Journal of proteome research

Article

pubs.acs.org/jpr

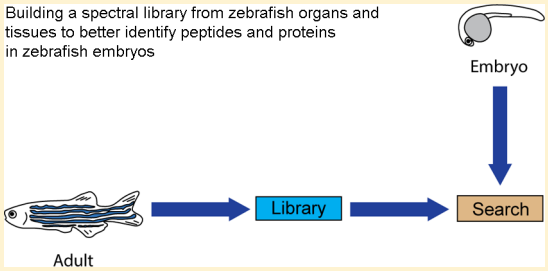
Identifying Proteins in Zebrafish Embryos Using Spectral Libraries Generated from Dissected Adult Organs and Tissues

Suzanne J. van der Plas-Duivesteijn,[†] Yassene Mohammed,[†] Hans Dalebout,[†] Annemarie Meijer,[‡] Anouk Botermans,[†] Jordy L. Hoogendijk,[†] Alex A. Henneman,[†] André M. Deelder,[†] Herman P. Spaink,[‡] and Magnus Palmblad*,[†]

[†]Center for Proteomics and Metabolomics, Leiden University Medical Center, Albinusdreef 2, Leiden 2300 RC, The Netherlands

[‡]Department of Molecular Cell Biology, University of Leiden, Leiden, The Netherlands

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 of 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.



KEYWORDS: *spectral libraries, SpectraST, peptide identification, dissection, embryos, adults, zebrafish, Taverna, workflows*

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 data sets. 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, and 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 are 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 are already publicly available in repositories such as PRIDE,[12] according to the authors' knowledge no spectral library for zebrafish is currently in the public domain (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 in solving important biological questions. Complementary to embryos and larvae, organs and tissues of

Received: October 28, 2013
Published: January 27, 2014

dx.doi.org/10.1021/pr4010585 | *J. Proteome Res.* 2014, 13, 1537−1544

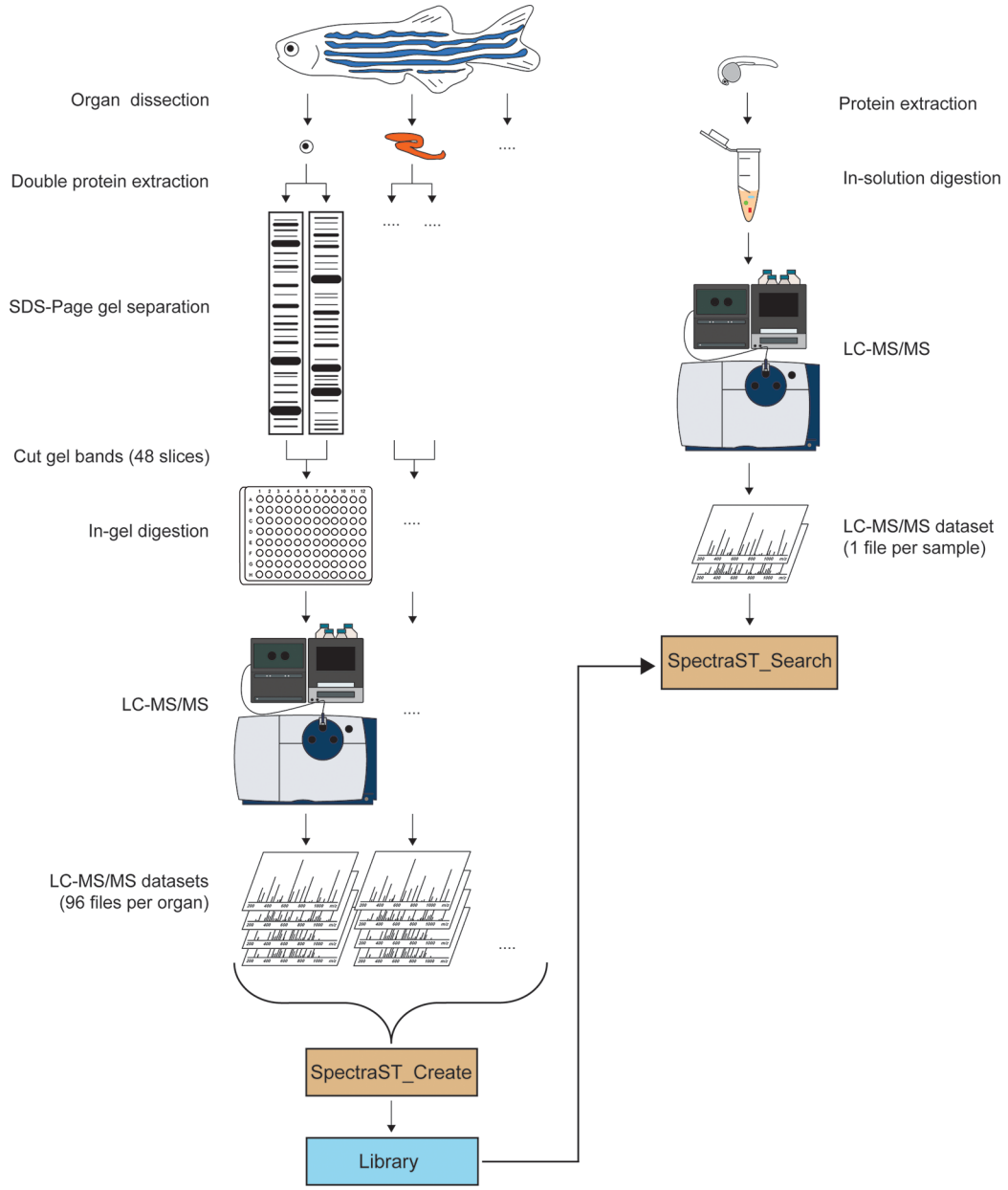


Figure 1. Experimental workflow of building and searching the zebrafish in-depth library. Tissues and organs of adult zebrafish were harvested to generate an in-depth MS/MS analysis (using gel- on top of LC- separation). An organ fractionated spectral library (A-library) was built. Spectra of 1−6 dpf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

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 includes 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 data sets 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 no. 11221. Organs of male and female of ∼4.5 month old zebrafish (nacre strain AB background) were dissected. Before dissection, the zebrafish were maintained at 28 °C on a 14/10 h light/dark cycle. The day before dissection, the fish were denied food to reduce the amount of food matter in the intestine.

dx.doi.org/10.1021/pr4010585 | J. Proteome Res. 2014, 13, 1537−1544

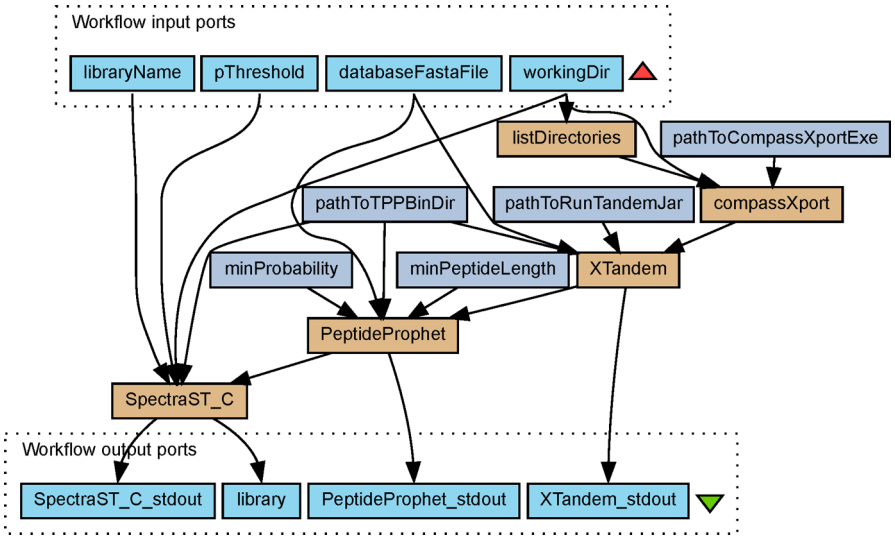


Figure 2. Scientific workflow for creating spectral libraries. The workflow consists of five Java BeanShell processors (brown). The *listDirectories* prepares a list of all .d directories used as input. The *compassXport* is a format converter and prepares the data into the right format for the *XTandem* 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.

Organ Dissection

The zebrafish were anaesthetized with ice water. Organ dissection started immediately after sedation.[15] Within 15 min the following organs were collected: ovaries, testes, muscle (tail region), eyes, brain, intestine, fins, gills, heart, skin, swim bladder, spleen, kidney, liver, pancreas, and gall bladder. Each organ was placed in a Safe-Lock microcentrifuge tube (Eppendorf), snap-frozen in liquid nitrogen, and stored at −80 °C. To get a convenient amount of protein with which to work, 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 to 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 organ and tissue 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 8 M urea, 75 mM NaCl, 50 mM Tris-HCL pH 8.2, 50 U/mL benzonase (E1014-5KU, Sigma-Aldrich), 2 mM $MgCl_2$, 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 min. Samples were placed at 4 °C for 30 min, and the supernatant was collected after centrifugation at 16 000 × *g* for another 30 min 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 mM $MgCl_2$ and one using 1% SDS, 50 U/mL benzonase, and 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 afterward. 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.

dx.doi.org/10.1021/pr4010585 | *J. Proteome Res.* 2014, 13, 1537−1544

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 splitless NanoLC-Ultra 2D plus system (Eksigent) with a linear 90 min (for in-solution digests) or 45 min (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 min 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 a library spectral libraries[19] (Figure 2). (The Taverna workflow and both libraries (A and AE) are available at https://www.ms-utils.org/zebrafish.) 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 mainly used 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 programs, 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 MS^n 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.

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 data sets of whole organ digests were converted to MGF using DataAnalysis 4.0 (Bruker) and a maximum of 4000 compounds. These were compared by compareMS[2],[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.

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 bloodstream 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 old 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, for example, 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

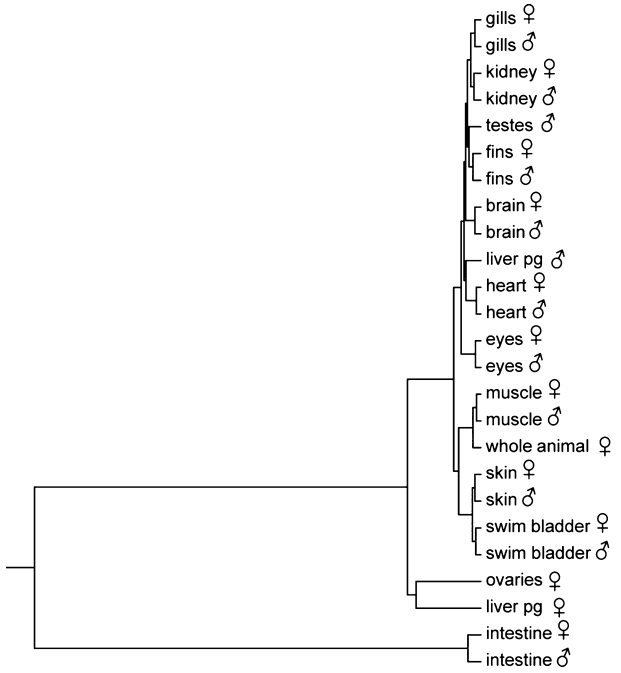


Figure 3. Relationship between whole-organ LC−MS/MS data sets from dissected zebrafish. Hierarchical clustering of LC−MS/MS data sets from dissected adult male and female zebrafish organs and tissues using a previously described method for phylogenetic analysis.[13] Most organs show little difference between sexes, with the salient exceptions of the male and female reproductive glands and livers. The liver samples included the pancreas and gall bladder.

dx.doi.org/10.1021/pr4010585 | J. Proteome Res. 2014, 13, 1537−1544

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 2). A Taverna[19] workflow was used to search all data against genome-derived sequences and combine the results into a spectral library (Figure 3). 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 h postfertilization (hpf) zebrafish embryos previously published by Lößner et al.[28] As small data-set probes, zebrafish developmental stages of 1−6 days postfertilization (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 PSMs at a 1% false discovery rate (FDR) compared with both the embryonic library and sequence database search. In total, 8563 spectra were identified from these 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

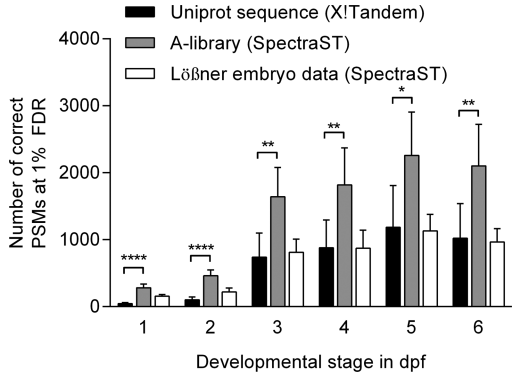


Figure 4. Improvement on peptide spectrum matches in embryos by using spectral libraries from adult organs. Searching data from different developmental stages (1−6 dpf) against the A library using SpectraST resulted in more correct peptide-spectrum matches (PSMs) compared with an X!Tandem search against a Uniprot sequence database from *Danio rerio*. The SpectraST search against a spectral library from embryos only (compiled from a data set previously published by Lößner et al.[28]) showed comparable results to the X!Tandem search. Each developmental stage was sampled six times and analyzed 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 data set 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.

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 with 44 and 100 PSMs, respectively) than the later developmental stages (3−6 dpf), showing a smaller difference between the two searches.

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 to 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 applications of the same statistical

dx.doi.org/10.1021/pr4010585 | J. Proteome Res. 2014, 13, 1537−1544

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 5280 and 8120 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 data sets 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 by 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 data sets could be searched using our libraries with at least as good or better results than a sequence database search.

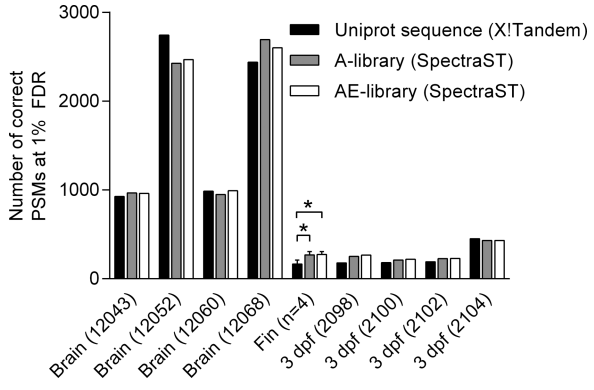


Figure 5. Validation of the A and AE libraries against data from other laboratories. Spectra from different data sets available from PRIDE were searched against our A and AE libraries using both SpectraST and X!Tandem with the zebrafish UniProt sequence database. Both libraries identified significantly 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 with the X!Tandem search in PRIDE data sets from zebrafish brain and 3 dpf embryos. The data-set numbers (n = 1) for brain and 3 dpf embryo data represent PRIDE references. The PRIDE data sets 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 nonredundant 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 and glycosylation.

The spectral libraries work best when applied to data from similar instruments in the same laboratory but also work 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 data set, 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.

■ CONCLUSIONS

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.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +31 71 5269526. Fax: +31 71 5266907. E-mail: n.m.palmblad@lumc.nl.

Notes

The authors declare no competing financial interest.

■ 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.).

■ ABBREVIATIONS

dpf, days post fertilization; A library, adult library; AE library, adult and embryonic library; PSMs, peptide-spectrum matches

■ REFERENCES

(1) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R. Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* **2008**, *5* (10), 873−875.

(2) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7* (5), 655−667.

(3) Lange, V.; Picotti, P.; Domon, B.; Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **2008**, *4*, 222.

(4) Deutsch, E. W.; Lam, H.; Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **2008**, *9* (5), 429−434.

(5) Brittijn, S. A.; Duivesteijn, 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. *Int. J. Dev. Biol.* **2009**, *53* (5−6), 835−850.

(6) Traver, D.; Herbomel, P.; Patton, E. E.; Murphey, R. D.; Yoder, J. A.; Litman, G. W.; Catic, A.; Amemiya, C. T.; Zon, L. I.; Trede, N. S. The zebrafish as a model organism to study development of the immune system. *Adv. Immunol.* **2003**, *81*, 253−330.

(7) Phelps, H. A.; Neely, M. N. Evolution of the zebrafish model: from development to immunity and infectious disease. *Zebrafish* **2005**, *2* (2), 87−103.

(8) Trede, N. S.; Langenau, D. M.; Traver, D.; Look, A. T.; Zon, L. I. The use of zebrafish to understand immunity. *Immunity* **2004**, *20* (4), 367−379.

(9) Cui, C. Infectious disease modeling and innate immune function in zebrafish embryos. *Methods Cell Biol.* **2011**, *105*, 273.

(10) Mione, M.; Meijer, A. H.; Snaar-Jagalska, B. E.; Spaink, H. P.; Trede, N. S. Disease modeling in zebrafish: cancer and immune responses–a report on a workshop held in Spoleto, Italy. *Zebrafish* **2009**, *6* (4), 445−451.

(11) Gerhard, G. S. Comparative aspects of zebrafish (Danio rerio) as a model for aging research. *Exp. Gerontol.* **2003**, *38* (11−12), 1333−1341.

(12) Vizcaíno, J. A.; Côté, R.; Csordas, A.; Dianes, J.; Fabregat, A.; Foster, J.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; O'Kelly, G.; Schoenegger, A.; Ovelleiro, D.; Pérez-Riverol, Y.; Reisinger, F.; Ríos, D.; Wang, R.; Hermjakob, H. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **2013**, *41* (Database issue), 1063−1069.

(13) Palmblad, M.; Deelder, A. M. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun. Mass Spectrom.* **2012**, *26* (7), 728−732.

(14) Westerfield, M. *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio rerio)*, 4th ed.; Univ. of Oregon Press: Eugene, OR, 2000.

(15) Gupta, T.; Mullins, M. C. Dissection of organs from the adult zebrafish. *J. Vis. Exp.* **2010**, No. 37, 1717.

(16) Link, V.; Shevchenko, A.; Heisenberg, C. P. Proteomics of early zebrafish embryos. *BMC Dev. Biol.* **2006**, *6*, 1.

(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.* **2009**, *394* (2), 177−185.

(18) Mostovenko, E.; Deelder, A. M.; Palmblad, M. Protein expression dynamics during Escherichia coli glucose-lactose diauxie. *BMC Microbiol.* **2011**, *11*, 126.

(19) Wolstencroft, K.; Haines, R.; Fellows, D.; Williams, A.; Withers, D.; Owen, S.; Soiland-Reyes, S.; Dunlop, I.; Nenadic, A.; Fisher, P.; Bhagat, J.; Belhajjame, K.; Bacall, F.; Hardisty, A.; Nieva de la Hidalga, A.; Balcazar Vargas, M. P.; Sufi, S.; Goble, C. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.* **2013**, *41*, 557−561.

(20) Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R. A common open representation of mass spectrometry data and its

application to proteomics research. *Nat. Biotechnol.* **2004**, *22* (11), 1459−1466.
(21) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.
(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.* **2002**, *74* (20), 5383−5392.
(23) Ma, K.; Vitek, O.; Nesvizhskii, A. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinf.* **2012**, *13* (Suppl 16), S1.
(24) Maddison, D. R.; Swofford, D. L.; Maddison, W. P., NEXUS: an extensible file format for systematic information. *Syst. Biol* **1997**, *46*, (4).
(25) Tamura, K.; Dudley, J.; Nei, M.; Kumar, S. *Mol. Biol. Evol.* **2007**, *24* (8), 1596−1599.
(26) Sokal, R.; Michener, C., A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **1958**, *38*.
(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.* **2003**, *2* (1), 4.
(28) Lößner, C.; Wee, S.; Ler, S. G.; Li, R. H.; Carney, T.; Blackstock, W.; Gunaratne, J. Expanding the zebrafish embryo proteome using multiple fractionation approaches and tandem mass spectrometry. *Proteomics* **2012**, *12* (11), 1879−1882.
(29) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150−9.
(30) Jessen, J. R.; Willett, C. E.; Lin, S. Artificial chromosome transgenesis reveals long-distance negative regulation of rag1 in zebrafish. *Nat. Genet.* **1999**, *23* (1), 15−16.
(31) Grandel, H.; Lun, K.; Rauch, G.-J.; Rhinn, M.; Piotrowski, T.; Houart, C.; Sordino, P.; Küchler, A. M.; Schulte-Merker, S.; Geisler, R.; Holder, N.; Wilson, S. W.; Brand, M. 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* **2002**, *129* (12), 2851−2865.
(32) Qi, H. H.; Sarkissian, M.; Hu, G.-Q.; Wang, Z.; Bhattacharjee, A.; Gordon, D. B.; Gonzales, M.; Lan, F.; Ongusaha, P. P.; Huarte, M.; Yaghi, N. K.; Lim, H.; Garcia, B. A.; Brizuela, L.; Zhao, K.; Roberts, T. M.; Shi, Y. Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. *Nature* **2010**, *466* (7305), 503−507.
(33) Abramsson, A.; Westman-Brinkmalm, A.; Pannee, J.; Gustavsson, M.; von, O. M.; Blennow, K.; Brinkmalm, G.; Kettunen, P.; Zetterberg, H. Proteomics profiling of single organs from individual adult zebrafish. *Zebrafish* **2010**, *7* (2), 161−168.
(34) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, (17).
(35) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I., iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10*, (12).
(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.* **2011**, *34* (2), 225−232.
(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 Bioinf.* **2010**, *3* (4), 135−142.
(38) Lucitt, M. B.; Price, T. S.; Pizarro, A.; Wu, W.; Yocum, A. K.; Seiler, C.; Pack, M. A.; Blair, I. A.; Fitzgerald, G. A.; Grosser, T. Analysis of the zebrafish proteome during embryonic development. *Mol. Cell. Proteomics.* **2008**, *7* (5), 981−994.
(39) Antberg, L.; Cifani, P.; Sandin, M.; Levander, F.; James, P. Critical Comparison of Multidimensional Separation Methods for Increasing Protein Expression Coverage. *J. Proteome Res.* **2012**, *11* (5), 2644−2652.