be available, to facilitate the use of this animal model in proteomics screening.

Protein extraction is followed by liquid chromatography (LC) separation of the peptide mixture, or SDS fractionation and thereafter LC separation. The last laboratory based step is the detection of the peptides by mass spectrometry. Methods applied to protein mixtures and mass spectrometer settings are well described in literature and do not require specific alterations for its use on zebrafish material, and therefore is not further described in this review.

Data analysis

Finally in these practical considerations, we review proteomics data analysis. Many different software tools are available to match tandem mass spectra to peptides, identifying the peptide and infer the protein (56).

For a standard bottom-up shotgun proteomics experiment, peptide identification can be achieved by at least three different methods: *de novo* sequencing, searching against sequence databases, and matching against real spectra in spectral libraries. *De novo* sequencing does not require any a priori sequence information from the species, but is the least sensitive method of these three. Database searches correlates acquired MS² spectra with *in silico* generated spectra from a protein sequence database, and works best for completely sequenced organisms, although false-discovery rates can be estimated for partially sequenced ones as well (57). Peptide identification using spectral libraries are based on the correlation between query spectra and previously acquired spectra. This is the most sensitive method, but requires both the sequence database and a diverse spectral library. The search space when using spectral libraries is typically smaller than in database searches, and faster when searching large datasets, especially when allowing for variable post-translational modifications (58, 59). In 2014, we released a spectral library for zebrafish (19). As is the case with the spectral libraries available for other model species, the library can be extended by inclusion of data from additional proteomics experiments, or rebuilt using data from more recent instruments. To be able to detect novel proteins, peptides or modifications, a combination of database and library search methods is advised (19). This can be practically accomplished in several freely available tools, such as the Trans-Proteomic Pipeline (60), using Comet for the sequence database search, SpectraST for the spectral library search, PeptideProphet to validate the peptide-spectrum matches, and iProphet to integrate the database and library search results.

After peptide identifications are aggregated and proteins inferred and quantified, zebrafish proteomics data can be further analysed and presented much like quantitative proteomics data from any other organism, including gene ontology and pathway enrichment analysis without necessitating conversion to orthologs from a different species. For example, there are currently (April 2018) 1,560 *Danio rerio* pathways in Reactome, which is comparable with mouse (1,603 pathways). The Comparative GO web tool was recently updated (61) and extended for zebrafish omics data. Though many of these tools were developed for genomics data, they work equally well for proteomics data, as proteomics data is typically gene-centric. Proteins are identified by association with a particular gene and each protein identification (using the UniProt FASTA file or our spectral library) also contains the gene name. This means the quantitative proteomics data can be directly plugged into these and many other tools for GO and pathway enrichment analysis and visualization.

PROTEOME PHYLOGENETICS SUGGESTS APPLICATION OF COMPLEMENTARY MODELS

Shared ancestry and similarities between humans and animal models are crucial in hypothesis determination, and therefore important to apply when designing experiments for human disease and drug development. The most currently used methods to study molecular evolution/phylogenetics, make use of DNA sequences (62). However, before high-throughput DNA sequencing methods were developed, patterns of proteins or enzymatically cleaved peptides (which were separated by chromatographic or electrophoretic methods) were used (63, 64). Recent proteomic studies that applied phylogenetics, showed that, by using stable and abundant proteins, it is possible to analyse much older specimens than with DNA-based methods (65-69). Additionally, proteomic phylogeny demonstrated new discoveries in evolutionary associations (70). The comparison between DNA and protein is not trivial, since one gene, via mRNA processing and post-translational modification (PTM) of the protein sequence, can give rise to more than one *proteoform* (and thus finally to more than one biological function). This in turn may affect the distance as measured by shared peptide or protein features, independent of the full DNA sequence for the same protein. A recent study explored phylogenetic relationships of primates, using a purely proteomics approach, with an algorithm that directly compares tryptic peptide tandem mass spectra between species (71). This method, based on a common bottom-up or “shotgun” proteomics method, may also be applicable in the design of proteomics experiments, and can be informative in choosing an animal model system.

In proteomics studying the frequency of shared tryptic peptides between organisms is of great importance, as these can cause interferences in some experiments while being beneficial in other studies. For example, shared peptides synthesized as stable isotope labelled standards can possibly be used for both human and animal studies, which will reduce the need to synthesize and validate such peptides for each (model) species separately. Furthermore it may be interesting to know whether spectra can be used across (model) species when identifying peptides using spectral libraries and no library for the species of interest is available.

In 2014, the first zebrafish spectral library was published (19), based on a novel approach of using dissected zebrafish organs to produce a wider proteome coverage and better spectral quality. Unfortunately, due to the small size of the zebrafish, not all organs could be easily dissected, for example it is difficult and extremely time consuming to dissect the different parts of the heart. In addition, embryos are even smaller, making it challenging to accomplish deep proteome analysis in single embryos, especially if one wants to perform protein fractionation or enrich PTMs. Moreover, human and mouse libraries contain spectra from specific cell lines as well. Even though zebrafish cell lines are available (33, 72), they were not added to the library at present. A recent study showed the applicability of highly conserved peptides, bridging the gap of the zebrafish models limitations due to its small size, by using a larger but closely related carp (*Cyprinus carpio*) (73). In addition, where zebrafish produce hundreds of eggs per

female, carp lay several hundreds of thousands of eggs per female. Especially in high throughput screens, the amount of eggs produced can serve as a limitation, therefore, another advantage of using carp as a complementary model to zebrafish, is the larger production of carp eggs. This study, in which the limitations of one model are addressed by the other, stresses the importance of applying “proteome phylogenetics”, especially since it was shown that applying highly conserved peptides between model species can be reapplied in related animal models (73).

RESOURCES

Many online recourses are available assisting the application of the zebrafish animal model for research purposes (74). For the field of proteomics no resource is available focussed on the zebrafish animal model alone. Instead, proteomics resources are arranged/centralized based on this omics field explicitly. The largest information resource for proteomics is Uniprot (75). This online resource provides proteomes for species that have their complete genome sequenced, containing a reference proteome for zebrafish. Uniprot provides high-quality data of protein sequences accompanied with functional information. PeptideAtlas (76), a public accessible compendium of peptides that are identified in large tandem mass spectrometry proteomic experiments, released a first zebrafish build on April 25, 2014 (77). Unfortunately, this resource is built on data from a pending publication and unpublished results. For spectral library searching, the National Institute of Standards and Technology (NIST) develops peptide mass spectral libraries with the purpose to provide peptide reference data for disease-related biomarker discovery (78). On September 1, 2015, NIST released the first zebrafish spectral library (19, 78). Our recently published omics data visualization tool, COMICS, can generate anatomical representations of gene and protein expression, including for zebrafish (figure 3), stimulating a multiple omics approach (79). Conglomerate resources for zebrafish only are also available, but are focussed on laboratory methods, strains, cell lines, genomics, and transcriptomics (table 1) (74). Via these well-known resources by researchers applying the zebrafish animal model, notifications to protein repositories can be mentioned. On April 20, 2015, the zebrafish spectral library was mentioned which redirected to the NIST webpage via an active link. With above described proteomics resources for the zebrafish animal model, and to avoid an ocean of online webpages developed for the zebrafish animal model, we see possibilities for implementation of proteomics data within existing genomic recourses.

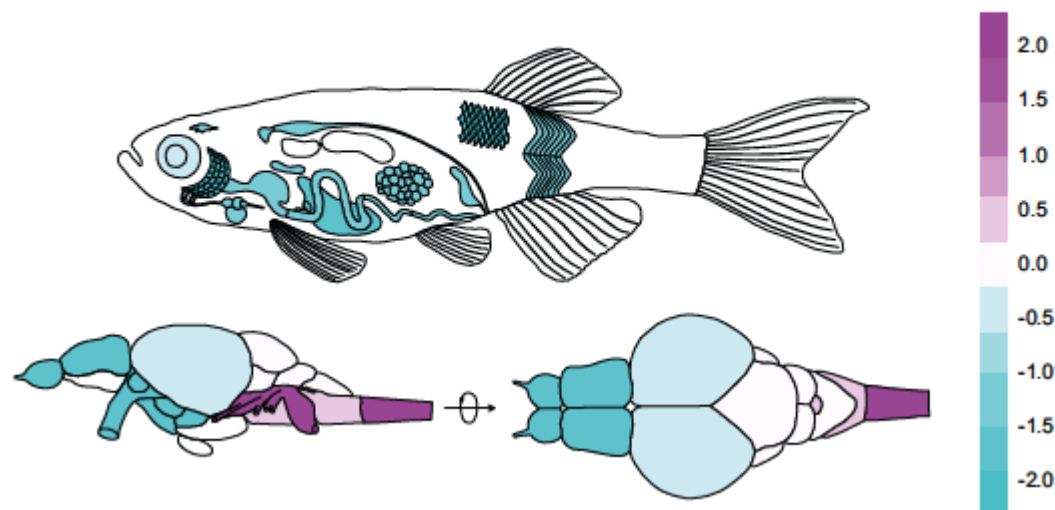


Figure 3: Log10 ratio of recorded observations of *ca2* (cyan) and *mbpa* (magenta) gene expression in ZFIN visualized by COMICS (79). This particular illustration primarily summarizes what information is available in the ZFIN gene expression database, but also says something about the spatial dependence of expression of these two genes. Any anatomically annotated omics data can be visualized by this software tool. The use of the ZFA anatomical ontology is critical to map the observations from multiple experiments to the same layer or level of granularity in the anatomogram.

Table 1: Zebrafish animal model resources

Name of resource	Aim of the resource	Webpage of the resource
ZFIN	zebrafish model organism database	https://zfin.org
zebrafish international resource centre (ZIRC)	repository for zebrafish strains, material and research information	http://zebrafish.org
Zebrafish book (45)	a practical handbook for zebrafish protocols	https://zfin.org/zf_info/zfbook/zfbk.html and https://wiki.zfin.org/display/prot/ZFIN+Protocol+Wiki
The Journal of visualized experiments (JOVE)	creating and publishing videos of scientific experiments allowing scientists, educators and students to see the intricate details of cutting-edge experiments rather than read them	https://www.jove.com

Zebrafish-specific and general resources for this animal model that are widely used among researchers in the field. ZFIN and ZIRC are sister organizations. JOVE is a repository of tutorial and practical video instructions that include videos on mass spectrometry, proteomics, data analysis as well as actual handling of the zebrafish.

EXAMPLE: ZEBRAFISH MODEL TO STUDY TUBERCULOSIS INFECTION

The zebrafish is a well described model to study infectious diseases, and often applied to study tuberculosis (TB) infection (2, 5, 80-82). To study mycobacterial infection in the zebrafish model, the bacterium *M. marinum* (a natural host of zebrafish) is used (81-84). Genetically, *M. marinum* is closely related to *M. tuberculosis* (85, 86), and both have the ability to replicate inside macrophages and induce the formation of granulomas, the hallmark pathological structures of TB (84, 87). The zebrafish mycobacterial infection model, using *M. marinum*, mimics many aspects of human tuberculosis infection (82, 84), and the use of this model has led to new insights into mycobacterial pathogenesis relevant to the human TB disease (88-90). With only 16 publications in 2016 (PubMed search November 2017), studies on the zebrafish TB infection model using *M. marinum* infections barely touched the field of proteomics. Nevertheless, proteomics seems a powerful tool to answer important remaining immune questions in relation to TB infection (17, 91, 92).

Injecting fluorescent labelled pathogens, e.g. *M. marinum*, into the yolk and vein of zebrafish embryos is a well described technique, resulting in easy detectable bacterial growth which eventually leads to the formation of aggregates of infected immune cells, representing the early stages of the granulomas characteristic for TB (39). Unfortunately, when applied to proteomics using the shotgun bottom-up technique, this did not result in the detection of proteins or protein groups that allowed to distinguish the infected from the non-infected zebrafish embryos significantly (S.J. van der Plas - Duivesteljn, personal communication). In addition, attempting to support a multiple omics approach, adult zebrafish infected with Mma20 and E11 *M. marinum* strains from a previously well published transcriptomics infection study (93), were examined on protein expression in a shotgun bottom-up approach (S.J. van der Plas - Duivesteljn, personal communication). Only 13 proteins were found to be significantly different between the infected and non-infected adult zebrafish in all three replicas. An additional 22 proteins were found to be significantly different between non-infected and infected individuals, but this significant difference could be found in only two out of three replicas, and therefore were marked as speculative results (S.J. van der Plas - Duivesteljn, personal communication). In comparison with previous studies where many significant differences were found on the genomic and transcriptomic level, the number of significant proteins found in this proteomics study was unexpectedly low (S.J. van der Plas - Duivesteljn, personal communication). The resulting low numbers of significantly different proteins found in this adult infection study, and the lack of a possible distinction between infected and non-infected individuals in the study with embryos, could be due to the low number of replicas used to study the proteomics effect. In contrast, the same amount of replicas used in the adult infection setup did result in sufficient genomic results (93), and the infected zebrafish embryos showed severe infection after fluorescent microscopic examination (S.J. van der Plas - Duivesteljn, personal communication). In addition, the lack of a validated protein extraction kit could also have affected these results. In particular, high levels of vitellogenins were found in the embryo infection experiment (S.J. van der Plas - Duivesteljn, personal communication), which can affect protein

detection of lower abundant proteins in the same detection range, emphasizing the need for a more efficient deyolking protocol. Finally, due to the time-consuming data analysis, we experienced the need for the development of software tools assisting in (more) automated screening of the data.

Several genomics studies include additional studies on cellular level of the zebrafish animal model. We therefore also applied a standard bottom-up proteomics approach on FACS sorted cells from dissociated zebrafish embryos. Dissociating immune cell populations from zebrafish using FACS has also been described in a previous study on RNA sequencing (94). Using FACS enables deeper proteome coverage and potentially simultaneous detection of zebrafish and bacterial proteins, as these can be easily distinguished, even on the tryptic peptide level. However, FACS is a slow process and requires additional expertise and equipment, and is not yet easily applicable in high-throughput workflows.

CONCLUSION

The zebrafish has been established as animal model in genomics and transcriptomics and now also emerges in proteomic studies. Since the zebrafish is used as a powerful model within biomedical research, there is a need for the development of HTS methodology. Different zebrafish strains and cell lines have been developed to meet biomedical hypotheses. The use of zebrafish embryos show great potential when applied to HTS. Harmonization of ethical legislation between countries would benefit the use of the zebrafish animal model and assist in faster development of HTS methodology.

From a practical point of view, key technology has been developed to increase the throughput specifically in experiments using zebrafish embryo and larval models. Examples include the ISpawn system for breeding and a fully automated injection system to automate embryonic infection (23). For proteomics applications, there is a need for the development of an automated deyolking protocol, a commercially available protein extraction kit and standardized protocols for homogenization. In addition, robust and automated analysis platforms are essential when applying proteomic HTS on the zebrafish or other animal models.

Besides the development of new methodology, research communities should investigate the possibility of applying complementary species in their studies of animal models near the boundaries of their proteomics design. Rather than developing entirely new online resources, existing ones could be expanded to handle additional model systems to avoid oceans of available resources in which researchers drown when searching for information.

Altogether, lots of work needs to be done to accomplish a successful proteomics HTS using zebrafish as animal model. However, developed techniques and procedures will also be beneficiary for other studies applying proteomics. With such a successful model system as the zebrafish and proteomics as an emerging, valuable and promising field of

research, we have no doubt above suggested inventions will be developed in the near future.

ACKNOWLEDGEMENTS

The authors wish to thank the members of the zebrafish facility team of the Institute of Biology for fish caretaking, and Dmitrii Travin and Iaroslav Popov for technical assistance with the COMICS software. SJPD and MP were supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 917.11.398 (MP).

REFERENCES

1. Streisinger, G., Walker, C., Dower, N., Knauber, D. & Singer, F. Production of clones of homozygous diploid zebra fish (*Brachydanio rerio*). *Nature* 291, 293-296 (1981).
2. Lieschke, G.J. & Currie, P.D. Animal models of human disease: zebrafish swim into view. *Nat Rev Genet* 8, 353-367 (2007).
3. Brittijn, S.A. et al. Zebrafish development and regeneration: new tools for biomedical research. *International Journal of Developmental Biology* 53, 835-850 (2009).
4. Ali, S., Champagne, D.L., Spaink, H.P. & Richardson, M.K. Zebrafish embryos and larvae: a new generation of disease models and drug screens. *Birth defects research. Part C, Embryo today : reviews* 93, 115-133 (2011).
5. Meijer, A.H. & Spaink, H.P. Host-pathogen interactions made transparent with the zebrafish model. *Curr. Drug Targets*. 12, 1000-1017 (2011).
6. Kari, G., Rodeck, U. & Dicker, A.P. Zebrafish: An Emerging Model System for Human Disease and Drug Discovery. *Clin Pharmacol Ther* 82, 70-80 (2007).
7. Zebrafish International Resource Center (ZIRC), University of Oregon, USA, (<http://zebrafish.org/home/guide.php>)
8. Howe, K. et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature advance online publication* (2013).
9. Reimers, M.J., Hahn, M.E. & Tanguay, R.L. Two zebrafish alcohol dehydrogenases share common ancestry with mammalian class I, II, IV, and V alcohol dehydrogenase genes but have distinct functional characteristics. *J Biol Chem* 279, 38303-38312 (2004).
10. Renier, C. et al. Genomic and functional conservation of sedative-hypnotic targets in the zebrafish. *Pharmacogenetics and genomics* 17, 237-253 (2007).
11. Li, C., Tan, X.F., Lim, T.K., Lin, Q. & Gong, Z. Comprehensive and quantitative proteomic analyses of zebrafish plasma reveals conserved protein profiles between genders and between zebrafish and human. *Scientific Reports* 6, 24329 (2016).
12. Meijer, A.H. et al. Expression analysis of the Toll-like receptor and TIR domain adaptor families of zebrafish. *Molecular immunology* 40, 773-783 (2004).
13. Stein, C., Caccamo, M., Laird, G. & Leptin, M. Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biol* 8, R251 (2007).
14. Traver, D. et al. The zebrafish as a model organism to study development of the immune system. *Adv. Immunol.* 81, 253-330 (2003).

15. Trede, N.S., Langenau, D.M., Traver, D., Look, A.T. & Zon, L.I. The use of zebrafish to understand immunity. *Immunity* 20, 367-379 (2004).
16. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology* 18, 83 (2017).
17. Overall, C.M. Can proteomics fill the gap between genomics and phenotypes? *Journal of Proteomics* 100, 1-2 (2014).
18. List, E.O. et al. The use of proteomics to study infectious diseases. *Infectious disorders drug targets* 8, 31-45 (2008).
19. Plas-Duivestijn, S.J. et al. Identifying proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *J Proteome Res* 13 (2014).
20. ZFIN. (<https://zfin.org/action/genotype/view/ZDB-GENO-960809-7>) (2016).
21. Balik-Meisner, M., Truong, L., Scholl, E.H., Tanguay, R.L. & Reif, D.M. Population genetic diversity in zebrafish lines. *Mammalian genome : official journal of the International Mammalian Genome Society* (2018).
22. Guryev, V. et al. Genetic variation in the zebrafish. *Genome Research* 16, 491-497 (2006).
23. Carvalho, R. et al. A high-throughput screen for tuberculosis progression. *PLoS one* 6, e16779 (2011).
24. Veneman, W.J. et al. A zebrafish high throughput screening system used for *Staphylococcus epidermidis* infection marker discovery. *BMC Genomics* 14, 255 (2013).
25. Pardo-Martin, C. et al. High-throughput in vivo vertebrate screening. *Nat Meth* 7, 634-636 (2010).
26. Lessman, C.A. The developing zebrafish (*Danio rerio*): a vertebrate model for high-throughput screening of chemical libraries. *Birth defects research. Part C, Embryo today : reviews* 93, 268-280 (2011).
27. Letamendia, A. et al. Development and validation of an automated high-throughput system for zebrafish in vivo screenings. *PLoS one* 7, e36690 (2012).
28. Spaink, H.P. et al. Robotic injection of zebrafish embryos for high-throughput screening in disease models. *Methods* 62, 246-254 (2013).
29. Miscevic, F., Rotstein, O. & Wen, X.Y. Advances in zebrafish high content and high throughput technologies. *Comb Chem High Throughput Screen* 15, 515-521 (2012).
30. Cui, C. et al. Infectious disease modeling and innate immune function in zebrafish embryos. *Methods Cell Biol.* 105, 273-308 (2011).
31. Lam, S.H., Chua, H.L., Gong, Z., Lam, T.J. & Sin, Y.M. Development and maturation of the immune system in zebrafish, *Danio rerio*: a gene expression profiling, in situ hybridization and immunological study. *Dev.Comp Immunol.* 28, 9-28 (2004).
32. White, R.M. et al. Transparent adult zebrafish as a tool for in vivo transplantation analysis. *Cell stem cell* 2, 183-189 (2008).
33. ZIRC zebrafish cell lines. (<http://zebrafish.org/fish/lineAll.php>) (2018)
34. Simonetti, R.B.M., L.S.; Streit, D.P.; Oberst E.R. Zebrafish (*Danio rerio*): Ethics in animal experimentation. *IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS)* 9, 106-110 (2016).
35. Lidster, K., Readman, G.D., Prescott, M.J. & Owen, S.F. International survey on the use and welfare of zebrafish *Danio rerio* in research. *Journal of Fish Biology* 90, 1891-1905 (2017).

36. Strahle, U. et al. Zebrafish embryos as an alternative to animal experiments—a commentary on the definition of the onset of protected life stages in animal welfare regulations. *Reproductive toxicology* (Elmsford, N.Y.) 33, 128-132 (2012).
37. Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B. & Schilling, T.F. Stages of embryonic development of the zebrafish. *Developmental dynamics : an official publication of the American Association of Anatomists* 203, 253-310 (1995).
38. Adatto, I., Lawrence, C., Thompson, M. & Zon, L.I. A new system for the rapid collection of large numbers of developmentally staged zebrafish embryos. *PLoS one* 6, e21715 (2011).
39. Benard, E.L. et al. Infection of zebrafish embryos with intracellular bacterial pathogens. *J.Vis.Exp.* (2012).
40. Henry, K.M., Loynes, C.A., Whyte, M.K. & Renshaw, S.A. Zebrafish as a model for the study of neutrophil biology. *Journal of leukocyte biology* 94, 633-642 (2013).
41. Takaki, K., Cosma, C.L., Troll, M.A. & Ramakrishnan, L. An in vivo platform for rapid high-throughput antitubercular drug discovery. *Cell reports* 2, 175-184 (2012).
42. Westerfield, M. *The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio)*. Edn. 4. (Univ. of Oregon Press, Eugene; 2000).
43. Link, V., Shevchenko, A. & Heisenberg, C.P. Proteomics of early zebrafish embryos. *BMC.Dev.Biol.* 6, 1 (2006).
44. Mandrell, D. et al. Automated zebrafish chorion removal and single embryo placement: optimizing throughput of zebrafish developmental toxicity screens. *Journal of laboratory automation* 17, 66-74 (2012).
45. Choudhury, D. et al. Fish and Chips: a microfluidic perfusion platform for monitoring zebrafish development. *Lab on a chip* 12, 892-900 (2012).
46. Zheng, C. et al. Fish in chips: an automated microfluidic device to study drug dynamics in vivo using zebrafish embryos. *Chemical communications (Cambridge, England)* 50, 981-984 (2014).
47. Wang, H., Tan, J.T.T., Emelyanov, A., Korzh, V. & Gong, Z. Hepatic and extrahepatic expression of vitellogenin genes in the zebrafish, *Danio rerio*. *Gene* 356, 91-100 (2005).
48. Lin, Y., Chen, Y., Yang, X., Xu, D. & Liang, S. Proteome analysis of a single zebrafish embryo using three different digestion strategies coupled with liquid chromatography-tandem mass spectrometry. *Anal.Biochem.* 394, 177-185 (2009).
49. Rahlouni, F., Szarka, S., Shulaev, V. & Prokai, L. A Survey of the Impact of Deyolking on Biological Processes Covered by Shotgun Proteomic Analyses of Zebrafish Embryos. *Zebrafish* 12, 398-407 (2015).
50. IgY RuBisCO columns, (<https://www.sigmaaldrich.com/technical-documents/articles/biology/seppro-rubisco.html>) (2018)
51. Irmeler, I., Schmidt, K. & Starck, J.M. Developmental variability during early embryonic development of zebra fish, *Danio rerio*. *Journal of experimental zoology. Part B, Molecular and developmental evolution* 302, 446-457 (2004).
52. Parichy, D.M., Elizondo, M.R., Mills, M.G., Gordon, T.N. & Engeszer, R.E. Normal table of postembryonic zebrafish development: staging by externally visible anatomy of the living fish. *Developmental dynamics : an official publication of the American Association of Anatomists* 238, 2975-3015 (2009).
53. Bennike, T.B. et al. Comparing the proteome of snap frozen, RNAlater preserved, and formalin-fixed paraffin-embedded human tissue samples. *EuPA Open Proteomics* 10, 9-18 (2016).

54. Baghirova, S., Hughes, B.G., Hendzel, M.J. & Schulz, R. Sequential fractionation and isolation of subcellular proteins from tissue or cultured cells. *MethodsX* 2, 440-445 (2015).
55. Havanapan, P.O. & Thongboonkerd, V. Are protease inhibitors required for gel-based proteomics of kidney and urine? *J Proteome Res* 8, 3109-3117 (2009).
56. Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.-C. & Yates, J.R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical Reviews* 113, 2343-2394 (2013).
57. Victor, B. et al. Partially sequenced organisms, decoy searches and false discovery rates. *J. Proteome Res.* 11 (2012).
58. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*. 7, 655-667 (2007).
59. Lam, H. et al. Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* 5, 873-875 (2008).
60. Deutsch, E.W. et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 10, 1150-1159 (2010).
61. Ebrahimie, E., Fruzangohar, M., Moussavi Nik, S.H. & Newman, M. Gene Ontology-Based Analysis of Zebrafish Omics Data Using the Web Tool Comparative Gene Ontology. *Zebrafish* 14, 492-494 (2017).
62. Hedges, S.B. The origin and evolution of model organisms. *Nat Rev Genet* 3, 838-849 (2002).
63. Zuckerkandl, E. The Evolution of Hemoglobin. *Sci Am* 212 (1965).
64. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J Theor Biol* 8 (1965).
65. Welker, F. et al. Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature* 522, 81-84 (2015).
66. Procopio, N., Chamberlain, A.T. & Buckley, M. Intra- and Interskeletal Proteome Variations in Fresh and Buried Bones. *J Proteome Res* 16, 2016-2029 (2017).
67. Jiang, X. et al. Method development of efficient protein extraction in bone tissue for proteome analysis. *J Proteome Res* 6, 2287-2294 (2007).
68. Buckley, M., Collins, M., Thomas-Oates, J. & Wilson, J.C. Species Identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* 23 (2009).
69. Orlando, L. et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* 499, 74-78 (2013).
70. Sigg, M.A. et al. Evolutionary Proteomics Uncovers Ancient Associations of Cilia with Signaling Pathways. *Developmental cell* 43, 744-762.e711 (2017).
71. Palmblad, M. & Deelder, A.M. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun. Mass Spectrom.* 26, 728-732 (2012).
72. ATCC FISH CELL LINES, (https://www.lgcstandards-atcc.org/Products/Cells_and_Microorganisms/Cell_Lines/Animal/Fish.aspx?geo_country=nl) (2017).
73. Kolder, I.C.R.M. et al. A full-body transcriptome and proteome resource for the European common carp. *BMC Genomics* 17, 701 (2016).
74. Smith, S.A. Zebrafish resources on the Internet. *ILAR Journal* 53, 208-214 (2012).
75. The UniProt Consortium UniProt: the universal protein knowledgebase. *Nucleic Acids Research* 45, D158-D169 (2017).
76. Deutsch, E.W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* 9, 429-434 (2008).
77. Peptide atlas: zebrafish (https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=408) (2018)

78. NIST zebrafish database. (<http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:zebrafish20150109>) (2018).
79. Travin, D. et al. COMICS: Cartoon Visualization of Omics Data in Spatial Context Using Anatomical Ontologies. *J Proteome Res* (2017).
80. Davis, J.M. et al. Real-time visualization of mycobacterium-macrophage interactions leading to initiation of granuloma formation in zebrafish embryos. *Immunity* 17, 693-702 (2002).
81. Lesley, R. & Ramakrishnan, L. Insights into early mycobacterial pathogenesis from the zebrafish. *Current opinion in microbiology* 11, 277-283 (2008).
82. Swaim, L.E. et al. Mycobacterium marinum infection of adult zebrafish causes caseating granulomatous tuberculosis and is moderated by adaptive immunity. *Infect.Immun.* 74, 6108-6117 (2006).
83. Clay, H. et al. Dichotomous role of the macrophage in early Mycobacterium marinum infection of the zebrafish. *Cell host & microbe* 2, 29-39 (2007).
84. Tobin, D.M. & Ramakrishnan, L. Comparative pathogenesis of Mycobacterium marinum and Mycobacterium tuberculosis. *Cellular microbiology* 10, 1027-1039 (2008).
85. Tonjum, T., Welty, D.B., Jantzen, E. & Small, P.L. Differentiation of Mycobacterium ulcerans, M. marinum, and M. haemophilum: mapping of their relationships to M. tuberculosis by fatty acid profile analysis, DNA-DNA hybridization, and 16S rRNA gene sequence analysis. *Journal of clinical microbiology* 36, 918-925 (1998).
86. Stinear, T.P. et al. Insights from the complete genome sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis. *Genome Res* 18, 729-741 (2008).
87. van der Vaart, M., Spaink, H.P. & Meijer, A.H. Pathogen recognition and activation of the innate immune response in zebrafish. *Advances in hematology* 2012, 159807 (2012).
88. Cronan, M.R. & Tobin, D.M. Fit for consumption: zebrafish as a model for tuberculosis. *Disease models & mechanisms* 7, 777-784 (2014).
89. Volkman, H.E. et al. Tuberculous granuloma induction via interaction of a bacterial secreted protein with host epithelium. *Science* 327, 466-469 (2010).
90. Volkman, H.E. et al. Tuberculous granuloma formation is enhanced by a mycobacterium virulence determinant. *PLoS Biol* 2, e367 (2004).
91. Verhoeckx, K.C.M. et al. A combination of proteomics, principal component analysis and transcriptomics is a powerful tool for the identification of biomarkers for macrophage maturation in the U937 cell line. *PROTEOMICS* 4, 1014-1028 (2004).
92. Hecker, M., Antelmann, H., Buttner, K. & Bernhardt, J. Gel-based proteomics of Gram-positive bacteria: a powerful tool to address physiological questions. *Proteomics* 8, 4958-4975 (2008).
93. van der Sar, A.M., Spaink, H.P., Zakrzewska, A., Bitter, W. & Meijer, A.H. Specificity of the zebrafish host transcriptome response to acute and chronic mycobacterial infection and the role of innate and adaptive immune components. *Mol. Immunol.* 46, 2317-2332 (2009).
94. Rougeot, J. et al. RNA Sequencing of FACS-Sorted Immune Cell Populations from Zebrafish Infection Models to Identify Cell Specific Responses to Intracellular Pathogens. *Methods Mol Biol* 1197, 261-274 (2014).



Chapter 6

Discussion and Summary

In this thesis, the zebrafish animal model is introduced to proteomics by developing new methodology and a number of resources, including a spectral library and data analysis workflows, with the goal to enable application of zebrafish models in biomedical studies using a proteomics design.

PROTEOMICS

A major goal of proteomics is the identification of all proteins expressed in a cell or tissue (1). All proteomics experiments for this thesis involve the so called shotgun bottom-up approach, where we attempted to identify as much proteins as possible. We show the applicability of more standardized proteomics techniques to an animal model for understanding the nature of disease for biomedical purposes. High throughput (HT) is of great value in biomedical research (2). By making use of the proteomics shotgun approach, a relatively HT can be achieved (3) compared to other MS-based proteomics technologies (such as 2D gel protein analysis).

Protein Identification

Library searching has become more important in biomedical research applying a proteomics design, due to its advantages in higher speed, improved sensitivity and lower error rates compared to other forms of spectral matching (4, 5). Since a higher search speed contributes to HTS, it is a valuable tool to apply in experimental designs. Spectral libraries of some model organisms (*E.coli*, *C. elegans*, *M. musculus*, and *R. norvegicus*) were already available at the start of this PhD study. However, no spectral library from zebrafish (*Danio rerio*) or any closely related organism was available. We therefore built the first zebrafish spectral library to assist the applicability of HTS for this animal model when applied in proteomic studies (**chapter 2**). Besides being the first zebrafish library created, it is the first library created using dissection of adult organs and tissues to cover as much of the proteome as possible. As library searching is generally more sensitive than database searches, the zebrafish library described in **chapter 2** contributed to increased sensitivity and coverage in subsequent experiments involving the zebrafish animal model.

To enable the accessibility, visibility, and reusability of the zebrafish library, we collaborated with NIST. Using their own pipeline, NIST reprocessed all our data resulting in a library available as: "the standard zebrafish spectral library" which is accessible via their own webpage, since January 9, 2015 (6). Scientists applying the zebrafish animal model, freely exchange knowledge and data via ZFIN and ZIRC, which are sister organizations (7, 8). In addition to its release on the NIST webpage, the scientific community was informed about our zebrafish library via a notification on the ZFIN webpage on April 20, 2015.

Manually creating spectral libraries is time consuming due to the amount of data and the number of steps involved. We therefore built a scientific workflow in Taverna and used this to create the zebrafish library (**chapter 2**). This workflow has been made freely available to the scientific community for building similar libraries of other organisms.

Automation, and so simplifying, of the library building process will increase its development and will contribute to a HT when researchers use this proteomics tool.

Though not emphasized in the publication, the created zebrafish library (**chapter 2**) is also unique as it contains spatial information from the dissection organs, namely the expression patterns as function of anatomy (as well as developmental stage from the embryo data). By separating the proteins using SDS-PAGE prior to tryptic digestion and LC/MS/MS analysis, the library also contains information on the apparent molecular weight and isoforms. In other words, the library and its by-products represent a resource that can be mined for additional information, such as organ-specific protein expression and post-translational events.

Quantitative information can always be looked at on the peptide (spectrum) level (9) and, depending on the analysis pipeline, on the level of unique proteoforms or groups of proteins sharing a set of identified peptides. These protein groups typically correspond to a protein (or gene) family. While using only ion trap tandem mass spectrometry data and peptide-spectrum matches, we also collated protein groups into super – protein groupings across all samples in an experiment (**chapter 5**; zebrafish adult infection). Such summation will provide differentially abundant proteins, but may obscure smaller changes in individual proteins. At the end of the day, it should always be kept in mind that the individual units of qualitative as well as quantitative information are at the peptide level in bottom-up proteomics.

Proteomic software tools

For automated and HT analysis of mass spectrometry data, efficient software tools are a necessity. While many tools are already available, we developed, applied or repurposed tools for our specific needs. For example, we repurposed CompareMS2, a software tool originally developed for molecular phylogenetics (10), in the experimental design for generating the zebrafish spectral library by comparing male and female organs and tissues to prioritize those that would add the most information to library creation (**chapter 2**). **Chapter 3** describes some of the many possible applications of CompareMS2 in a wide range of experimental designs. Unlike other tools, including the recent MICId (11), CompareMS2 does not use genomic information for microbial (or other species) identification or classification. This makes CompareMS2 less complicated to use and directly applicable in different contexts, and is independent from other information. For example CompareMS2 not even needs the information that the spectra derive from peptides. CompareMS2 quickly gives information on similarities across proteomic databases which can always be used as a quick first analysis step and therefore also contributes to a higher throughput. For future perspectives, CompareMS2 could be implemented in automated analysis platforms to study its applicability in HT data analysis.

The introduction of this thesis (**chapter 1**) started with the statement that a multidisciplinary approach, where “omics” fields are combined, complement each other in favour of addressing challenges in biomedical research. The efficiency of such a

multidisciplinary approach is highlighted in several studies applied to both human and the zebrafish animal model (3, 12). **Chapter 4** supports this multidisciplinary approach, by describing the correlation between uniquely mapped peptides and RNA-seq of carp organs, resulting in a valuable prediction for its use as marker genes to be applied in future comparative genomics. In **chapter 4** we also highlight that in-depth analysis of mRNA-protein results is desired for future studies, which also includes the development of additional software tools to realize such a study. In **chapter 5** we highlight a multidisciplinary approach in which we expanded the results of a microarray study with an additional proteomics study. Both **chapters, 4 and 5**, emphasize the need to develop automated software tools that assist in automated and multidisciplinary data analysis which will result in HT and improved applicability.

As mentioned before, the zebrafish spectral library (**chapter 2**), contains valuable data of organ specific peptides and proteins. In **chapter 4** a beta-version of an Integrative data viewer was described, in which an R script is used to map gene and protein expression to individual organs of the carp body plan. The resulting developed software tool COMICS describes the actual integrative data viewer mentioned in **chapter 4** and was built using the work described in **chapter 2 and 4** of this thesis (13). COMICS is applicable to data available from the zebrafish spectral library (**chapter 2**), which enables mapping of organ specific protein expression. Information achieved using this software tool, which combines protein expression (zebrafish spectral library; **chapter 2**) and gene expression data, can contribute to a better understanding of tissue/organ related expression and the development of possible markers for future research purposes. Additionally, this tool allows researchers to re-use available published datasets, preventing the sacrifice of animals for research purposes, and so contributes to the reduction of the use of animals.

Experimental protocols

Besides resources like spectral libraries and software tools, we also experienced the need for several experimental protocols, which we highlighted in **chapter 5**. Where for genomic studies, DNA and RNA extraction kits are commercially available, also for proteomics, protein extraction kits are available. The ReadyPrep protein extraction kit (BIORAD) allows total protein extraction from cellular material, but does not facilitate whole organ or embryo extraction. ProteoExtract® Kits (Merck) offers kits for subcellular, membrane, transmembrane and protein extraction, which are collected in fractions, and therefore does not allow a HT design. The Total Protein Extraction kit (Merck), which promises to purificate proteins from 5 grams of any tissue and cell, shows high potential but has not been tested on zebrafish yet. Another disadvantage of the usage of these kits is that they are relatively expensive compared to buffers produced in your own laboratory.

Protein extraction buffers that have been described in literature result in extraction of proteins from different origin, e.g. cell membrane proteins, cytoplasmic proteins and nuclear proteins (**chapter 5**), resulting in a multiple proteomics approach to address a full cover of all proteins, which does not support HT processing (14, 15). To enable total protein coverage implemented in HT processing on zebrafish (tissue), available total

protein kits should be validated for HTS and their applicability to this animal model. Furthermore, additional literature describing and validating standardized extraction protocols will be necessary. Moreover, the constant development of new kits that can be applied to proteomics, like the Rapid Digestion–Trypsin (Promega), which shortens trypsin digestion from 18 hours to 60 minutes (16), will contribute to reach the HT status for proteomics studies.

Instrumentation

The proteomics effect of zebrafish infection (**chapter 5**) was studied using a state-of-the-art FTICR-ion trap cluster, using the ion traps for peptide identification and the FTICR for label-free quantitation (17). We used the FTICR, in combination with microscale liquid chromatography, as it is a very stable detection system with high mass measurement accuracy and reproducible signal intensities across large sample studies (9). This setup meets with high throughput screening (HTS) designs which are required in biomedical research (18). The system was in operation at the start of this project, and the necessary algorithms for label-free quantitation were developed, including alignment and “match between runs” (to prevent missing value problems). The liquid chromatography LC systems used in the studies described in this thesis, require less frequent maintenance than for example nano-LC systems, making them more suitable for HTS. Above described instrumentation combined in a LC/MS/MS setup created a starting point for experimental design which meets the requirements for HTS.

When making use of liquid chromatography (LC) to separate complex peptide mixtures, one can vary many parameters, e.g. the separation time of a sample, washing time of columns and syringes, type of running buffers, etc. In addition, the setting of the mass spectrometer itself, set to detect LC separated peptides, allows many variable settings as well. While LC and mass spectrometry are well known techniques within proteomics, varying parametric settings within an experimental design is not well described yet. Future research to standardization of parameters and settings of the used instrumentation will contribute to multidisciplinary approaches, reproducibility and HTS.

ZEBRAFISH

(Zebra)fish models within proteomics

The zebrafish is described as a widely used animal model for biomedical research (19–24). There are many questions to be answered within this field, requiring different approaches and models. Zebrafish cannot replace other animal models, since every animal model has its own advantages and disadvantages (19, 25), but they can serve as an acceptable addition to unravel scientific questions. We therefore would like to emphasize that the zebrafish is and never will be *the* perfect animal model, but *an* animal model with great potential for biomedical research. In addition, the combination of using multiple models will improve the understanding of disease pathogenesis in biomedical research (19). Compared to human anatomy, zebrafish lack some organs which reflects their evolutionary distance. This disadvantage underlines the carefulness in translation

to humans. But when applying an animal model one should consider its advantages, e.g. for the zebrafish animal model: the similarity of most organ systems and gene sequences to humans, its transparency at the early life stages, fast development, and cost reduction (when using embryos for experiments), and in addition the possibility to apply automated/high-throughput screens in drug toxicity and testing (23, 24, 26-36).

Within this thesis we primarily made use of wild type zebrafish (AB line), a genetically heterogeneous population. Genetically heterogeneous populations can cause large standard deviations and variations in most studies. On the opposite, heterogeneity reflects the natural situation in humans, with all its differences between individuals (37). By carefully standardizing experimental conditions (35) any effects other than heterogeneity should be minimized.

Zebrafish embryos

When making use of the embryo and/or larval stages of the zebrafish model within a proteomics design, it is advised to remove the yolk (a procedure called *deyolking*) before starting protein extraction (38, 39). Deyolking is a technique to remove the most abundant protein (vitellogenin) from a zebrafish embryo or larva (1-4 days post fertilization), to enable the detection of less abundant proteins in the same range (molecular weight in kD) as vitellogenin. We therefore applied the deyolking method on 1-4 dpf embryos for creating the zebrafish spectral library (**chapter 2**), and to study the effect of mycobacterium infection (**chapter 5**). In 2015, the impact of deyolking was described, showing that deyolking depleted molecular pathways associated with calcium signaling and immune response signaling (40). It was therefore advised to use both yolk-intact and deyolged embryos to maximize protein coverage for shotgun proteomics studies (40). According to our own experience with deyolking, a technique which is time consuming and has a low accuracy, we would advise to invest in future studies on automated and faster methods, for example by the development of columns that can withdraw vitellogenin proteins.

Zebrafish cells

Protocols have been developed to extract specific cells (like macrophages) from zebrafish embryos by using fluorescent labeled zebrafish lines, dissociate cells from zebrafish embryos, and apply these to FACS (41). Extracting the specific cells (e.g. macrophages or other immune cells) from zebrafish embryos is a time consuming technique (taking a few working days from breeding to FACS) which does not meet HT status (**chapter 5**). An entire batch of embryos from one lay is required, resulting in one or two tubes with cells (**chapter 5**). A solution for the limitation caused by the required amount of embryos can be taken from **chapter 4**, in which we suggest to use one animal model to solve the limitations of another. In other words, use carp eggs instead, since carp produce many more eggs in one lay. Cultured cells could also be an option, but those are lacking information of immune interaction, which will be available when making use of dissociated cells.

Proteomic “phylogenesis”

In **chapter 1**, we mentioned we were interested in whether spectra can be used across (model) species when identifying peptides using spectral libraries, if no spectral library from the species of interest is available. The applicability of spectral libraries between animal models within proteomics can be explained by conserved peptides. For example, AIAELGIYPAVDPLDSTSR, a peptide from mitochondrial ATP synthase subunit beta, is identical between human, *D. rerio*, *D. melanogaster*, and even *C. elegans*. The tandem mass spectra of this peptide in both *D. rerio* and *C. carpio*, two close relatives, shows similar spectral patterns, though the signal-to-background is better in one than in the other (figure 1). The conservation of peptides between species indicates that at least some tandem mass spectra of tryptic peptides can be shared across many model systems. **Chapter 4** demonstrates the possibility of bridging the limitations of the animal models carp and zebrafish by complementing these close related model species, based on their highly conserved peptides. For example, the spectral library of zebrafish (**chapter 2**) can be used to complement the lack of a spectral library for carp. And the advantage of the large body size of the carp might be used to bridge the limitations of the zebrafish model due to its small organ size and limited amount of cells. In summary, **chapter 4** demonstrates that spectra can be used across (model) species, which shows high potency for other animal models, like rats and mice. Future research on shared peptides should demonstrate additional applications, for example the possibility shared use of stable isotope labeled standards across species could be studied, to eventually reduce species specific synthesis and validation of such peptides.

Even though shared tryptic peptides enable the translation within closely related species, at the same time unique tryptic peptides are necessary to distinguish one from the other, for example in distinguishing bacterial proteins for species determination (figure 2 of **chapter 1**, and **chapter 3**). Or to distinguish human proteins from xenotransplants for example, to predict possible interferences between the two proteomes. And in addition, to complete traditional classification methods, using so called paleoproteomics (S.J. van der Plas – Duivesteljn, personal communication). Such proteomic “phylogenesis” can also be calculated *in silico*. Since *in silico* experiments can be used to predict and explore an experimental design within the field of proteomics (42, 43), future research to both shared and unique tryptic peptides on many different levels (species, organ, cellular level) can be of great importance in creating cost effective experimental designs. We would like to stress the importance of using phylogenetic and “protein phylogeny” information when applying animal models.

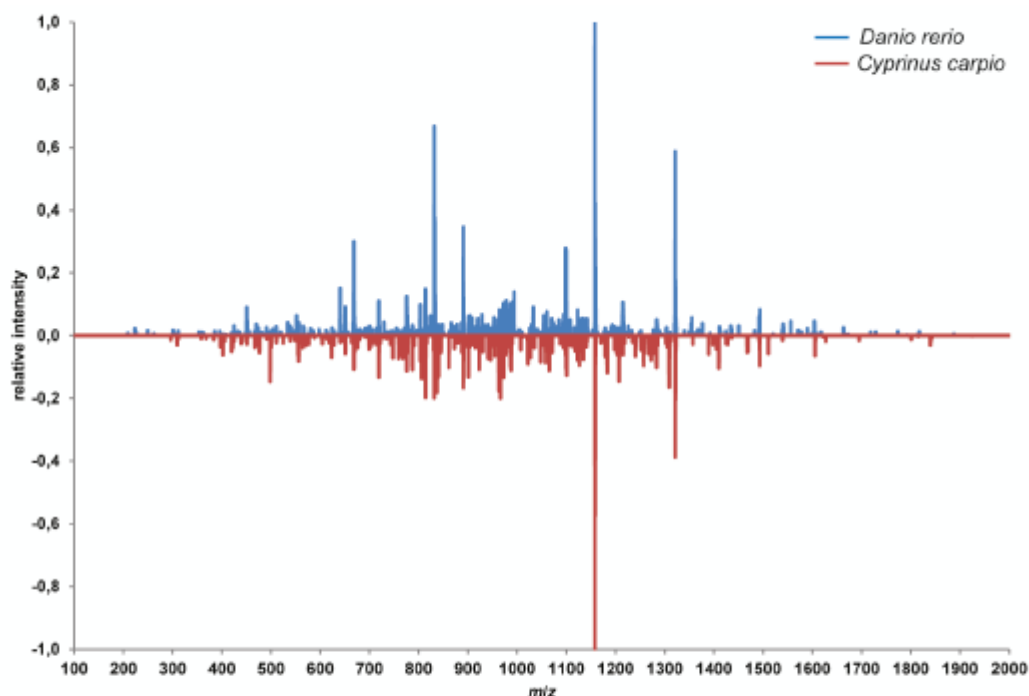


Figure 1: Matching tandem mass spectra from zebrafish (top) and carp (bottom) acquired on an amaZon ion trap mass spectrometer. Both spectra were confidently identified as a peptide AIAELGIYPVDPLDSTSR from mitochondrial ATP synthase subunit beta. The absolute signal in this particular carp spectrum is a factor ~157 lower than the zebrafish spectrum. They could nevertheless be confidently matched (SpectraST dot product 0.797 and PeptideProphet probability 0.9976), demonstrating the possibility of shared use of available libraries between both species.

TB infection model applied to proteomics

A secondary goal of this thesis was to explore use of the zebrafish *M. marinum* infection model in a proteomics design. When studying *M. marinum* infection in zebrafish adult and embryos in a standard bottom-up approach (mentioned in **chapter 5**) it has become clear which inventions are necessary to enable such studies. While development of these inventions is in progress, infected zebrafish embryos could be studied using other mass spectrometry techniques. For example, applying mass spectrometry imaging (44, 45) on histological sections of *M. marinum* infected zebrafish adults or embryos might allow histological annotation of peptides and proteins to discover biomarker candidates. Even though this technique will probably not result in HTS, it will add valuable information to TB research. In addition it would be very interesting to investigate if *in silico* mediated experiments could assist in exploring experimental designs for such studies, saving time consuming preliminary studies and prevent sacrificing animals.

CONCLUDING REMARKS

The work described in this thesis was part of a study which aims to research new tools to study infectious diseases (e.g. TB) and its application in biomedical studies. The research goal assigned to this thesis was described as: introduce the zebrafish TB infection model to proteomics. The introduction of zebrafish to proteomics in general is set as a first step to reach this goal and is described within this thesis. Future work within this project will include the introduction of the zebrafish as model to study TB infection, combined with automated designs for this animal model.

The work presented in this thesis provided the scientific community with the first zebrafish spectral library, which is an important first step to introduce the zebrafish to proteomics research. The results of this study can be used in additional studies like to discover new biomarkers (**chapter 2**).

Proteomics studies designed to research biological hypotheses which make use of open source software, requires more user friendly software tools than available at present. Also there is a high necessity for the development of additional software tools that facilitate the field of proteomics. Development of these resources will lead to improved or automated workflows that enables HTS, and gain a HT in multidisciplinary studies. This thesis provides such resources, namely software to automate spectral library building to simplify the creation of spectral libraries (**chapter 2**), and Compare MS2 which can be used to provide information on similarities across proteomics databases. Compare MS2 was shown to be a resource as quick first analysis step which can be applied to a wide range of experimental designs (**chapter 3**).

Multidisciplinary approaches, addressing multiple omics, will expand within future biomedical research. The work described in **chapter 4** demonstrated the advantages and applicability of such an approach, exemplified by using two closely related animal models (common carp and zebrafish). The start of the build of an integrative data viewer, which was the foundation for the development of a user friendly resource tool, is a perfect example of a multidisciplinary approach. In addition, an overview is given on advantages and disadvantages of the zebrafish animal model within proteomics (**chapter 5**).

Our practical approach to zebrafish models in proteomics described in this thesis has resulted in experimental protocols, a standard spectral library and data analysis tools, and significantly contributed to maturing of this model system for proteomics research.

REFERENCES

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* 422, 198-207 (2003).
2. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 4 (2005).
3. Alli Shaik, A. et al. Functional mapping of the zebrafish early embryo proteome and transcriptome. *J Proteome Res* 13, 5536-5550 (2014).
4. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7, 655-667 (2007).
5. Lam, H. et al. Building consensus spectral libraries for peptide identification in proteomics. *Nat.Methods* 5, 873-875 (2008).
6. NIST zebrafish database. <http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:zebrafish20150109>
7. Zfin <https://zfin.org/action/genotype/view/ZDB-GENO-960809-7>
8. Zebrafish International Resource Center (ZIRC), University of Oregon, USA. <http://zebrafish.org/home/guide.php>
9. Johansson, A. et al. Identification of genetic variants influencing the human plasma proteome. *Proc Natl Acad Sci U S A* 110, 4673-4678 (2013).
10. Palmblad, M. & Deelder, A.M. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun.Mass Spectrom.* 26, 728-732 (2012).
11. Alves, G. et al. Identification of Microorganisms by High Resolution Tandem Mass Spectrometry with Accurate Statistical Significance. *Journal of The American Society for Mass Spectrometry* 27, 194-210 (2015).
12. Edfors, F. et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Molecular systems biology* 12 (2016).
13. Travin, D. et al. COMICS: Cartoon Visualization of Omics Data in Spatial Context Using Anatomical Ontologies. *J Proteome Res* (2017).
14. Zhang, C.G., Chromy, B.A. & McCutchen-Maloney, S.L. Host-pathogen interactions: a proteomic view. *Expert Rev Proteomics* 2, 187-202 (2005).
15. Schmidt, F. et al. Complementary analysis of the Mycobacterium tuberculosis proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology. *Mol Cell Proteomics* 3, 24-42 (2004).
16. Kobs, G. Same-Day Mass Spec Sample Prep Is Now a Reality: Shorten Digestion Time to as Little as 60 Minutes. *American Medical Association, Manual of Style* 10 (2018).
17. Palmblad, M., van der Burgt, Y.E., Mostovenko, E., Dalebout, H. & Deelder, A.M. A novel mass spectrometry cluster for high-throughput quantitative proteomics. *J.Am.Soc.Mass Spectrom.* 21, 1002-1011 (2010).
18. Liu, Y. et al. Quantitative variability of 342 plasma proteins in a human twin population. *Molecular systems biology* 11, 786 (2015).
19. Lieschke, G.J. & Currie, P.D. Animal models of human disease: zebrafish swim into view. *Nat Rev Genet* 8, 353-367 (2007).
20. Brittijn, S.A. et al. Zebrafish development and regeneration: new tools for biomedical research. *International Journal of Developmental Biology* 53, 835-850 (2009).
21. Ali, S., Champagne, D.L., Spaink, H.P. & Richardson, M.K. Zebrafish embryos and larvae: a new generation of disease models and drug screens. *Birth defects research. Part C, Embryo today : reviews* 93, 115-133 (2011).

22. Meijer, A.H. & Spalink, H.P. Host-pathogen interactions made transparent with the zebrafish model. *Curr. Drug Targets*. 12, 1000-1017 (2011).
23. Kari, G., Rodeck, U. & Dicker, A.P. Zebrafish: An Emerging Model System for Human Disease and Drug Discovery. *Clin Pharmacol Ther* 82, 70-80 (2007).
24. Love, D.R., Pichler, F.B., Dodd, A., Copp, B.R. & Greenwood, D.R. Technology for high-throughput screens: the present and future using zebrafish. *Current Opinion in Biotechnology* 15, 564-571 (2004).
25. Myllymaki, H., Niskanen, M., Oksanen, K.E. & Ramet, M. Animal models in tuberculosis research - where is the beef? *Expert opinion on drug discovery* 10, 871-883 (2015).
26. Carvalho, R. et al. A high-throughput screen for tuberculosis progression. *PloS one* 6, e16779 (2011).
27. Veneman, W.J. et al. A zebrafish high throughput screening system used for *Staphylococcus epidermidis* infection marker discovery. *BMC Genomics* 14, 255 (2013).
28. Pardo-Martin, C. et al. High-throughput in vivo vertebrate screening. *Nat Meth* 7, 634-636 (2010).
29. Lessman, C.A. The developing zebrafish (*Danio rerio*): a vertebrate model for high-throughput screening of chemical libraries. *Birth defects research. Part C, Embryo today : reviews* 93, 268-280 (2011).
30. Letamendia, A. et al. Development and validation of an automated high-throughput system for zebrafish in vivo screenings. *PloS one* 7, e36690 (2012).
31. Spalink, H.P. et al. Robotic injection of zebrafish embryos for high-throughput screening in disease models. *Methods* 62, 246-254 (2013).
32. Miscevic, F., Rotstein, O. & Wen, X.Y. Advances in zebrafish high content and high throughput technologies. *Comb Chem High Throughput Screen* 15, 515-521 (2012).
33. Cui, C. Infectious disease modeling and innate immune function in zebrafish embryos. *Methods in cell biology* 105, 273 (2011).
34. White, R.M. et al. Transparent adult zebrafish as a tool for in vivo transplantation analysis. *Cell stem cell* 2, 183-189 (2008).
35. Benard, E.L. et al. Infection of zebrafish embryos with intracellular bacterial pathogens. *J. Vis. Exp.* (2012).
36. Adatto, I., Lawrence, C., Thompson, M. & Zon, L.I. A new system for the rapid collection of large numbers of developmentally staged zebrafish embryos. *PloS one* 6, e21715 (2011).
37. Hammaren, M.M. et al. Adequate Th2-type response associates with restricted bacterial growth in latent mycobacterial infection of zebrafish. *PLoS pathogens* 10, e1004190 (2014).
38. Link, V., Shevchenko, A. & Heisenberg, C.P. Proteomics of early zebrafish embryos. *BMC. Dev. Biol.* 6, 1 (2006).
39. Lin, Y., Chen, Y., Yang, X., Xu, D. & Liang, S. Proteome analysis of a single zebrafish embryo using three different digestion strategies coupled with liquid chromatography-tandem mass spectrometry. *Anal. Biochem.* 394, 177-185 (2009).
40. Rahlouni, F., Szarka, S., Shulaev, V. & Prokai, L. A Survey of the Impact of Deyolking on Biological Processes Covered by Shotgun Proteomic Analyses of Zebrafish Embryos. *Zebrafish* 12, 398-407 (2015).
41. Rougeot, J. et al. RNA Sequencing of FACS-Sorted Immune Cell Populations from

Zebrafish Infection Models to Identify Cell Specific Responses to Intracellular Pathogens. *Methods Mol Biol* 1197, 261-274 (2014).

42. Victor, B. et al. Partially sequenced organisms, decoy searches and false discovery rates. *J Proteome Res* 11, 1991-1995 (2012).
43. Cagney, G., Amiri, S., Premawaradana, T., Lindo, M. & Emili, A. In silico proteome analysis to facilitate proteomics experiments using mass spectrometry. *Proteome science* 1, 5 (2003).
44. Lou, S. et al. High-grade sarcoma diagnosis and prognosis: Biomarker discovery by mass spectrometry imaging. *Proteomics* 16, 1802-1813 (2016).
45. McDonnell, L.A. et al. Peptide and protein imaging mass spectrometry in cancer research. *Journal of Proteomics* 73, 1921-1944 (2010).



Appendices

Nederlandse samenvatting

List of abbreviations

List of publications

Curriculum Vitae

Dankwoord

NEDERLANDSE SAMENVATTING

Ons lichaam bestaat uit cellen die met elkaar communiceren door middel van signalen. Door het in kaart brengen van signalen die cellen afgeven, ontvangen en/of verwerken kan men te weten komen hoe bepaalde ziekteverwekkers ons ziek maken. Deze kennis kunnen we ook gebruiken voor het ontwikkelen van vaccins en medicatie. Celsignalering vindt plaats op verschillende niveaus, bijvoorbeeld op DNA-niveau (wat onderzocht wordt binnen het onderzoeksveld genetica), of op mRNA-niveau (transcriptie-onderzoek), of op eiwitniveau (eiwitkundig onderzoek), enz.. Celsignalering wordt vaak onderzocht met behulp van een diemodel. In dit proefschrift is de toepassing van het diemodel zebrafish voor eiwitkundig onderzoek beschreven. Deze twee termen vereisen wat uitleg voordat het beschreven werk kan worden samengevat.

Eiwitkunde

Eiwitkunde (proteomics) is de studie naar alle eiwitten van een organisme of deel van een organisme. Maar wat is een eiwit eigenlijk? DNA uit de kern van een cel wordt afgelezen en vertaald in mRNA. mRNA wordt op zijn beurt weer vertaald in verschillende aminozuren (de bouwstenen van een eiwit). Deze aminozuren vormen peptides en meerdere peptides vormen polypeptide ketens. Verbindingen tussen deze ketens zorgen voor de draaiing en het vouwen van het eiwit. Het eindresultaat is een driedimensionale structuur. Door de veelzijdige functies en de enorme variëteit aan eiwitten en eiwittypen, bestaan er verschillende typen eiwitkundig onderzoek. Zo kan er bijvoorbeeld onderzoek worden verricht naar de structuur of de aanwezigheid van eiwitten of de interactie van eiwitten in een organisme of deel van een organisme.

Voor de experimenten beschreven in dit proefschrift is er gebruik gemaakt van de zogeheten "shotgun bottom-up" aanpak. Voor deze aanpak worden eiwitten uit de cellen van het te onderzoeken orgaan of organisme gehaald; dit proces wordt extractie genoemd. Na de extractie worden de eiwitten in stukjes geknipt door een enzym, bijvoorbeeld trypsin; een proces dat we digestie noemen. De geknipte stukjes eiwit worden vervolgens gescheiden op basis van de oplosbaarheid met behulp van een scheidingskolom. De gescheiden eiwitstukjes worden omgezet in deeltjes met lading; ionisatie genoemd. Deze geladen deeltjes worden uiteindelijk geanalyseerd door middel van massaspectrometrie. Met een massaspectrometer meten we eigenlijk de ratio van de massa van het deeltje en de lading ervan; "mass-to-charge ratio", oftewel m/z . In de massaspectrometer wordt eerst een massa-spectrometrische (MS) scan uitgevoerd. Daarna worden de deeltjes met het meest intense signaal die zich op dat moment in de massaspectrometer bevinden afgescheiden van de rest, kleiner gemaakt (gefragmenteerd) en nogmaals geanalyseerd; dit noemen we MS/MS of MS^2 . Deze analyses worden weergegeven in spectra, die door middel van softwareprogramma's kunnen worden gekoppeld aan peptide-identificaties (stukjes aaneengesloten aminozuren; oftewel aminozuursequentie) die behoren tot een bepaald eiwit.

De identificatie van de peptides kan worden uitgevoerd door middel van het doorzoeken van een sequentie database, door het koppelen van spectra met behulp van

een spectrale bibliotheek, of door middel van *de novo* sequencing (vanuit het nieuwe). In dit proefschrift beschrijven we alleen experimenten die geanalyseerd zijn met de eerste twee manieren van identificatie. Bij identificatie door middel van een database worden de geanalyseerde MS² spectra van een experiment gecorreleerd aan *in silico* (theoretisch) gegenereerde spectra. Een spectrale bibliotheek is een collectie van eerder geïdentificeerde peptide-spectra. Identificatie door middel van een bibliotheek berust op de correlatie van de spectra in de bibliotheek met de spectra van het uitgevoerde experiment. Doordat een spectrale bibliotheek informatie bevat (bijvoorbeeld over hoe het spectrum eruitziet) en een sequentie database niet, worden identificaties met behulp van een bibliotheek als een meer gevoelige methode bestempeld (1, 2). Tevens gaat het identificeren met behulp van een bibliotheek sneller dan met behulp van een sequentie database en is een bibliotheek-methode minder fout gevoelig (1, 2).

Diermodellen

Om celsignalering in kaart te kunnen brengen worden diverse soorten materiaal gebruikt. Zo kan er bijvoorbeeld onderzoek gedaan worden op cellen, of op weefsel materiaal. Om celsignalering tussen meerdere typen cellen en systemen in kaart te kunnen brengen, wordt soms gebruik gemaakt van diermodellen. Een experiment dat met behulp van diermodellen wordt uitgevoerd moet aan bepaalde eisen voldoen. Een diermodel waarbij gebruik gemaakt wordt van organismen die onder de proefdierwet vallen, vereist goedkeuring van een dierexperimentencommissie die toetst of het beschreven experiment aan de wettelijke eisen voldoet. Een zogenaamde dierproef wordt daarom te allen tijde weloverwogen uitgevoerd.

Er bestaan veel verschillende soorten diermodellen, zoals apen, muizen, ratten, slakken, wormen, enz.. De zebravis (*Danio rerio*) wordt veel toegepast als diermodel binnen biomedisch onderzoek. Ondanks het feit dat een zebravis qua uiterlijk niet op een mens lijkt zijn er wel degelijk overeenkomsten: 70% van het genetisch materiaal komt overeen met dat van de mens (3) en zebravissen beschikken over vergelijkbare organen. Doordat het immuun signaleringssysteem van de zebravis sterke overeenkomsten vertoont met dat van de mens (4-6), wordt dit model veel toegepast in onderzoek naar infectieziekten zoals tuberculose.

Onderzoek beschreven in dit proefschrift

Ondanks het feit dat in biomedische studies de zebravis een veelgebruikt diermodel is, is het een relatief "nieuw" diermodel binnen het eiwitkundig onderzoek. Binnen het eiwitkundig onderzoek zijn er diverse spectrale bibliotheken beschikbaar van verschillende diermodellen. Zoals eerder vermeld kunnen de spectra van het uitgevoerde experiment geïdentificeerd worden aan een bepaald eiwit met behulp van spectrale bibliotheken. Echter voor de zebravis ontbrak zo een spectrale bibliotheek. In **hoofdstuk 2** is de ontwikkeling van een spectrale bibliotheek voor het zebravis-model beschreven. Daarnaast is er een methode ontwikkeld om het creëren en valideren van toekomstige spectrale bibliotheek te kunnen automatiseren, wat tijdswinst oplevert voor de onderzoeker. Tevens is de bibliotheek beschikbaar gesteld via the National Institute of Standards and Technology's (NIST), een website die o.a. spectrale bibliotheken van

verschillende diersmodellen beheerd (7), zodat deze voor iedereen vrij toegankelijk is.

De mate waarin een diersmodel op de mens lijkt of de mate waarin diersmodellen onderling op elkaar lijken, kunnen worden weergegeven met behulp van fylogenie (een soort familie stamboom). Eiwitkundig onderzoek leidt met enige regelmaat tot nieuwe fylogenetische inzichten. Voor een eerder beschreven methode, die directe vergelijkingen maakt tussen massaspectra van peptiden, werd gesuggereerd dat deze op vele manieren toepasbaar zou zijn binnen bijvoorbeeld de fylogenie, experimentele ontwerpen en forensisch onderzoek. In **hoofdstuk 3** worden de verschillende toepassingsmogelijkheden van deze methode beschreven.

De zebravis is een klein dier. Dit kan problemen opleveren wanneer bijvoorbeeld de verschillende onderdelen van het hart van dit dier moeten worden onderzocht, of wanneer bepaalde cellen uit het lichaam gehaald moeten worden waarbij een bepaald volume moet worden behaald. De genetisch gerelateerde en veel grotere soortgenoot van de zebravis, de gewone karper (*C. caprio*), zou een oplossing kunnen zijn voor bovengenoemde beperkingen. In **hoofdstuk 4** wordt de mogelijkheid beschreven om een genetisch vergelijkbare soortgenoot in te zetten om beperkingen van de ander mee op te vullen.

De zebravis is een veelvuldig beschreven diersmodel binnen genetische studies (DNA, RNA, mRNA, enz.). Echter, de zebravis is minder frequent beschreven binnen het eiwitkundig onderzoek. In **hoofdstuk 5** wordt een overzicht beschreven van de mogelijkheden, beperkingen en praktische overwegingen met betrekking tot het gebruik van het diersmodel zebravis voor toepassing binnen het eiwitkundig onderzoek. Ook worden beschikbare methoden en hulpmiddelen die toepasbaar zijn voor dit type onderzoek beschreven. Daarnaast wordt in dit hoofdstuk de toepasbaarheid beschreven van een eiwitkundig onderzoek naar tuberculose infecties met de zebravis als diersmodel.

Conclusies

In dit proefschrift is de toepasbaarheid van het diersmodel zebravis binnen het eiwitkundig onderzoek verder ontwikkeld door het genereren van een spectrale-zebravis-bibliotheek, data-analyse platforms, nieuwe onderzoeksmethodes en hulpmiddelen. Het doel hiervan was om het diersmodel zebravis beter toe te kunnen passen in biomedisch gerelateerd eiwitkundig onderzoek.

Als zodanig heeft het in dit proefschrift beschreven onderzoek bijgedragen aan de ontwikkeling van het zebravis diersmodel voor eiwitkundig onderzoek, waarin de publicatie van de eerste spectrale bibliotheek van zebravissen een belangrijke eerste stap is geweest.

REFERENTIES

- 1 Lam, H. et al. Development and validation of a spectral library searching method for peptide Identification from MS/MS. *Proteomics*. 7, 655-667 (2007).
- 2 Lam, H. et al. Building consensus spectral libraries for peptide Identification in proteomics. *Nat.Methods* 5, 873-875 (2008).
- 3 Howe, K. et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* advance online publication, doi:10.1038/nature12111 (2013).
- 4 Traver, D. et al. The zebrafish as a model organism to study development of the immune system. *Adv.Immunol.* 81, 253-330 (2003).
- 5 Trede, N. S., Langenau, D. M., Traver, D., Look, A. T. & Zon, L. I. The use of zebrafish to understand immunity. *Immunity*. 20, 367-379 (2004).
- 6 Meijer, A. H. & Spaink, H. P. Host-pathogen interactions made transparent with the zebrafish model. *Curr.Drug Targets*. 12, 1000-1017 (2011).
- 7 NIST zebrafish database (<http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:lib:zebrafish20150109>) (2018).

LIST OF ABBREVIATIONS, ACRONYMS, AND PROGRAM NAMES

AB	Wild type zebrafish line, derived from two lines, A and B, purchased by Streisinger at different times from a pet shop in Albany, Oregon, USA.
<i>ab initio</i>	Latin for “from the beginning”
A-Library	Adult zebrafish spectral library
AE-library	Adult and Embryonic zebrafish spectral library
AMPureXP	Agencourt AMPure XP (PCR purification mixture)
ATP	Adenosine triphosphate
AUGUSTUS	An eukaryote gene prediction program which predicts on the basis of pure sequences (<i>ab Initio</i>)
BLASTP	Basic Local Alignment Search Tool for proteins to proteins
BLASTX	Basic Local Alignment Search Tool. A database which can be used to search the six-frame conceptual translated nucleotide query products (both strands) against protein sequences databases
bp	Base pair
cDNA	Copy or complement DNA
CLIP	Cre-LoxP Inverse PCR
contigs	A set of overlapping DNA segments. The set represents a consensus region of DNA
Comet	Open source tandem mass spectrometry sequence database search engine
COMICS	A webtool to visualize omics data in a cartoon
CompassXport	Software tool used for converting Mass Spectrometry data to mzXML format
CompareMS ²	Open source software tool for direct comparison of tandem mass spectra
CID	Collision Induced Dissociation
CPM	Counts per million
2D	Two dimensional
Da	Dalton (atomic mass unit)
DataAnalysis	Software tool from Bruker for direct analysis of mass spectrometry data from Bruker Instruments
DEC	Dier Experimenten Commissie (Animal welfare committee)
<i>de novo</i>	Latin for “of new”
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol

dpf	Days post fertilization
E-value	The Expected value after a BLAST search describes the number of expected hits by chance when searching a database. Value from 0-1. The closer to 0 means a more “significant” BLAST match
EDTA	Ethylenediaminetetraacetic acid
edgeR	Software tool for Empirical Analysis of Digital Gene Expression Data in R
Ensembl	A vertebrate genome browser, used for annotation, computing and prediction
ESI	Electrospray Ionization
ETD	Electron-transfer dissociation
FACS	Fluorescent Activated Cell Sorting
FASTA	A text based file format for nucleotide or peptide sequences
FDR	False Discovery Rate
FTICR	Fourier Transform Ion Cyclotron Resonance Mass Spectrometer
GB	GigaByte
GO	Gene Ontology. Created to make gene and gene product attributes comparable across all species
gplots	Software tool for plotting data in R
GraphPad Prism	Software tool used for statistics
HCl	Hydrogen Chloride
hcTnC	Human cardiac troponin C
hcTnI	Human cardiac troponin I
hcTnT	Human cardiac troponin T
HIV	Human Immunodeficiency virus
hpf	Hours post fertilization
HTS	High Throughput Screening
HTSeq-count	High-Throughput sequencing count (software tool)
HyStar	Software tool that coordinates experiments when a liquid chromatography system is coupled with spectroscopic methods
<i>in silico</i>	Latin for performed on a computer, or computer simulated
<i>in vivo</i>	Latin for within the living
InterProphet	Software tool that further processes PeptideProphet output results from several experiments and search engines before processing by ProteinProphet
InterProScan	Software package allowing classification of proteins.
IOD	Iodoacetamide
IT	IonTrap

JOVE	Journal of Visualized Experiments
LC	Liquid Chromatography
LCP	LytR-CspA-Psr family protein
LTB	Latent Tuberculosis
MAKER	A genome annotation pipeline which is easily configurable
MDR-TB	Multi Drug Resistant Tuberculosis
MEGA	Molecular Evolutionary Genetics Analysis (software tool)
mg	Milligram
mgf	Data format for mass spectrometry data to enable data analysis
MICId	Microorganism Classification and Identification software
mM	Millimolar
MOPS	A buffer consisting of 3-(N-morpholino) propanesulfonic acid
MS	Mass spectrometry
MS ¹	First MS scan which registers the <i>m/z</i> ratio and intensities of a batch of peptides
MS ² (or MS/MS)	Second MS scan of selected and fragmented precursor ion from MS ¹
<i>m/z</i>	Mass-to-charge
mzXML	Data format; which is an XML open standard for storing MS ⁿ data
mRNA	Messenger Ribonucleic Acid
Mya	Millions of years ago
MYD88	Myeloid differentiation primary response 88
MΦ	Macrophages
NCBI	National Center for Biotechnology Information
NEXUS	Software tool for building phylogenetic trees.
NIST	National Institute of Standards and Technology's
O/N	Overnight
PacBio	Pacific Biosciences
PBJelly	PBJelly is a highly automated pipeline that aligns long sequencing reads (such as PacBio RS reads or long 454 reads in fasta format) to high-confidence draft assemblies
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PepAlign	A software program that aligns peptide identifications from LC-MS/MS (e.g. in pepXML)
PeptideProphet	A software tool for generating statistical validation of MS/MS search engines' spectra-to-peptide sequence assignments

PepWarp	Software program that warps MS/MS data (e.g. in pepXML)
pH	A numeric scale for aqueous solutions to specify its acidity or basicity (alkalinity)
PreCR	Repair Mix. An enzyme cocktail formulated to repair damaged template DNA prior to its use in the polymerase chain reaction
PRIDE	The PRoteomics IDentifications database
ProteinProphet	A software tool for generating probabilities for protein identifications based on MS/MS data, making use of PeptideProphet output data files
PSM	Peptide-spectrum match
PTM	Post Translational Modifications
QGIS	Software tool: open-source geographic Information system (GIS) for desktops
R	Free software environment for statistical computing and graphics
Repbase	Database of repetitive DNA elements
RepeatModeler	A software package for de-novo repeat family identification and modelling
RMBlast	A version of RepeatMasker which is compatible with the standard NCBI blast program
RNA	Ribonucleic acid
RNA-seq	RNA-sequencing (or whole transcriptome shotgun sequencing)
ROC curve	Receiver Operating Characteristic curve
RPKM	Reads per million per kilobase
SAMTOOLS	Sequence Alignment/Map software tool. A generic format for large nucleotide sequence alignments storage
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
SMRT	Single molecule real time sequencing
SSPACE	SSPACE standard is a stand-alone program for scaffolding pre-assembled contigs using NGS paired-read data
SpectraST	Spectra Search Tool
Swiss-Prot	Protein database of the Swiss Institute of Bioinformatics
TAE buffer	Tris-acetate-EDTA buffer
Taverna	Software tool: open source and domain-independent Workflow Management System
TB	Tuberculosis
TFA	Trifluoroacetic acid
TLRs	Toll Like Receptors

Tophat	Software tool to identify splice junctions by mapping the aligning RNA-seq reads to mammalian-sized genomes
TPP	Trans Proteomic Pipeline
TrapControl	Software tool that controls the IonTrap system
TrEMBL	Translated EMBL Nucleotide Sequence Data Library
TRIS	Tris(hydroxymethyl)aminomethane
Uniprot	Universal Protein Resource
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
μL	Microliter
XDR-TB	Extensively Drug Resistant Tuberculosis
X!Tandem	Open source software that can match tandem mass spectra with peptide sequences, for protein identification
UPGMA	Unweighted Pair Group Method with Arithmetic Mean (a agglomerative hierarchical clustering method)
ZFA	Zebrafish Anatomy Ontology
ZIRC	Zebrafish International Research Center
ZFIN	The Zebrafish Information Network

LIST OF PUBLICATIONS

van der Plas – Dulvesteljn, S.J.; de Rooij, L.P.M.H.; Wu, Y.; Dabhoiwala, N.F.; Wallner, C.; Lamers, W.H. Developmental anatomy of the murine urethral sphincter complex. In preparation.

van der Plas – Dulvesteljn, S.J.; Benard, E.L.; Kolder, I.C.R.M.; Stam, K.A.; Meijer, A.H.; Palmblad, M. Zebrafish Models in Proteomics. In preparation. (This thesis)

Horn, I.R.; Kenens, Y.; Palmblad N.M.; **van der Plas – Dulvesteljn, S.J.**; Langeveld, B.W.; Meijer, H.J.M.; Dalebout, H.; Fischer, A.; Florens, F.B.V.; Niemann, J.; Rijdsdijk, K.F.; Schulp, A.; Gravendeel, B. Palaeoproteomics of bird bones for taxonomic classification. Submitted, September 2018.

de Keijzer J.; **van der Plas – Dulvesteljn, S.J.**; Mulder, A.; de Zwaan, R.; de Ru1, A.H.; van Veelen, P.A.; van Soolingen D.; Palmblad, N.M. Rapid Identification of phylogenetic relationships in the Mycobacterium genus by direct comparison of shared spectral content. Submitted, August 2016.

Travin, D.; Popov, I.; Güler, A.Z.; Medvedev, D.; **van der Plas – Dulvesteljn, S.J.**; Álvarez, M.V.; Kolder, I.C.R.M.; Meijer, A.H.; Spalink, H.P.; Palmblad, N.M. COMICS: Cartoon Visualization of Omics Data in Spatial Context Using Anatomical Ontologies. *Journal of Proteome Research* 2018, 17, 739-744.

Kolder, I.C.R.M.; **van der Plas – Dulvesteljn S.J.**; Tan, G.; Wiegertjes, G.F.; Forlenza, M.; Güler, A.T.; Travin, D.Y.; Nakao, M.; Moritomo, T.; Irnazarow, I.; den Dunnen, J.T.; Anvar, S.Y.; Jansen, H.J.; Dirks, R.P.; Palmblad, N.M.; Lenhard, B.; Henkel, C.V.; Spalink, H.P. A full-body transcriptome and proteome resource for the European common carp. *BMC genomics* 2016, 17, 701. (This thesis)

van der Plas – Dulvesteljn, S.J. and Kraaljeveld, K, Response to Kopps & Palsboll. *Journal of Sea Research* 2016, 110, 16.

van der Plas – Dulvesteljn, S.J.; Wulff, T.; Klychnikov, O.; Ohana, D.; Dalebout, H.; van Veelen, P.; de Keijzer, J.; Nessen, A.M.; van der Burgt, Y.E.M.; Deelder, A.M.; Palmblad, N.M., Differentiating Samples and Experimental Protocols by Direct Comparison of Tandem Mass Spectra, *Rapid Communications In Mass Spectrometry* 2016, 30, 731-738. (This thesis)

van der Plas – Dulvesteljn, S.J.; Smit, F.J.L.; van Alphen, J.J.M.; Kraaljeveld, K, Harbor porpoise Phocoena phocoena strandings on the Dutch coast: No genetic structure, but evidence for inbreeding. *Journal of Sea Research* 2015, 97, 24-27.

van der Plas – Dulvesteljn, S. J.; Mohammed, Y.; Dalebout, H.; Meijer, A.; Botermans, A.; Hoogendijk, J. L.; Henneman, A. A.; Deelder, A. M.; Spalink, H. P.; Palmblad M. Identifying

proteins in zebrafish embryos using spectral libraries generated from dissected adult organs and tissues. *Journal of Proteome Research* 2014, 13, 1537–1544. (This thesis)

Brittijn, S. A.; **Dulvesteljn, S. J.**; Belmamoune, M.; Bertens, L. F. M.; Bitter, W.; De Bruijn, J. D.; Champagne, D. L.; Cuppen, E.; Flik, G.; Vandenbroucke-Grauls, C. M.; Janssen, R. A. J.; De Jong, I. M. L.; De Kloet, E. R.; Kros, A.; Meijer, A. H.; Metz, J. R.; van der Sar, A. M.; Schaaf, M. J. M.; Schulte-Merker, S.; Spaink, H. P.; Tak, P. P.; Verbeek, F. J.; Vervoordeldonk, M. J.; Vonk, F. J.; Witte, F.; Yuan, H. P.; Richardson, M. K., Zebrafish development and regeneration: new tools for biomedical research. *International Journal of Developmental Biology* 2009, 53, (5-6), 835-850.

CURRICULUM VITAE

Suzanne van der Plas – Duivesteljn was born on November the 10th, 1985 in 's-Gravenhage, the Netherlands. Suzanne graduated at the Bonaventuracollege in Leiden in 2003. Driven by the ambition to become a marine biologist (to study cetaceans), she started the study Biology and Medical Laboratory education at the Hogeschool in Leiden at the end of the summer of 2003. Next to this study, Suzanne followed the Ecocorps Summer Internship Programme with the Cetacean Research and Rescue Unit, Scotland, to gain experience on cetacean research. Within 4 years, she received her Bachelor of Applied Sciences with an additional article 12 diploma in 2007 (which allowed her to work with experimental animals). Directly after graduation she was employed as research technician at the department of Integrative Zoology of Leiden University where she continued working on her bachelor Internship project: Zebrafish Caudal Fin Regeneration. This work resulted in her first publication in the International Journal of Developmental Biology in 2009. Suzanne started her master in Biology at Leiden University in 2007, and graduated within two years (2009) with the addition of an art. 9 diploma (allowing her to design experiments that require the use of experimental animals). During this master, Suzanne organized the European Cetacean Society (ECS) conference 2008 in the Netherlands.

Directly following her studies, Suzanne was employed at Integrative Zoology (Leiden University) for five months, as teaching and research assistant, while integrating into the cetacean research field in the Netherlands. The financial crisis kicked in and Suzanne was forced to change her plans from a career in marine biology into another ambition. Therefore, in 2009, she started a PhD project at the Academic Medical Center at the department of anatomy and embryology to study pelvic floor development and aging symptoms, with emphasis on incontinence problems. After a promising first year the project unfortunately had to finish prematurely due to insufficient funding, which forced Suzanne to change her career plans again. Between November 2010 and April 2011, Suzanne gained experience as research assistant, animal caretaker, and teaching assistant. In April 2011, she got the opportunity to start her PhD trajectory at the Leiden University in the Center of Proteomics and Metabolomics, supervised by Prof. Dr. André Deelder, Prof. Annemarie Meijer (from Leiden University), and Dr. Magnus Palmblad, entitled: "Advancing Zebrafish Models in Proteomics".

In her spare time Suzanne published her master traineeship project (a genetic study on harbour porpoises in the Dutch part of the North Sea) in the Journal of Sea research in 2015. As a volunteer at the Dorsal Fin Foundation (Stichting Rugvin) where Suzanne participates the study on Harbour Porpoises in Dutch coastal waters, she is still involved in cetacean research. Suzanne is also active as drama actor at drama club St.Theodulfus.

Suzanne and her husband married in 2011, and became parents of a son (spring 2014) and daughter (summer 2016). Suzanne is currently employed at the department of Parasitology at the LUMC. During her career Suzanne experienced that she likes managing, organizing, and teaching the most. She likes to work on topics involving anatomy, embryology, cetacean, zebrafish and medical research.

DANKWOORD

Dit proefschrift is het resultaat van een aantal jaren hard werken. Naast alle wetenschappelijke kennis en ervaring ben ik een onschatbare mix van ervaringen rijker. Dat je een proefschrift nooit alleen kunt maken bewijst het enorme aantal mensen met wie ik heb samengewerkt en wie ik hiervoor graag wil bedanken.

Prof. Deelder, André, Prof. Meijer, Annemarie en Dr. Palmblad, Magnus, dank jullie wel voor alle ondersteuning, wijsheid, gastvrijheid, het vertrouwen en voor de vrijheid die ik kreeg om zelfstandig mijn experimenten uit te mogen voeren op verschillende locaties (LUMC, Gorlaeus en Sylvius). Magnus, *tack så mycket för allt*.

Paranimfen kies je om een speciale reden. Bernadette, ooit begon mijn PhD carrière samen met jou op het AMC. Ik heb veel van je geleerd en dat heeft mij geholpen om mijn promotie onderzoek te volbrengen. Je bent een waardevolle vriendin. Dana, dank je wel voor een onvergetelijke PhD tijd. Ik heb een hoop van je geleerd. Dank je wel voor al je advies, gesprekken je (h)eerlijke directheid en de lol op de kamer en in het lab. Dames, op onze vriendschap!

A huge thanks for the laughs, endless support, and everything else during the last 4 years to my room mates: Hans, Simone, Kate, Dana, Linda, Irene, Anton, Tugce, Clara, Karin, Guinevere, Yassene, Jan, Bjorn and Kristell. Hans, een speciaal woord van dank gaat uit naar jou, zonder jou had ik de LC-MS systemen vast nooit zo goed begrepen als nu. Dank je voor je tijd en enorme geduld die je had bij het uitleggen en oplossen van problemen op het P-O lab. Computer languages and statistics involving programming are complicated for me. Tugce, Iris, Koen, Yassene, Rob, and Alex, many thanks for your support on these topics. I could always walk into your office for questions. Without your help I could not have analysed the data for this thesis. Leonie en Kirsten bedankt voor jullie luisterend oor en medeleven, dit heeft mij geholpen om erdoorheen te komen. Many thanks for the advice that helped me being a mom and a scientist to Sha, Tiziana, and Rosina. Manu thank you for everything. Working in such a large group as the CPM, I will never be able to thank everyone in person. Still, I'd like to thank all my colleagues for their help, advice, fun, laughs, "blikje-tijd", CPM drinks, Christmas celebrations (and its organization), and being a Quality Crew-member. My PhD trajectory started when the units of parasitology and CPM were still joint, and I am happy to be a member of the PARA team again. Thanks to all PARA's for your support, with a special thanks to Prof. Maria Yazdanbakhsh, Leonie, Sanne, and Koen.

Ik vind het leuk om studenten te begeleiden. Jordy en Anouk, als team hebben we gezorgd voor de totstandkoming van de eerste publicatie voor dit proefschrift. Bedankt voor deze mooie ervaring. Coen ook jou heb ik mogen begeleiden wat o.a. resulteerde in de "fylogenie" figuur in de introductie van deze thesis. Dank je wel.

Op het Gorlaeus laboratorium kreeg ik het privilege om volledig zelfstandig te mogen werken binnen de Immunobiologie-groep en het Animal Sciences & Health research cluster. Beste Prof. Spaijk en Prof. Meijer, Herman en Annemarie, ik wil jullie bedanken

voor jullie gastvrijheid en vertrouwen en ook voor jullie hulp als ik vragen had over het onderzoek. *At the Gorlaeus I felt welcome and for that I'd like to thank the entire group for making me feel part of your team! With a special thanks to Erica, Michiel, Chao, Ania, Zakia, Malgorzata, Anita, Wouter, Saskia, Ulricke, and Davy, who helped me with all my questions on the fish, bacteria, and injection-techniques. Thanks for still making me feel "at home" at Integrative Zoology (where I spend many lunches due to all zebrafish related experiments at the Sylvius laboratory), with a special thanks to Prof. Michael Richardson, Merijn, Ilse, Janine, Freek, Jacco, Marjolijn, Peter, Faiza, Shaukat, Laurie, and Eric, thanks for your "gezelligheid" and "luisterend oor"! Gerda dank je wel voor je tijd en uitleg van de fluorescentie-microscopen en je behulpzame adviezen tijdens mijn studie en promotietijd. Alle medewerkers van ZF-screens, met name Peti, wil ik danken voor hun samenwerking (zichtbaar en onzichtbaar). I also worked with many persons in joint ventures. Iris, Christiaan, Marja, Tune, Barbara, Ivo, Fabrice, Jeroen, Dmitrii, Iaroslav, and Kirsten, thanks for the fruitful collaborations.*

Er was geen promotie geweest zonder de steun van mijn lieve familie. Pappa en Mamma, van jullie mocht ik studeren wat ik wilde en kreeg hiervoor alle steun die een kind kan wensen. Waar mogelijk stonden jullie voor mij klaar. Mijn dank is niet in woorden te omschrijven. Lieve Mattijs en Anna, jullie grote aanstekelijke glimlach is een fantastische thuiskomst na elke werkdag. Lieve Peter, aan jou draag ik dit proefschrift op om te veel redenen die ik hier kan noemen. Jij hebt mij samen met de kinderen geleerd dat een liefdevol thuis dat jullie mij bieden zoveel belangrijker is dan werk, maar ook dat zo'n goede thuisbasis heel belangrijk is voor een vruchtbare carrière. Zonder jouw steun was dit proefschrift er zeker niet geweest.

Suzanne

