Published in final edited form as:

Proteomics. 2015 August ; 15(15): 2678–2690. doi:10.1002/pmic.201400606.

Machine learning reveals sex-specific 17 β -estradiol-responsive expression patterns in white perch (*Morone americana*) plasma proteins

Justin Schilling¹, Angelito I. Nepomuceno², Antonio Planchart^{3,4}, Jeffrey A. Yoder^{4,5,6}, Robert M. Kelly⁷, David C. Muddiman^{2,6}, Harry V. Daniels¹, Naoshi Hiramatsu⁸, and Benjamin J. Reading¹

¹Department of Applied Ecology, North Carolina State University, Raleigh, NC, USA

²W. M. Keck FTMS Laboratory for Human Health Research, Department of Chemistry, North Carolina State University, Raleigh, NC, USA

³Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA

⁴Center for Human Health and the Environment, North Carolina State University, Raleigh, NC, USA

⁵Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, NC, USA

⁶Center for Comparative Medicine and Translational Research, North Carolina State University, Raleigh, NC, USA

⁷Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC, USA

⁸Faculty of Fisheries Sciences, Hokkaido University, Hakodate, Hokkaido, Japan

Abstract

With growing abundance and awareness of endocrine disrupting compounds (EDCs) in the environment, there is a need for accurate and reliable detection of EDC exposure. Our objective in the present study was to observe differences within and between the global plasma proteomes of sexually mature male and female white perch (*Morone americana*) before (Initial Control, IC) and after 17 β -estradiol (E₂) induction. Semiquantitative nanoLC-MS/MS data were analyzed by machine learning support vector machines (SVMs) and by two-way ANOVA. By ANOVA, the expression levels of 44, 77, and 57 proteins varied significantly by gender, treatment, and the interaction of gender and treatment, respectively. SVMs perfectly classified male and female perch IC and E₂-induced plasma samples using the protein expression data. E₂-induced male and female perch plasma proteomes contained significantly higher levels of the yolk precursors vitellogenin Aa and Ab (VtgAa, VtgAb), as well as latrophilin and seven transmembrane domain-containing

Correspondence: Dr. Benjamin J. Reading, 127 David Clark Laboratories, Department of Applied Ecology, North Carolina State University, Raleigh, NC 27695-7617, USA, **Fax:** +1-919-515-5327, bjreadin@ncsu.edu.

Additional supporting information may be found in the online version of this article at the publisher's web-site

The authors have declared no conflict of interest.

protein 1 (Eltd1) and kininogen 1 (Kng1). This is the first report that Eltd1 and Kng1 may be E₂-responsive proteins in fishes and therefore may be useful indicators of estrogen induction.

Keywords

Animal proteomics; Endocrine disrupting compounds; Machine learning; Plasma; Semiquantitative proteomics; Support vector machines

1 Introduction

There is growing concern over the increasing prevalence of natural and anthropogenic hormone mimics and antagonists present in waterways around the world. Known as environmental endocrine-disrupting compounds (EDCs), these pollutants can alter gene expression and adversely affect sexual development and reproductive function in a wide range of fishes and other aquatic vertebrate species [1]. EDCs can mimic endogenous hormones and affect normal patterns of hormone synthesis and metabolism as well as hormone receptor expression and function [2,3]. The endogenous female hormones estradiol, estrone, and estriol play central roles in regulating vertebrate reproductive processes. Of these, 17 β -estradiol (E₂) is the most biologically active form of estrogen synthesized naturally in vertebrate ovaries and its synthetic derivative, 17 α -ethynylestradiol (EE₂), is the primary form used in oral contraceptives and hormone replacement therapies for humans [4–7]. After their excretion by humans, estrogenic steroids often persist after sewage treatment and are detectable in municipal wastewater effluents at levels that could pose health risks to aquatic vertebrates and to humans as well [8–10].

Aquatic vertebrates are good models for EDC exposure because industrial and municipal effluents that drain into waterways they inhabit often contain EDCs [11]. White perch (*Morone americana*) are native to fresh and brackish waters throughout Northeastern North America and thus are a good model for studying the effects of exposure to EDCs and other environmental contaminants in fishes found in a wide variety of habitats [12–14]. Current strategies for monitoring EDC exposure in fishes rely heavily on recommendations put forward by the Organization for Economic Cooperation and Development (OECD; Paris, France) in 2002 [15, 16]. In the past, such studies/strategies have relied primarily on identification and use of biomarkers of EDC exposure, such as the presence of the female yolk precursors, vitellogenins, in the blood of immature fishes or sexually mature males exposed to estrogen mimicking EDCs [17–19]. Other female specific plasma proteins, such as choriogenins, also have been successfully used in this regard [20]. These particular biomarker approaches offer simple and effective solutions for the detection of estrogenic EDC exposure in populations of aquatic vertebrates. However, they have limitations. For example, it is difficult to discern the exposure impact of sexually mature female fishes to estrogen mimicking EDCs based on the presence of vitellogenins in the blood as females typically produce these yolk precursors at high levels during ovarian growth [18, 21–23]. Additionally, proper biomarker characterization is required prior to any assay implementation. In the case of fish vitellogenins, this is often complex as there are several forms and they exhibit different sensitivities to estrogen induction [19, 24, 25]. Therefore,

proper use and interpretation of vitellogenins as biomarkers of estrogen or estrogenic EDC exposure require thorough characterization of the yolk protein precursors in a particular fish or aquatic species.

Due to recent computational advances in data mining, the power of machine learning has been realized for use in the analysis of transcriptomics and proteomics data. While classical linear statistical analysis methods require a sample size $N = n + 1$ where n is the number of proteins measured to fully assess the univariates [26], machine learning works by identifying patterns in complex datasets and it may be used to classify biological samples by treatment based on the expression patterns of many hundreds or thousands of genes or proteins. We consider such machine learning evaluations to provide transcriptomic or proteomic “fingerprints”. As examples, artificial neural networks (ANNs) are a type of machine learning that have been used to evaluate environmental perturbations of oysters [26, 27] and to predict egg quality in fish [28] using transcriptomics data. Recently, another type of machine learning called support vector machines (SVMs) has been used to evaluate EDC effects measured using microarray gene expression analysis of zebrafish (*Danio rerio*) [29]. We also have used SVMs to accurately classify the proteomes in different growth stages of striped bass (*Morone saxatilis*) ovaries [30] and to differentiate between white perch ovary cell fractions [22]. Thus, proteomics methods paired with pattern recognition by machine learning potentially offer superior alternatives for evaluation of hormone or EDC responses compared to prior biomarker-based approaches.

Our objective in the present study was to observe differences within and between the global plasma proteomes of mature male and female white perch before and after E₂ induction. As such, we performed three technical replicate nanoLC-MS/MS analyses on each of four groups of biological sample pools: 1) Initial Control (IC) male plasma, 2) IC female plasma, 3) E₂-induced male plasma, and 4) E₂-induced female plasma. Here, we report that using SVMs we were able to identify: 1) sex-specific differences between the IC plasma proteomes of mature male and female white perch, 2) differences between the plasma proteomes of IC white perch and perch induced with E₂, and 3) sex-specific differences between the plasma proteomes of mature male and female white perch induced with E₂.

2 Materials and methods

2.1 Sample collection and preparation

Three year old adult white perch (body weight 127.63 ± 33.02 g; total length 207.43 ± 17.3 mm; all values listed as average $\pm$ standard deviation) were reared in recirculating aquaculture systems under controlled photothermal conditions at the North Carolina State University Aquaculture Facility (Raleigh, NC) [31, 32]. Water temperature was maintained between 10–12°C during the course of the study and for 30 days prior to induction with E₂. All animals were handled according to approved institutional animal care and use committee (IACUC) protocols. Sexually mature female and male fish (N = 3 each) were anesthetized with Fiquel MS-222 (Argent Chemical Laboratories, Redmond, WA), and IC blood was collected from the caudal vein by heparinized needle and syringe. Blood was transferred into 12×75 mm tubes containing 25 μ L of aprotinin solution (40 TIU/mL aprotinin in 0.9% NaCl) (all chemicals were purchased from Sigma Aldrich, St. Louis, MO unless otherwise

stated). Samples were centrifuged and the plasma was collected and stored at -80°C until use. Sexually mature male ($N = 15$) and female ($N = 13$) white perch were injected with 0.1 mL of 5mg/mL E_2 dissolved in propylene glycol, per 100 g body weight. Additional E_2 injections were administered at seven and fourteen days after the initial injection based on established protocols [18,21,33]. This induction procedure was based on the finding that estrogen administration to male oviparous vertebrates produces lasting changes in vitellogenin expression patterns, termed a cellular “memory effect”, whereby re-stimulation with estrogen after an initial dosage drives far greater expression levels of vitellogenins [34]. Final bleeds were taken 16 days following the initial injection. E_2 -induced blood samples were centrifuged and the plasma was collected and stored at -80°C until use.

2.2 Filter-aided sample preparation and nanoLC-MS/MS

Plasma samples were thawed and diluted with Tris-buffered saline (20 mM Tris-HCL pH 8.0, 150 mM NaCl, and 2 mM CaCl_2) to final protein concentrations of 5 mg/mL as determined by Bradford assay [35,36]. Four biological sample pools were prepared from the plasmas as follows [37]: (1) IC male plasma ($N = 3$ fish), (2) IC female plasma ($N = 3$ fish), (3) E_2 -induced male plasma ($N = 15$ fish), and (4) E_2 -induced female plasma ($N = 13$ fish). Each of these four pooled plasma samples was separately prepared for tandem mass spectrometry using a modified filter-aided sample preparation (FASP) protocol [22, 38]. DTT was used to reduce disulfide bonds by incubating samples with 3 μL of 50 mM DTT per 30 μL (100 μg) of sample. Samples were incubated with DTT for 30 min at 56°C . Proteins were denatured by adding 200 μL of 8 M urea in 0.1 M Tris-HCl pH 8.5 to each sample. Each sample was then transferred onto a Vivacon 500 30 kDa MW cutoff filter (Sartorius Stedim Biotech, Goettingen, Germany) and centrifuged at $14\,000 \times g$ for 15 min at 21°C . To ensure denaturation, an additional 200 μL of 8 M urea in 0.1M Tris-HCl pH 8.5 was added to each sample before centrifuging again as above. Each sample was on-filter alkylated by adding 100 μL of 50 mM iodoacetamide prepared in 8 M urea atop each filter. Samples were incubated in the dark at room temperature for 20 min and then centrifuged at $14\,000 \times g$ for 10 min at 21°C . Each filter was washed three times with 100 μL of 8M urea by centrifugation for 10 min at $14\,000 \times g$ followed by three washes with 100 μL of 0.1 M Tris pH 7.5. Each sample was then placed in a new centrifuge tube and modified trypsin freshly prepared in 0.1 M Tris pH 7.5 was added to each sample at an enzyme to protein ratio of 1:100. Following overnight digestion at 37°C , tryptic peptides were collected by centrifugation at $14\,000 \times g$ for 10 min at 21°C .

All nanoReversed phase chromatography and MS/MS methods were identical to those described previously [22]. Peptides were separated using a Thermo Scientific EASY nLC II system (Thermo Scientific, San Jose, CA) in line with a cHiPLC nanoflex system (AB Sciex, Framingham, MA). A ChromXP C18-CL 3 μm trap column and a ChromXP C18-CL 75 $\mu\text{m} \times 15$ cm analytical column were used for nanoLC peptide separation in a vented column configuration [39] coupled to a quadrupole orbitrap mass spectrometer (Q-Exactive, Thermo Scientific, San Jose, CA) [40]. Each biological sample pool was analyzed in triplicate.

All spectra were collected in profile mode. Mass resolving power of the precursor peptide ions were acquired at $70k_{FWHM}$ at $m/z = 200$. The automatic gain control (AGC) target for MS acquisitions was set to $1E6$ with a maximum ion injection time of 30 ms. The scan range was set from 400 to $1600m/z$. Tandem MS spectra were acquired with $17.5k_{FWHM}$ resolving power at $m/z = 200$. The AGC target for MS/MS acquisition were set to $2E4$ with a maximum injection time of 120 msec. Data dependent acquisition was performed on the top 12 candidates with dynamic exclusion time set to 60 seconds. An underfill ratio of 1.0% was used as a threshold for the candidate peptides selected for MS/MS with an isolation width of $2.0 m/z$. A normalized collision energy (NCE) of 27 was used for fragmentation for all peptides. Microscans were set to 1 msec for both MS and MS/MS. Charge state screening was enabled. Unassigned and 1+ charge states were rejected from MS/MS isolation and activation. The ambient ion used for lock mass was $m/z = 445.12003$.

2.3 Protein identifications and semi-quantitative spectral counting

Chromatogram files (.RAW) were processed into a peak list format (.MGF) using Proteome Discoverer (Thermo Scientific, San Jose, CA) and batch searched using MASCOT [41] (Matrix Science, Boston, MA) against 21829 open reading frames (ORFs) predicted from the striped bass genome sequence assembly (Reading et al., *unpublished data*) translated in all six ORFs using AUGUSTUS [42]. Polypeptide sequences for human keratins and porcine trypsin were included in the database. A target-reverse protein database was made to determine false discovery rate. MASCOT data search parameters consisted of carbamidomethyl (C) as a fixed modification and carbamyl (K, N-term), deamidation (NQ), oxidation (M), and phospho (STY) as variable modifications. The MASCOT precursor ion search tolerance was 5 ppm and the fragment ion tolerance was ± 0.02 Da. ProteoIQ was used to filter protein identifications using a 1% false discovery rate [43–45].

Semiquantitative proteomics analysis was carried out using normalized tryptic peptides of each protein (normalized spectral counts). Normalized spectral counts (NSpC) were calculated using the total spectral counts within the replicates of each sample and followed with the maximum spectral counts obtained from each IC and E_2 -induced plasma replicate analysis for each of the genders. These NSpC for each of the three technical replicates per biological sample pool were exported from ProteoIQ and transformed to account for zero values [$\log_{10}(y + 1)$, where $y = \text{NSpC}$] [46]. Proteins with > 0.25 average NSpC (NSpC_{avg}) across all twelve technical replicates (three technical replicates per treatment) were analyzed by two-way ANOVA [47]. Predicted striped bass proteins corresponding to the white perch peptides identified by MASCOT were analyzed by BLAST at the National Center for Biotechnology Information (NCBI) and annotated.

2.4 Western blotting

Polyclonal antisera were raised against the three forms of white perch vitellogenin using two unique synthetic peptides from the lipovitellin heavy chain per target protein (Gene-Tel Laboratories LLC, Madison, WI). In order to provide an orthogonal confirmation of vitellogenin expression, Western blotting of male white perch IC and E_2 -induced plasmas was performed using these antibodies [21, 23, 48].

2.5 Data analysis

Comparisons of protein expression levels between samples were evaluated by two-way ANOVA ($\alpha = 0.05$) of $\log_{10}(y + 1)$ transformed NSpC values using the general linear model (GLM) procedure in the SAS/STAT software (version 9.3, SAS Institute Inc., Cary, NC, USA). The two-way ANOVA was used for the comparisons of: 1) treatment (IC versus E₂-induced plasmas), 2) gender (IC and E₂-induced male plasma versus IC and E₂-induced female plasma, and 3) the interaction of treatment and gender (IC male and female by E₂-induced male and female). We employed two independent methods of α correction following two-way ANOVA: 1) standard Bonferroni correction and 2) Benjamini-Hochberg method [22]. ANOVA residual values were input for modulated modularity clustering (MMC) analysis performed using Pearson correlation coefficient [49]. MMC graphs were generated using residual values from 1) the entire dataset (all four groups) and for the following specific pair-wise comparisons of groups: 2) IC male and IC female, 3) IC male and E₂-induced male, 4) IC female and E₂-induced female, and 5) E₂-induced male and E₂-induced female.

For each of the MMC graphs, the DAVID Functional Classification Tool [50] was used to identify Gene Ontology enrichment of proteins within all modules. Default parameters for DAVID were used, and approved gene abbreviations for all white perch proteins were manually collected from the NCBI or GeneCards using approved gene names for zebrafish [51].

We used WEKA (version 3.6.7, www.cs.waikato.ac.nz/ml/weka/) to perform the sequential minimal optimization (SMO) support vector machines (SVMs) machine learning classifier algorithm [52]. The SVMs were used to classify the $\log_{10}(y + 1)$ transformed NSpC semiquantitative protein expression data as IC male plasma, E₂-induced male plasma, IC female plasma, or E₂-induced female plasma using the complete perch plasma protein expression dataset and no preprocessing step. Additionally, a *k*-means clustering algorithm was included as a preprocessing step prior to performing SVMs in the analyses of the complete protein expression dataset. Data were clustered in *k* = 3 clusters (where each of the clusters represented either IC male plasma, IC female plasma, or E₂-induced plasmas of both genders) and *k* = 2 clusters (where each of the clusters represented E₂-induced male plasma or E₂-induced female plasma). We employed two cross-validation strategies for the comparisons to estimate classifier performance as follows: 1) a percentage split whereby 66% of the data were randomly selected and used for training and the remaining 33% of the data were input as a cross-validation and 2) a stratified hold-out with *k*-fold where one fold was used for cross-validation and *k*–1-fold of the randomly reordered dataset were used for training. A 10-fold stratified hold-out was used for SVMs cross-validation when the samples were classified as IC male plasma, E₂-induced male plasma, IC female plasma, or E₂-induced female plasma using the complete protein expression dataset and when the *k*-means algorithm clustered the dataset into three clusters (IC male plasma, IC female plasma, or E₂-induced plasmas of both genders). A *k*-fold stratified hold-out cross-validation was not performed for comparison of E₂-induced male and E₂-induced female plasmas that were clustered into only two groups by *k*-means preprocessing. In all cross-validations, the SVMs

were naive to the hold-out data and percent correct assignment was used to evaluate classifier performance.

All data classes were properly represented in the machine learning training and cross-validation datasets. The performance of the SVMs was evaluated as a percent of correct classification during the cross-validations. Additionally, the Kappa statistic and area under the receiver operating characteristic curve (AUROC) of each model also were reported. As a negative control of machine learning, the NSpC values for the whole dataset were randomly reordered ten-times and each was entered into SVMs as described above, with or without *k*-means preprocessing, as appropriate.

3 Results

3.1 Tandem mass spectrometry

A total of 104 unique proteins were detected by tandem mass spectrometry in the white perch plasma proteomes (Supplemental Table 1). A list of the complete nucleotide sequences for open reading frames encoding these proteins, and found in the striped bass genome sequence are included as Supplemental Table 2. Of these proteins, 94 were detected with NSpC_{avg} ≥ 0.25 (Table 1). All 58 proteins detected in IC male plasma were also detected in IC female plasma, however, 39 proteins were found in IC female plasma but not IC male plasma (Fig. 1A). A comparison of proteins detected in IC male and E₂-induced male plasmas revealed 26 proteins exclusive to IC male plasma, 15 exclusive to E₂-induced male plasma, and 35 found in both (Fig. 1B). Similarly, comparing proteins detected in IC female and E₂-induced female plasmas detected 48 proteins exclusive to IC female plasma, six exclusive to E₂-induced female plasma, and 49 found in both (Fig. 1C). Lastly, comparing proteins detected in E₂-induced male and E₂-induced female plasmas revealed two proteins exclusive to E₂-induced male plasma, seven exclusive to E₂-induced female plasma, and 48 found in both (Fig. 1D).

3.2 Western blotting

Western blotting confirmed that none of the three forms of vitellogenin were expressed in IC male white perch plasma (Fig. 2). After induction with E₂, however, expression of all three forms of white perch vitellogenin (VtgAa, VtgAb, and VtgC) was evident by Western blotting (Fig. 2).

3.3 Inferential statistics

Three comparisons of white perch plasma protein expression were made by two-way ANOVA using proteins with NSpC_{avg} ≥ 0.25: 1) Treatment (IC male and female versus E₂-induced male and female), 2) Gender (IC and E₂-induced male versus IC and E₂-induced female), and 3) Treatment by Gender (IC male and female by E₂-induced male and female) (Table 2). When protein expression levels of IC male and female versus E₂-induced male and female plasma were compared by two-way ANOVA after α correction, 57 of the 94 (60.6%, p = 0.000531, Bonferroni) and 77 of the 94 (81.9%, p = 0.034, Benjamini–Hochberg) proteins detected were found to vary significantly by gender. When protein expression levels between IC and E₂-induced male versus IC and E₂-induced female plasma

were compared as above, 26 of the 68 (38.2%, $p = 0.000735$, Bonferroni) and 44 of the 68 (64.7%, $p = 0.029$, Benjamini–Hochberg) proteins detected were found to vary significantly by gender. When protein expression levels between IC male and female by E₂-induced male and female plasma were compared as above, 32 of the 93 (34.4%, $p = 0.0003$, Bonferroni) and 57 of the 93 (61.2%, $p = 0.0263$, Benjamini–Hochberg) proteins detected were found to vary significantly by treatment and gender. Differentially expressed proteins are listed in Supporting Information (Supplemental Table 3).

3.4 MMC and DAVID analysis

The entire white perch plasma protein expression dataset also was analyzed by two-way ANOVA and the resulting residual values for each protein were input for MMC. A table of these residual values is included in Supplemental Materials (Supplemental Table 4).

ANOVA residual values for predicted striped bass proteins corresponding to white perch plasma proteins detected by mass spectrometry were organized on the basis of covariable expression using MMC. MMC graphs were generated representing proteins from the entire mass spectrometry dataset (Fig. 3A), IC male and IC female (Fig. 3B), IC male and E₂-induced male (Fig. 3C), IC female and E₂-induced female (Fig. 3D), and E₂-induced male and E₂-induced female (Fig. 3E). The output from each MMC analysis is included in Supporting Information (Supplemental Table 5).

When all residual values from two-way ANOVA were input into MMC, a total of 15 modules were generated (Fig. 3A). Analysis by DAVID revealed three proteins (F2, Tfa, Eno2) in module 14 with significant GO enrichment for metal ion binding. The MMC graph of two-way ANOVA residual values from IC male and IC female contains 27 modules, the largest of which (module 15) contains 11 proteins (Fig. 3B). Module 15 was significantly enriched for GO terms associated with glycolysis (Aldoa, Eno2, Gapdh). The MMC graph of IC male and E₂ male residuals contains six modules, the largest of which (module 2) contains 23 proteins, four of which (Cp, Eno2, Rock2b, zgc:163025) are involved in metal ion binding (Fig. 2C). Module 5 was significantly enriched for GO terms associated with glycolysis (Aldoa, Eno2, Tpi1b) (Fig. 3C). MMC generated eight modules when residuals from IC female and E₂ female were input (Fig. 3D). Modules 2 and 3 are the largest, each containing 16 proteins, yet neither module shows significant GO-term enrichment. Modules 5 and 7, however, show significant GO-term enrichment for proteins related to SERPIN (serine protease inhibitor) activity (Agt, Serpina1, Serpind1) and blood coagulation (F2, Fgg, Plg), respectively. When E₂ male and E₂ female residuals were input into MMC, a total of 10 modules were generated, one of which (module 9) shows significant GO-term enrichment for proteins related to extracellular space, signal peptide, and secretion (Apoa2, F2, Serpina1) (Fig. 3E).

Supplemental Figure 1 depicts glycolysis/gluconeogenesis KEGG Pathway analysis of the entire white perch plasma protein dataset as measured by mass spectrometry. Detected white perch plasma proteins are indicated by a red box.

3.5 Machine learning

When $\log_{10}(y + 1)$ transformed NSpC values from the complete white perch plasma proteome dataset were input directly into SVMs with no preprocessing step, all four groups (IC male, E₂-induced male, IC female, and E₂-induced female) were classified with perfect accuracy (100% correct classifier assignment) using 10-fold or 66%-split cross-validations (Table 3). The Kappa statistics and AUROCs were both optimal at 1.00. When the same dataset was preprocessed by *k*-means clustering, the algorithm was unable to correctly group the entire dataset into four clusters. Instead, the *k*-means algorithm clustered the entire white perch plasma protein dataset into three clusters that included IC male plasma, IC female plasma, or E₂-induced plasmas of both sexes. This indicates that the *k*-means algorithm was not sensitive enough to differentiate between the E₂-induced sex-specific responses in protein expression. Following *k*-means clustering, the SVMs were again able to classify all data with perfect accuracy using both cross-validation strategies (Table 3). When only the E₂-induced male and E₂-induced female protein expression data were preprocessed by *k*-means into two clusters and then fed into SVMs, the two groups were classified with perfect accuracy using the 66%-split cross-validation (100% correct assignment) (Table 3). With just two clusters, the 10-fold cross validation was not possible for this comparison.

Ten randomized datasets were generated for each of the IC and E₂-induced comparisons to serve as negative controls for machine learning; the resulting SVM classifier performance is reported in Table 4. The correct classification by SVMs of the randomized datasets ranged from 5.0 to 40.0%, which was slightly less than that predicted from random assignment into the respective four, (25.0%), three (33.3%), and two groups (50.0%). Kappa statistics ranged from -0.156 to 0.022 and AUROC ranged from 0.210 to 0.707. A Kappa statistic of less than 0.200 is considered poor performance [53]. An AUROC of 0.500 is comparable to random assignment [52]. A negative Kappa statistic and/or an AUROC of less than 0.500 indicate that machine learning classifier performance is worse than that predicted by random assignment.

4 Discussion

The SVMs were able to distinguish the plasma proteomes of IC white perch from that of E₂-induced fish and this approach should be amenable to assessing the effects of E₂ on white perch and possibly other aquatic vertebrates. These findings are similar to a previous report where SVMs were used to classify microarray gene expression data from zebrafish exposed to different types of EDCs [29]. However, these authors report inconsistent classifier performance, which they attribute to varying data complexity related to small biological sample size. These authors state that the classifiers used less than 100 input attributes (i.e., gene transcripts). Here, the SVMs classifier performance was 100% correct when the expression of 104 unique proteins was considered, further supporting the contention that a considerable number of inputs are required for optimal classification. It also has been shown that the cross-validation performance of ANNs when classifying microarray gene expression data decreases dramatically when the input attributes are restricted from 250 to 100 genes [28] as sample size and attribute number are both important considerations for successful machine learning performance.

In the future, similar studies could be conducted whereby white perch or other sentinel species of fishes are exposed to a panel of different estrogenic EDCs and/or mixtures tested to determine whether machine learning can classify the treatments or exposures. For instance, different doses of EDCs may elicit different plasma proteome responses and combinations of EDCs may have synergistic effects on the responses of plasma proteomes. These findings set the stage for using machine learning to classify exposure of aquatic animals to complex mixtures of EDCs. Since wastewater, for instance, may contain a mixture of different EDCs, it becomes difficult to implement the practical identification of particular EDC(s) to which an organism has been exposed. Machine learning may offer a possible solution to this problem as pattern recognition could be used to identify the particular gene or protein expression signatures diagnostic of exposure to one or more particular EDCs. Further research will be required to validate this possibility. While we only evaluated the effects of E₂ on white perch, importantly SVMs did detect subtle sex-specific responses to E₂ in the plasma proteomes, which to our knowledge has not been demonstrated at this scale in any other fish. Such resolution of machine learning suggests that it may be possible to decipher gene or protein expression differences between organisms treated with different EDCs or different concentrations of the same EDC in future studies.

SVMs classification of sexually mature male and female white perch plasma proteomes proves that these methods may be employed to identify the gender of fishes from blood and/or plasma samples using a semiquantitative proteomics approach. This also opens the possibility that these or similar methods may be used for gender identification of sexually immature fish as well. Hidden Markov models have been used to identify age- and gender-related differences in plasma metabolomics data of human children [54]. In practice, gender identification of sexually immature fish usually requires lethal sampling and gonad histology. Therefore, the machine learning and proteomics procedures described here would be beneficial for the study of stock structures of imperiled fishes, where lethal sampling is not a desirable approach.

Blood plasma contains an extremely large number of proteins across an even larger dynamic range of abundances. As such, plasma has been regarded as an especially difficult sample for proteomics analysis [55] and here we detected just 94 distinct proteins with $NSpC_{avg} = 0.25$. While the white perch blood samples used in this study were centrifuged as described in order to isolate the plasma and later subjected to FASP, no additional fractionation methods were utilized prior to analysis by mass spectrometry. Following E₂ induction, the increase in abundance of a few proteins, particularly VtgAb, likely impaired detection of less abundant proteins and thus limited the depth of protein detection. In spite of this, it is notable that SVMs were still able to perfectly distinguish between male and female plasma samples, both before and after E₂ induction. The spike in abundance of a few predominant plasma proteins following E₂ induction likely affected the outcomes of the plasma protein comparisons as analyzed by ANOVA. Specifically, many of the proteins that were detected at low levels in IC male and IC female plasma were not detected following E₂ induction and were found to vary significantly by two-way ANOVA following Bonferroni and/or Benjamini–Hochberg correction (Supplemental Tables 1 and 2).

When the entire white perch plasma protein dataset is examined, many proteins directly involved in or closely associated with glycolysis are generally apparent (e.g., Aldoa, Eno1a, Eno2, Eno3, Gapdh, Gapdhs, Pgam2, Tpi1b). Consistent with the findings in the present study, estrogens have been shown to have strong anabolic effects in vertebrates [56]. A key enzyme in muscle glycogenolysis, Pygma (Contig440-0.0) expression levels varied significantly between IC and E₂-induced perch plasma (p = 0.034).

Additionally, other factors generally representative of metabolic state were apparent in the plasma proteomes. Expression levels of Creatine kinase, muscle b (Ckmb, Contig15929-0.2), a metabolic enzyme that plays a key role in energy homeostasis [57], varied significantly between IC and E₂-induced plasma by two-way ANOVA (p = 0.0002) and by the interaction of treatment and gender (p = 0.0002) (Supplemental Table 2).

Expression levels of Eltd1 (Contig838-0.0), an epidermal growth factor (EGF) seven-transmembrane receptor, varied significantly between male and female plasma (p < 0.0001) and by the interaction of treatment and gender (p = 0.0026) (Supplemental Table 3). We believe this study to be the first to suggest that Eltd1 may be an E₂-responsive protein in a fish. Interestingly, the *Eltd1* locus is adjacent to one of the vitellogenin gene loci (*vtg1*) in the genome sequences of platypus (*Ornithorhynchus anatinus*), chicken (*Gallus gallus*), and teleost fishes [58]. The biological relevance of this association is unclear at this time, however vitellogenins are the predominant egg yolk precursors in oviparous vertebrates and their hepatic synthesis is primarily responsive to E₂ [58–65]. While males possess vitellogenin genes, expression will occur only if induced by an estrogenic compound. Accordingly, the two main forms of vitellogenin in white perch, VtgAa and VtgAb, were not detected in IC male plasma by MS/MS (Supplemental Table 1). While the third form and incomplete type vitellogenin, VtgC, was detected by Western blotting in male plasma following E₂ induction (Fig. 2), VtgC was not detected in either male or female plasma by nanoLC-MS/MS as it is known to be expressed at significantly lower levels than VtgAa and VtgAb in white perch [22]. Following induction with E₂, however, VtgAa and VtgAb levels increased markedly in male plasma (p < 0.001) (Supplemental Table 1) and VtgAb was the most abundant protein detected following induction with E₂ in both male and female plasmas (Supplemental Table 1). Two putative VtgAb proteins were detected using the AUGUSTUS predicted striped bass ORFs as a reference database (Contig26928-0.0 and Contig26928-0.1) in IC female plasma and E₂-induced male and female plasma (Supplemental Tables 1 and 2). While multiple *Vtg* genes have been found in *D. rerio* [65], subsequent curation of the striped bass genome consolidated both of the VtgAb contigs above to the same ORF (Reading et al., *unpublished data*). Detection of VtgAb in the plasma of IC female perch is consistent with the stage of their reproductive cycle at the time of their sampling (during vitellogenesis) based on photothermal conditioning [32].

Members of three proteinase inhibitor families were detected in perch plasma by tandem mass spectrometry: 1) Serpina1 and Serpind1 (Serpina1, Serpind1, serpin family); 2) Fetuin B (Fetub, fetuin family), and 3) Kininogen 1 (Kng1, Type 3 cystatin family). Serpins are primarily found in blood plasma where they help to maintain homeostasis by regulating proteolytic enzymes [66]. Two putative Serpina1 proteins and one putative Serpind1 protein were detected by tandem mass spectrometry in white perch plasma of both genders

(Contig12216-0.0, Contig1392-0.0, and Contig10856-0.0, respectively; see Supplemental Table 1). Expression of the two detected Serpina1 proteins varied significantly between IC plasma and E₂-induced plasma ($p < 0.0001$ for Contig1392-0.0 and Contig12216-0.0) (Supplemental Table 3). Expression levels of the Serpind1 protein varied significantly between IC plasma and E₂-induced plasma ($p < 0.0001$); however, Serpina1 (Contig1392-0.0) and Serpind1 were only detected in IC male and IC female plasmas (Supplemental Table 1). Serpina1 (Contig12216-0.0) was detected in all plasma samples, however, as with both putative Serpina1 proteins, the number of peptides detected decreased in male and female plasmas following E₂-induction. In addition to bovine uterine SERPINA14, a number of serpins have been shown to be estrogen-responsive in the laying hen, specifically SERPINB3, SERPINB11, and SERPINB12, which play a key role in the secretion and/or deposition of egg white proteins [66–69].

Fetuin-B (Fetub) is a liver-derived cysteine protease inhibitor with roles in a diverse array of cellular processes including fatty acid metabolism, osteogenesis and bone resorption, regulation of the insulin and hepatocyte growth factor receptors, response to systemic inflammation, and prevention of premature hardening of the zona pellucida surrounding mammalian oocytes [70]. Two putative forms of Fetub (Contig13043-0.0 and Contig13357-0.1) were detected. Contig13357-0.1 was not detected in IC male plasma but was detected in E₂-induced male plasma as well as in both IC and E₂-induced female plasma and its protein levels decreased after E₂ induction. When analyzed by two-way ANOVA for the interaction between treatment and gender, Contig13357-0.1 varied significantly ($p = 0.0017$) (Supplemental Table 3). It has been previously shown that E₂ induction in tilapia (*Oreochromis mossambicus*) suppresses the growth hormone/insulin-like growth factor 1 axis [24] and it would be interesting to further investigate the putative role of Fetub as a potential E₂-induced regulator of growth factors in this process.

Kininogen 1 (Kng1) is a cysteine proteinase inhibitor that displays a wide range of functions including blood coagulation and bradykinin transport [71]. In this study, Kng1 (Contig9596-0.0) was detected at significantly higher levels in both male ($p = 0.0013$) and female plasma following E₂ induction ($p < 0.0001$) compared to IC male and IC female plasma, respectively (Supplemental Tables 1 and 3). While kininogens have previously been isolated and characterized from a number of fish species [71–75] and shown to be estrogen responsive in the rat (*Rattus rattus*) [76], we believe this study to be the first to suggest that Kng1 may be an E₂-responsive protein in a fish.

5 Concluding remarks

Following induction with E₂, both male and female white perch plasma proteomes contained significantly higher levels of VtgAa, VtgAb, Eltd1, and Kng1 than IC plasmas. It appears that E₂ induction generally resulted in an increase in proteins directly or indirectly involved in glycolysis in the plasmas of both male and female perch. SVMs were able to perfectly classify male and female plasma samples as either IC or E₂-induced based on protein expression data as detected by mass spectrometry. Additional studies of the effects of EDCs on vertebrates may benefit from employing SVMs and/or other forms of machine learning in data analysis as these methods appear robust.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was funded in part by the North Carolina Agricultural Research Service, Hatch Project number 06819. The authors would like to thank Brad Ring and John Davis (North Carolina State University Aquaculture Facility) for care and maintenance of the white perch.

Abbreviations

AGC	automatic gain control
ANNs	artificial neural networks
AUROC	area under the receiver operating characteristic curve
BFW	Bonferroni family wise error
BH	Benjamini-Hochberg false discovery rate
CV	10-fold cross-validations
EDCs	endocrine disrupting compounds
E₂	17 β -estradiol
EE₂	17 α -ethynylestradiol
EGF	epidermal growth factor
GO	Gene Ontology
IC	initial control
MMC	modulated modularity clustering
NCE	normalized collision energy
ORF	open reading frame
SVMs	support vector machines

References

1. Tyler CR, Jobling S, Sumpter JP. Endocrine disruption in wildlife: a critical review of the evidence. *Crit. Rev. Toxicol.* 1998; 28:319–361. [PubMed: 9711432]
2. Vos JG, Dybing E, Greim HA, Ladefoged O, Lambré C, et al. Health Effects of Endocrine-Disrupting Chemicals on Wildlife, with Special Reference to the European Situation. *Crit. Rev. Toxicol.* 2000; 30:71–133. [PubMed: 10680769]
3. Schoeters G, Hond Den E, Dhooge W, vanLarebeke N, Leijs M. Endocrine disruptors and abnormalities of pubertal development. *Basic Clin. Pharmacol. Toxicol.* 2008; 102:168–175. [PubMed: 18226071]

4. Kuster M, deAlda MJL, Barceló D. Analysis and distribution of estrogens and progestogens in sewage sludge, soils and sediments. *TrAC*. 2004; 23:790–798.
5. Williams RJ, Johnson AC, Smith JLL, Kanda R. Steroid estrogens profiles along river stretches arising from sewage treatment works discharges. *Environ. Sci. Technol.* 2003; 37:1744–1750. [PubMed: 12775044]
6. Pojana G, Bonfà A, Buseti F, Collarin A, Marcomini A. Estrogenic potential of the Venice, Italy, lagoon waters. *Environ. Toxicol. Chem.* 2004; 23:1874–1880. [PubMed: 15352475]
7. Muller M, Rabenoelina F, Balaguer P, Patureau D, Lemenach K, et al. Chemical and biological analysis of endocrine-disrupting hormones and estrogenic activity in an advanced sewage treatment plant. *Environ. Toxicol. Chem.* 2008; 27:1649–1658. [PubMed: 18315391]
8. Manickum T, John W. Occurrence, fate and environmental risk assessment of endocrine disrupting compounds at the wastewater treatment works in Pietermaritzburg (South Africa). *Sci. Total Environ.* 2014; 468–469:584–597.
9. Esteban S, Gorga M, Petrovic M, González-Alonso S, Barceló D, et al. Analysis and occurrence of endocrine-disrupting compounds and estrogenic activity in the surface waters of Central Spain. *Sci. Total Environ.* 2014; 466–467:939–951.
10. Luo Y, Guo W, Ngo HH, Nghiem LD, Hai FI, et al. A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. *Sci. Total Environ.* 2014; 473–474:619–641.
11. Ying G-G, Kookana RS, Ru Y-J. Occurrence and fate of hormone steroids in the environment. *Environ. Int.* 2002; 28:545–551. [PubMed: 12503920]
12. Barnthouse LW, Glaser D, DeSantis L. Polychlorinated Biphenyls and Hudson River White Perch: Implications for Population-Level Ecological Risk Assessment and Risk Management. *Integr. Environ. Assess. Manag.* 2009; 5:435–444. [PubMed: 20050031]
13. Kamman NC, Burgess NM, Driscoll CT, Simonin HA. Mercury in Freshwater Fish of Northeast North America – A Geographic Perspective Based on Fish Tissue Monitoring Databases. *Ecotoxicology*. 2005; 14:163–180. [PubMed: 15931965]
14. Kavanagh RJ, Balch GC, Liparissis Y, Niimi, Sherry J, et al. Endocrine disruption and altered gonadal development in white perch (*Morone americana*) from the lower Great Lakes region. *Environ. Health Persp.* 2004; 112:898–902.
15. Hecker M, Hollert H, Cooper R, Vinggaard AM, Akahori Y, et al. The OECD validation program of the H295R steroidogenesis assay: Phase 3. Final inter-laboratory validation study. *Environ. Sci. Pollut. Res. Int.* 2011; 18:503–515. [PubMed: 20890769]
16. Gelbke HP, Kayser M, Poole A. OECD test strategies and methods for endocrine disruptors. *Toxicology*. 2004; 205:17–25. [PubMed: 15458786]
17. Kishida M, Specker JL. Vitellogenin in the surface mucus of tilapia (*Oreochromis mossambicus*): Possibility for uptake by the free-swimming embryos. *J. Exp. Zool.* 1994; 268:259–268.
18. Reading BJ, Hiramatsu N, Sullivan CV. Disparate binding of three types of vitellogenin to multiple forms of vitellogenin receptor in white perch. *Biol. Reprod.* 2011; 84:392–399. [PubMed: 20926805]
19. Hiramatsu N, Matsubara T, Fujita T, Sullivan CV, Hara A. Multiple piscine vitellogenins: biomarkers of fish exposure to estrogenic endocrine disruptors in aquatic environments. *Mar. Biol.* 2006; 149:35–47.
20. Hong L, Fujita T, Wada T, Amano, Hiramatsu N, et al. Choriogenin and vitellogenin in red lip mullet (*Chelon haematocheilus*): purification, characterization, and evaluation as potential biomarkers for detecting estrogenic activity. *Comp. Biochem. Phys. C*. 2009; 149:9–17.
21. Reading BJ, Hiramatsu N, Sawaguchi S, Matsubara T, Hara A, et al. Conserved and variant molecular and functional features of multiple egg yolk precursor proteins (vitellogenins) in white perch (*Morone americana*) and other teleosts. *Mar. Biotechnol.* 2009; 11:169–187. [PubMed: 18766402]
22. Schilling J, Nepomuceno A, Schaff JE, Muddiman, Daniels HV, et al. Compartment Proteomics Analysis of White Perch (*Morone americana*) Ovary Using Support Vector Machines. *J. Proteome Res.* 2014; 13:1515–1526. [PubMed: 24494930]

23. Williams VN, Reading BJ, Amano H, Hiramatsu N, Schilling J, et al. Proportional accumulation of yolk proteins derived from multiple vitellogenins is precisely regulated during vitellogenesis in striped bass (*Morone saxatilis*). *J. Exp. Zool.* 2014; 321:301–315.
24. Davis LK, Hiramatsu N, Hiramatsu K, Reading BJ, Matsubara T, et al. Induction of three vitellogenins by 17beta-estradiol with concurrent inhibition of the growth hormone-insulin-like growth factor 1 axis in a euryhaline teleost, the tilapia (*Oreochromis mossambicus*). *Biol. Reprod.* 2007; 77:614–625. [PubMed: 17507692]
25. Tyler CR, van Aerle R, Hutchinson TH. An *in vivo* testing system for endocrine disruptors in fish early life stages using induction of vitellogenin. *Environ. Toxicol. Chem.* 1999; 18:337–347.
26. Chapman RW, Mancia A, Beal M, Veloso A, Rathburn C, et al. A transcriptomic analysis of land-use impacts on the oyster, *Crassostrea virginica*, in the South Atlantic bight. *Mol. Ecol.* 2009; 18:2415–2425. [PubMed: 19457208]
27. Chapman RW, Mancia A, Beal M, Veloso A, Rathburn C, et al. The transcriptomic responses of the eastern oyster, *Crassostrea virginica*, to environmental conditions. *Mol. Ecol.* 2011; 20:1431–1449. [PubMed: 21426432]
28. Chapman RW, Reading BJ, Sullivan CV. Ovary transcriptome profiling via artificial intelligence reveals a transcriptomic fingerprint predicting egg quality in striped bass, *Morone saxatilis*. *PLoS ONE.* 2014; 9:e96818. [PubMed: 24820964]
29. Wang RL, Bencic D, Biales A, Flick R, Lazorchak J, et al. Discovery and validation of gene classifiers for endocrine-disrupting chemicals in zebrafish (*danio rerio*). *BMC Genomics.* 2012; 13:358. [PubMed: 22849515]
30. Reading BJ, Williams VN, Chapman RW, Williams TI, Sullivan CV. Dynamics of the Striped Bass (*Morone saxatilis*) Ovary Proteome Reveal a Complex Network of the Translasome. *J. Proteome Res.* 2013; 12:1691–1699. [PubMed: 23414552]
31. King WV, Berlinsky D, Sullivan CV. Involvement of gonadal steroids in final oocytematuration of white perch (*Morone americana*) and white bass (*M. chrysops*): in vivo and in vitro studies. *Fish Physiol. Biochem.* 1995; 14:489–500. [PubMed: 24197645]
32. Jackson LF, Sullivan CV. Reproduction of White Perch: The Annual Gametogenic Cycle. *T. Am. Fish. Soc.* 1995; 124:563–577.
33. Heppell SA, Jackson LF, Weber GM, Sullivan CV. Enzyme-linked immunosorbent assay (ELISA) of vitellogenin in temperate basses (genus *Morone*): plasma and in vitro analyses. *T. Am. Fish. Soc.* 1999; 128:532–541.
34. Baker HJ, Shapiro DJ. Rapid Accumulation of Vitellogenin Messenger RNA during Secondary Estrogen Stimulation of *Xenopus laevis*. *J Biol. Chem.* 1978; 253:4521–4524. [PubMed: 659432]
35. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* 1976; 72:248–254. [PubMed: 942051]
36. Noble JE., Bailey MJA. Quantitation of Protein. In: Burgess, RR., Deutscher, MP., editors. *Methods in Enzymology*. Elsevier; 2009. p. 73-95.
37. Karp NA, Lilley KS. Design and Analysis Issues in Quantitative Proteomics Studies. *Proteomics.* 2007; 7:42–50. [PubMed: 17893850]
38. Wi niewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. *Nat. Meth.* 2009; 6:359–362.
39. Andrews GL, Shuford CM, Burnett JC, Hawkrigde AM, Muddiman DC. Coupling of a vented column with splitless nanoRPLC-ESI-MS for the improved separation and detection of brain natriuretic peptide-32 and its proteolytic peptides. *J. Chromatogr. B.* 2009; 877:948–954.
40. Michalski A, Damoc E, Hauschild J-P, Lange O, Wieghaus A, et al. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics.* 2011; 10 M111.011015.
41. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.* 1999; 20:3551–3567. [PubMed: 10612281]
42. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 2005; 33:W465–W467. [PubMed: 15980513]

43. Weatherly DB, Atwood JA, Minning TA, Cavola C, Tarleton RL, et al. A Heuristic Method for Assigning a False-discovery Rate for Protein Identifications from Mascot Database Search Results. *Mol. Cell. Proteomics*. 2005; 4:762–772. [PubMed: 15703444]
44. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A Statistical Model for Identifying Proteins by TandemMass Spectrometry. *Anal. Chem*. 2003; 75:4646–4658. [PubMed: 14632076]
45. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem*. 2002; 74:5383–5392. [PubMed: 12403597]
46. Wu X, Tseng CW, Edwards N. HMMatch: peptide identification by spectral matching of tandem mass spectra using hidden Markov models. *J. Comput. Biol*. 2007; 14:1025–1043. [PubMed: 17985986]
47. Gokce E, Shuford CM, Franck WL, Dean RA, Muddiman DC. Evaluation of Normalization Methods on GeLC-MS/MS Label-Free Spectral Counting Data to Correct for Variation during Proteomic Workflows. *J. Am. Soc. Mass Spectrom*. 2011; 22:2199–2208. [PubMed: 21952779]
48. Williams VN, Reading BJ, Hiramatsu N, Amano H, Glassbrook, et al. Multiple vitellogenins and product yolk proteins in striped bass, *Morone saxatilis*: molecular characterization and processing during oocyte growth and maturation. *Fish Physiol. Biochem*. 2014; 40:395–415. [PubMed: 24005815]
49. Stone EA, Ayroles JF. Modulated modularity clustering as an exploratory tool for functional genomic inference. *PLoS Genet*. 2009; 5:e1000479. [PubMed: 19424432]
50. Dennis G Jr, Sherman BT, Hosack DA, Yang J. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*. 2003; 4:R60.1–R60.11.
51. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*. 1998; 14:656–664. [PubMed: 9789091]
52. Witten, IH., Frank, E., Hall, MA. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier; Boston: 2011.
53. McGinn T, Wyer PC, Newman TB, Keitz S, Leipzig R, et al. Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). *Can. Med. Soc. J*. 2004; 171:1369–1373.
54. Nikkilä J, Sysi-Aho M, Ermolov A, Seppänen-Laakso T, et al. Gender-dependent progression of systemic metabolic states in early childhood. *Mol. Syst. Biol*. 2008; 4:197. [PubMed: 18523432]
55. Anderson NL. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. *Molecular & Cellular Proteomics*. 2002; 1:845–867. [PubMed: 12488461]
56. Lange IG, Hartel A, Meyer HHD. Evolution of oestrogen functions in vertebrates. *J. Steroid biochem*. 2002; 83:219–226.
57. Wallimann T, Wyss M, Brdiczka D, Nicolay K, Eppenberger HM. Intracellular compartmentation, structure and function of creatine kinase isoenzymes in tissues with high and fluctuating energy demands: the ‘phosphocreatine circuit’ for cellular energy homeostasis. *Biochem. J*. 1992; 281:21–40. [PubMed: 1731757]
58. Babin PJ. Conservation of a vitellogenin gene cluster in oviparous vertebrates and identification of its traces in the platypus genome. *Gene*. 2008; 413:76–82. [PubMed: 18343608]
59. Tata JR. The expression of the vitellogenin gene. *Cell*. 1976
60. Sumpter JP, Jobling S. Vitellogenesis as a biomarker for estrogenic contamination of the aquatic environment. *Environ. Health Persp*. 1995; 103:173–178.
61. Denslow ND, Chow MC, Kroll KJ, Green L. Vitellogenin as a Biomarker of Exposure for Estrogen or Estrogen Mimics. *Ecotoxicology*. 1999; 8:385–398.
62. Davis LK, Hiramatsu N, Hiramatsu K, Reading BJ, Matsubara T, et al. Induction of Three Vitellogenins by 17beta-Estradiol with Concurrent Inhibition of the Growth Hormone-Insulin-Like Growth Factor 1 Axis in a Euryhaline Teleost, the *Tilapia* (*Oreochromis mossambicus*). 2007; 77:614–625.
63. Hiramatsu N, Matsubara T, Fujita T, Sullivan CV. Multiple piscine vitellogenins: biomarkers of fish exposure to estrogenic endocrine disruptors in aquatic environments. *Mar. Biol*. 2006; 149:35–47.

64. Rose J, Holbech H, Lindholst C, Nørum U, Povlsen A, et al. Vitellogenin induction by 17 β -estradiol and 17 α -ethinylestradiol in male zebrafish (*Danio rerio*). *Comp. Biochem. Phys. C*. 2002; 131:531–539.
65. Wang H, Tan JTT, Emelyanov A, Korzh V, Gong Z. Hepatic and extrahepatic expression of vitellogenin genes in the zebrafish, *Danio rerio*. *Gene*. 2005; 356:91–100. [PubMed: 15979250]
66. Potempa J, Korzus E, Travis J. The serpin superfamily of proteinase inhibitors: structure, function, and regulation. *J. Biol. Chem*. 1994; 269:15957–15960. [PubMed: 8206889]
67. Jo G, Lim W, Bae S-M, Bazer FW, Song G. Avian SERPINB12 expression in the avian oviduct is regulated by estrogen and up-regulated in epithelial cell-derived ovarian carcinomas of laying hens. *PLoS ONE*. 2014; 9:e99792. [PubMed: 25020046]
68. Lim W, Ahn SE, Jeong W, Kim J-H, Kim J, et al. Tissue specific expression and estrogen regulation of SERPINB3 in the chicken oviduct. *Gen. Comp. Endocrinol*. 2012; 175:65–73. [PubMed: 22026959]
69. Ulbrich SE, Frohlich T, Schulke K, Englberger E, Waldschmitt N, et al. Evidence for Estrogen-Dependent Uterine Serpin (SERPINA14) Expression During Estrus in the Bovine Endometrial Glandular Epithelium and Lumen. *Biol. Reprod*. 2009; 81:795–805. [PubMed: 19494250]
70. Dietzel E, Wessling J, Floehr J, Schäfer C, Ensslen S, et al. Fetuin-B, a liver-derived plasma protein is essential for fertilization. *Dev. Cell*. 2013; 25:106–112. [PubMed: 23562279]
71. Ylönen A, Rinne A, Herttuainen J, Bøggwald J, Jarvinen M, et al. Atlantic salmon (*Salmo salar* L.) skin contains a novel kininogen and another cysteine proteinase inhibitor. *Eur. J. Biochem*. 1999; 266:1066–1072. [PubMed: 10583403]
72. Ylönen A, Kalkkinen N, Saarinen J, Bøggwald J, Helin J. Glycosylation analysis of two cysteine proteinase inhibitors from Atlantic salmon skin: di-O-acetylated sialic acids are the major sialic acid species on N-glycans. *Glycobiology*. 2001; 11:523–531. [PubMed: 11447131]
73. Ylönen A, Helin J, Bøggwald J, Jaakola A, Rinne A, et al. Purification and characterization of novel kininogens from spotted wolffish and Atlantic cod. *Eur. J. Biochem*. 2002; 269:2639–2646. [PubMed: 12047371]
74. Zhou L, Li-Ling J, Huang H, Ma F, Li Q. Phylogenetic analysis of vertebrate kininogen genes. *Genomics*. 2008; 91:129–141. [PubMed: 18096361]
75. Zhou L, Liu X, Jin P, Li Q. Cloning of the kininogen gene from *Lampetra japonica* provides insights into its phylogeny in vertebrates. *J. Genet. Genomics*. 2009; 36:109–115. [PubMed: 19232309]
76. Chen LM, Chung P, Chao S, Chao L, Chao J. Differential regulation of kininogen gene expression by estrogen and progesterone *in vivo*. *Biochim. Biophys. Acta*. 1992; 1131:145–151. [PubMed: 1610894]

Figure 1.

Venn diagrams depicting the total number of white perch plasma proteins detected by nanoLC-MS/MS in (A) male (left) and female (right) Initial Control (IC) plasmas, (B) IC male plasma (left) and 17 β -estradiol (E₂)-induced male plasma (right), (C) IC female plasma (left) and E₂-induced female plasma (right), and (D) E₂-induced male (left) and E₂-induced female plasma (right) using the AUGUSTUS predicted open reading frames from the striped bass genome as a reference database.

Figure 2. Western blotting of male white perch IC and E₂-induced plasma using polyclonal antisera raised against two unique synthetic peptides per target protein (VtgAa, VtgAb, and VtgC). No reactivity for any of the vitellogenins was seen in IC plasma, while all three forms of vitellogenin were detected in male white perch plasma following induction with E₂.

Figure 3.

Modulated modularity clustering (MMC) of two-way ANOVA residual values from the white perch plasma protein expression dataset (left) and a table listing MMC module(s) with significant Gene Ontology (GO) enrichment within MMC modules using DAVID (right) for: The entire white perch plasma protein expression data set (A), IC male and IC female white perch plasma protein expression data (B), IC male and E₂-induced male white perch plasma protein expression data (C), IC female and E₂-induced female white perch plasma protein expression data (D), and E₂-induced male and E₂-induced female white perch plasma protein expression data (E). An asterisk after the Contig ID indicates that the protein varied

significantly by at least one two-way ANOVA comparison after Benjamini-Hochberg false discovery rate correction.

Table 1

Number of proteins detected in the initial IC male, IC female, 17 β -estradiol (E₂)-induced male, and E₂-induced female white perch plasma samples by tandem mass spectrometry

Plasma sample	Total proteins identified	Proteins with NSpC _{avg} 0.25
IC	58	58
IC	97	88
E ₂ -induced	50	49
E ₂ -induced	55	55

Proteins with 0.25 average normalized spectral counts (NSpC_{avg}) across all twelve technical replicates (three technical replicates per treatment) were further analyzed by two-way ANOVA.

Table 2

Two-way ANOVA *p*-value range from comparisons between male and female white perch plasma proteins detected in internal IC and 17 β -estradiol (E₂)-induced samples by tandem mass spectrometry using the predicted proteins from the striped bass genome as a reference database

<i>p</i> -value range	Treatment (IC/E ₂ - induced)	Gender (male/female)	Interaction (treatment by gender)
<0.0001	47	14	23
0.0001–0.05	31	33	36
>0.05	16	21	34
Significant by BFW	57	26	32
Significant by BH	77	44	57
Total Proteins	94	68	93

The number of proteins determined to be significantly different after the Bonferroni familywise error (BFW) and Benjamini–Hochberg false discovery rate (BH) corrections for multiple tests also are reported.

Table 3

Performance of support vector machines (SVMs) classifiers used to evaluate white perch plasma proteomes

Treatment groups and machine learning preprocessing steps	Classifier performance statistics	66%-split CV	10-fold CV
No preprocessing (IC male, IC female, E ₂ -induced male or E ₂ -induced female plasmas)	Correct classification	100.0	100.0
	Kappa statistic	1.0	1.0
	AUROC	1.0	1.0
<i>k</i> -means 3 clusters (IC male, IC female, or E ₂ -induced plasmas of both genders)	Correct classification	100.0	100.0
	Kappa statistic	1.0	1.0
	AUROC	1.0	1.0
<i>k</i> -means 2 clusters (E ₂ -induced male and E ₂ -induced female plasmas)	Correct classification	100.0	not performed
	Kappa statistic	1.0	not performed
	AUROC	1.0	not performed

Treatment groups were initial control (IC) male, IC female, estradiol-17 β (E₂)-induced male, and E₂-induced female. Data were clustered into either three or two clusters by *k*-means clustering or this preprocessing step was omitted. 'Correct classification' refers to SVMs classifier performance during 66%-split or 10-fold cross-validation (CV) and is given as a percent (%). 'AUROC' is the area under the receiver operating characteristic curve.

Table 4

Performance of support vector machines (SVMs) classifiers used to evaluate white perch plasma proteomes

Treatment groups and machine learning preprocessing steps	Classifier performance statistics	66%-split CV	10-fold CV
No preprocessing (IC male, IC female, E ₂ -induced male or E ₂ -induced female plasmas)	Correct classification	10.0 ± 24.2%	14.2 ± 14.7%
	Kappa statistic	0.022 ± 0.128	-0.144 ± 0.196
	AUROC	0.707 ± 0.190	0.389 ± 0.141
<i>k</i> -means 3 clusters (IC male, IC female, or E ₂ -induced plasmas of both genders)	Correct classification	5.0 ± 21.1%	13.3 ± 12.5%
	Kappa statistic	0.00 ± 0.105	-0.156 ± 0.167
	AUROC	0.635 ± 0.240	0.376 ± 0.121
<i>k</i> -means 2 clusters (E ₂ -induced male and E ₂ -induced female plasmas)	Correct classification	40.0 ± 20.1%	not performed
	Kappa statistic	0.000 ± 0.000	not performed
	AUROC	0.400 ± 0.210	not performed

Treatment groups were initial control (IC) male, IC female, estrogen (E₂)-induced male, and E₂-induced female. Data were clustered into either three or two clusters by *k*-means clustering or this preprocessing step was omitted. 'Correct classification' refers to SVMs classifier performance during 66%-split or 10-fold cross-validations (CV) and is given as a percent (%). 'AUROC' is the area under the receiver operating characteristic curve. Values reported represent the average ± standard deviation of SVMs models generated from ten randomly ordered datasets.