**Running head:**

Fathead minnow microarray interlaboratory comparison

**Corresponding author:**

Christopher J Martyniuk

Center for Environmental and Human Toxicology, Department of Physiological Sciences, UF Genetics Institute, College of Veterinary Medicine, University of Florida, Gainesville, Florida 32611-8525

(352) 294-4636

cmartyn@ufl.edu

**How consistent are we? Interlaboratory comparison study in fathead minnows using the model estrogen 17α-ethinylestradiol to develop recommendations for environmental transcriptomics**

April Feswick †, Meghan Isaacs †, Adam Biales ‡, Robert W. Flick ‡ , David C. Bencic ‡, Rong-Lin Wang ‡, Chris Vulpe §, Marianna Brown-Augustine §, Alex Loguinov §, Francesco Falciani ||, Philipp Antczak ||, John Herbert ||, Lorraine Brown #, Nancy D. Denslow ††, Kevin J. Kroll ††, Candice Lavelle ††, Viet Dang ††, Lynn Escalon ‡‡, Natàlia Garcia-Reyero ‡‡§§ ,Christopher J Martyniuk† †† \*,Kelly R. Munkittrick †

† Department of Biology, University of New Brunswick, Saint John, New Brunswick E2L 4L5

‡Molecular Indicators Research Branch, National Exposure Research Laboratory, Cincinnati, Ohio 45268

§Department of Nutritional Science and Toxicology, University of California-Berkeley, Berkeley, CA 94720-3104

|| University of Liverpool, Institute of Integrative Biology, Liverpool, United Kingdom L697ZB

# Pacific Environmental Science Centre, 2645 Dollarton Highway, North Vancouver, British Columbia V7H 1B1

††Center for Environmental and Human Toxicology, Department of Physiological Sciences, UF Genetics Institute, College of Veterinary Medicine, University of Florida, Gainesville, Florida 32611-8525

‡‡ US Army Engineer Research & Development Center, 3909 Halls Ferry Rd, Vicksburg MS 39180, USA

§§ Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Starkville, Mississippi, United States

\*Corresponding author: Chris Martyniuk (cmartyn@ufl.edu)

## **Abstract** Fundamental questions remain about the application of omics in environmental risk assessments, such as the consistency of data across laboratories. The objective of this study was to determine the congruence of transcript data across six independent laboratories. Male fathead minnows (FHM) were exposed to a measured concentration of 15.8 ng /L EE2 for 96 h. Livers were divided equally and sent to the participating laboratories for transcriptomic analysis using the same FHM microarray. Each laboratory was free to apply bioinformatics pipelines of their choice. There were 12,491 transcripts that were identified by one or more of the laboratories as responsive to EE2. Of these, 587 transcripts (4.7%) were detected by all laboratories. Mean overlap for differentially expressed genes among laboratories was ~50%, which improved to ~59.0% using a standardized analysis pipeline. Dynamic range of fold change estimates were variable between laboratories, but ranking transcripts by their relative fold difference resulted in a positive relationship for comparisons between any two laboratories (mean R2 > 0.9, p<0.001). Ten estrogen-responsive genes encompassing a fold change range from dramatic (>20-fold) (e.g. vitellogenin) to subtle (~2 fold) (i.e. Block of Proliferation 1) were identified as differentially expressed, suggesting that laboratories can consistently identify transcripts that are known *a priori* to be perturbed by a chemical stressor. Thus, attention should turn towards identifying core transcriptional networks using focused arrays for specific chemicals. Additionally, agreed upon bioinformatics pipelines and the ranking of genes based upon fold change (as opposed to p-value), should be considered in environmental risk assessment. These recommendations are expected to improve comparisons across laboratories and advance the use of omics in regulations.

**Keywords:** transcriptomics, risk assessment, endocrine disruptors, estrogenic compounds, interlaboratory comparison

**Introduction**

The use of biological indicators for environmental regulations, risk assessments, or monitoring programs requires careful investigation of the experimental method to generate reliable and reproducible data. Interlaboratory assays are conducted to validate the effectiveness and reliability of a method, and to demonstrate how an established protocol performs under different conditions. In the context of ecotoxicology and environmental monitoring, interlaboratory comparisons have been conducted for physiological endpoints such as testosterone (T), 17β-estradiol (E2) and 11-ketotestosterone (11KT) plasma steroids [1, 2], ethoxyresorufin-O-deethylase; ERODs [3, 4] and plasma vitellogenin [5], as well as for acute and short-term chronic whole effluent toxicity test methods [6].

The ability to generate consistent and reliable data widens applicability and improves predictability. More than 60 fish species are used in Canada’s Environmental Effects Monitoring (EEM) program [7, 8] and each species can show divergent biological responses to endpoint perturbation [9]. Due to this inherent species variability, a 25% or 2 standard deviation change from “normal” or baseline data in most whole organism endpoints is the recognized EEM threshold which triggers a change in the intensity, frequency, or focus of monitoring to further evaluate potential impacts at a particular site [10]. This magnitude of response was found to be consistent using a variety of approaches to define a ‘critical effects size” [11], although the definition of “unacceptable” change in physiological measurements can be developed via a multi-stakeholder negotiation process [12].

To date, the inherent variability in physiological endpoints, the inability to narrowly define an ecologically-relevant threshold level, and species differences in distribution and sensitivity, limits their utility in large-scale monitoring programs. The acceptable standardization of endpoints commonly used as biomarkers in endocrine disrupting compound assessments has been a struggle for nearly two decades [13]. However, despite variability, physiological endpoints (e.g. plasma Vtg) can be successfully linked to ecologically relevant endpoints (e.g fecundity) using modeling approaches from laboratory experiments [14]. Such a linkage requires consistent measurement of physiological endpoints, since the predictive model is driven by the magnitude of physiological change.

A concerted effort by researchers to incorporate omics-based tools in environmental risk assessment is underway and stakeholders continue to debate the criteria needed to facilitate the acceptance of omics approaches into a regulatory framework. For molecular endpoints, challenges related to reproducibility, variability and interpretation have been well documented in both laboratory studies [15, 16] and field studies [17]. To include molecular tools more broadly in monitoring and risk assessment, it is necessary to standardize and validate methods via interlaboratory studies, to better develop an understanding of the magnitude of change which is meaningful, and to better understand baseline variability [17].

Transcriptomic studies generally identify a list of differentially expressed genes (DEGs) related to specific biological processes that are responsive to a treatment. The successful identification of impacted genes relies heavily on the minimization of technical variation. For example, many non-biological variables have been identified that can influence interlaboratory comparisons of microarray data, such as sample processing, batch effects, platform type, microarray labelling, hybridization, filtering, data acquisition and normalization, and analysis [18-22]. Standard operating procedures and the introduction of reporting standards (Minimum Information About a Microarray Experiment) [23] has aided in the cross-comparison of microarray generated data. The comparison of a list of DEGs and the reproducibility of this list across laboratories appears to be the major goal of interlaboratory studies, and supports the notion that the list of DEGs are consistent among laboratories. Evidence that the standardization of microarray protocols can result in high levels of technical reproducibility, both within and between laboratories, has been demonstrated by the MicroArray Quality Control (MAQC) study [20].

To our knowledge, only two studies in ecotoxicology have conducted an interlaboratory study for microarray data (though it is acknowledged that other fields have conducted similar cross-laboratory comparisons). Gene expression profiles of HepG cells exposed to benzo(a)pyrene were characterized by two individual laboratories using two different platforms [24]. Although the study was limited by a small sample size, it concluded that there existed a significant correlation in the identified differentially expressed transcripts in both laboratories following BaP exposure as well as congruence in treatment-specific effects, despite data variability. Building on this, six laboratories with experience in microarray analysis conducted a study that exposed amphipods to cyfluthrin-spiked or control sediments [25]. The conclusion of this study was that a subset of genes could be consistently identified as differentially expressed across laboratories, despite variability in factors such as RNA quality.

Our objective was to evaluate the reproducibility of data collected from microarray analysis across multiple laboratories using the fathead minnow (*Pimephales promelas*, FHM), a widely utilized North American toxicological model in both laboratory ecotoxicology testing [26], and established monitoring programs developed by Canada’s EEM [8] and the USA’s EPA [27]. Our questions centered on the ability of independent laboratories to generate comparable differentially expressed gene lists as well as to identify a suite of biomarkers of exposure exhibiting different magnitudes of response, despite differences in data processing and bioinformatics analyses. To this end, FHMs were exposed to a low, measured dose (~15 ng/L) of 17α-ethinylestradiol (EE2), as this pharmaceutical is measured in North American water systems (typically detected in the low ng/L range) [28-30]. Livers were subsampled and expression profiles compared among six laboratories using platforms from the same production batch. The reasons for the use of this pharmaceutical were the following: (1) EE2 is a model compound for inducing estrogenic effects in fishes [31-34], specifically in the hepatic transcriptome in a number of small-bodied fish species [35-37]; (2) EE2 has been widely studied using transcriptomics [35-38], thus there is a strong body of literature that can be leveraged for future comparisons; (3) EE2 has been demonstrated to affect physiological parameters such as steroid levels [39, 40] and vitellogenin [31, 40], as well as fecundity, secondary sexual characteristics and GSI [40]. It is also associated with population level impacts in FHMs in wild populations [41]. Thus the nexus between transcriptomic, physiological and reproductive impacts using this pharmaceutical are most pertinent to monitoring programs.

## **Methods**

### *Experimental animals and EE2 exposure*

All experimental procedures were approved by the Animal Care Committee (protocol number 2013-3s-09) and carried out at the Canadian Rivers Institute at the University of New Brunswick, Saint John, NB, Canada. Carbon-filtered, dechlorinated city water was used for waterborne exposures. Reproductively mature, lab-reared male FHMs aged 1.5 years were housed in 400 L fiberglass holding tanks at a density of approximately 60 fish per tank under a photoperiod of 16 h light:8 h dark cycle. Temperature of the holding tanks was kept at 24-26 °C and dissolved O2 was 95-100%. Fish were fed ad libitum 1-2 times a day. We acknowledge that fish were older than those used in laboratory toxicity testing (~6 month), however the response of these animals to EE2 was dramatic and as expected, and the objective was to assess reproducibility of the transcriptome data among laboratories from the same liver.

EE2 (catalog number E4876, CAS 57-63-6, ≥98% purity) was purchased from Sigma Chemical Company (St. Louis, MO). A stock solution of 1 mg/mL EE2in ethanol was diluted to a final nominal concentration of 25 ng/L EE2 and 0.00025% ethanol. Control dose (vehicle control) without EE2 contained 0.00025% ethanol. There were ten control and eleven treatment 18 L glass aquaria, each with one male fish, all of equal size. The same temperature, photoperiod, oxygenation and feeding conditions as stated above also apply for the exposure and were measured daily. Fish were fed after the completion of each water change. Test waters were renewed at 24 h, 48 h and 72 h by total draining of the tank (100% water change), and the experiment was terminated after 96 h.

### *Water Chemistry*

Water samples were collected from three randomly chosen tanks for both control and EE2 treatments as fresh preparations of solutions (i.e. immediately after water change) on Day 2 and Day 4. Samples were stored in brown glass bottles with Teflon caps in the dark at 4 °C until chemical analysis, which occurred approximately one month later. Measurements of EE2 were performed in triplicate using an enzyme-linked immunosorbent assay (ELISA) kit (Abraxis, Los Angeles, CA) following the manufacturer's instructions. Briefly, 1 L of the water was concentrated on ODS-C18 columns (AccuBond II, Agilent Technologies, Palo Alto, CA, USA), eluted with dichloromethane, and evaporated to dryness with nitrogen and heated water bath. Samples were then reconstituted with methanol followed by water, resulting in 2.5 mL of a 10% methanol solution (400-fold concentration factor). The working range of this sandwich ELISA is 0.05 - 3 µg/L. Inter-assay and intra-assay variation (CV) using positive controls is typically 9.9% and 4.5%, respectively.

### *Experiment Termination*

At the conclusion of the exposures, fish were anesthetised using 250 mg/L buffered tricaine methanesulfonate (MS-222; Sigma Chemical Company), and fish fork length and weight were recorded. Blood samples were collected from the caudal sinus into heparinized hematocrit tubes (Fisher Scientific, Ottawa, ON, CAN) for plasma vitellogenin analysis and centrifuged at 2000 × g for 10 min with the resulting plasma being stored at −20 °C for vitellogenin quantification. Fish were then euthanized by spinal severance. Liver and testis weight (±0.01 g) were recorded in order to calculate morphometric parameters to ensure there were no differences between treatment groups in gonadosomatic index (GSI; gonad weight/body weight × 100) or liversomatic index (LSI; liver weight/body weight × 100). Each fish liver was then immediately subdivided into seven equally-sized pieces and each subdivided liver was flash frozen in individual Eppendorf® 1.5 ml centrifuge tubes on dry ice. Multiple pieces of hepatic tissue were harvested from alongside the intestine, and scooped into a single pile. Thus, tissue pieces were mixed as effectively as possible prior to distribution into the separate tubes. Tissues for gene expression analysis were stored at −80 °C until RNA extraction. Eight control fish and eight treated fish were chosen at random to be analyzed by every laboratory for the microarray analysis; one remaining set was kept as a back-up. Therefore, every laboratory received a liver subsample from each of 8 control and 8 exposed fish, totaling 16 liver sub-samples to be used with two slides containing 8x60K probes (detailed below).

### *Plasma vitellogenin levels*

Plasma vitellogenin was measured at the University of New Brunswick, Saint John laboratory using a commercial FHM vitellogenin ELISA kit (Cayman Chemical Company, Ann Arbor, MI, #V01018401). Control samples (n=8) were diluted 1:1,000 and 1:1,000,000, while treatment samples (n=8) were diluted 1:100,000 and 1:1,000,000 using the provided dilution buffer. Two treatment samples were too concentrated to obtain an absorbance reading and were not used. All samples were analyzed in duplicate and the coefficient of variation between duplicates was <20% (average CV 3%). The working range and level of detection of this sandwich ELISA is 0.05- 50 ng/ml and 0.02 ng/ml, respectively.

### *Participating laboratories*

Sixteen liver samples (8 control and 8 EEs exposed) were shipped on dry ice by courier in numbered (blinded) vials to six laboratories with experience in microarray analysis. Participants were asked to check the integrity of the parcel, the presence of dry ice, and the status of the samples immediately upon receipt. Each laboratory was given specific instructions on sample processing, detailed in each section below. Microarrays were also provided to each laboratory. The custom 60K probe fathead minnow array Agilent platform used was GPL15775 [42]. Flexibility was given in the Standard Operating Procedures to realistically model variation present in sample preparation and microarray analysis that is currently present across ecotoxicology laboratories.

### *RNA isolation and RNA integrity*

Each laboratory extracted RNA from liver tissue using a phenol and guanidine isothiocyanate based method (e.g TriZOL®, TriReagent, STAT60) according to the manufacturers’ protocols, and pellets were reconstituted based upon visual inspection of the pellet. Total RNA was quantified and purity was assessed using the Thermo Scientific™ NanoDrop by each laboratory. RNA integrity was assessed by each laboratory using the method of their choosing. RNA integrity values for all six laboratories, as well as the method used, are found in Table 1.

### *Microarray processing and analysis by each individual laboratory*

Each laboratory labeled 100 ng total RNA (laboratory 4 was an exception; 58.5 ng RNA in 1.5 µl was used due to low yields in some samples) and hybridized for 17 h at 65 ° C with two washes. The Agilent One-Color Microarray-Based Gene Expression Analysis (Low input quick Amp Labeling Kit and RNA Spike in Kit) protocol was followed as per the manufacture’s instruction. Each laboratory was asked to scan and report parameters for a 60K microarray. This microarray consists of 65,528 targets, and of these, there were 32,155 unique gene symbols, and 20,265 targets on the platform with full gene names. All microarray data have been deposited to the Gene Expression Omnibus (GEO) at the National Centre for Biotechnology Information as a Super Series (NCBI GSE 81544).One laboratory submitted their data independently to GEO (GSE70807). Note: The letters in the GEO submission do not correspond to the laboratory designations here.

After each laboratory had performed microarray hybridization using different data processing and analysis methods, sample identifications were released. There were no restrictions on algorithms or methods for microarray analysis, as a goal of the project was to assess congruency in the number, ID, fold change, and p-value of gene probes identified as differentially and not differentially expressed between control and treatment groups across laboratories. Each laboratory used a variety of methods to analyze the data. Details on the data processing and analysis conducted by individual laboratories are provided in Supplemental Table 1. Both raw and normalized data were then forwarded by each laboratory to laboratory 1 (the coordinating laboratory).

### *Microarray analysis by laboratory one*

Laboratory 1 was selected as the relative benchmark for interlaboratory comparisons, in order to determine the relative contribution of data processing to the variability in the response. Each individual dataset from a laboratory was analyzed separately and identically by laboratory 1. Spot intensity and coefficient of variation (CV) data were calculated by first normalizing the within-experiment microarrays using quantile normalization. To reduce the effect of noise across arrays and to standardize the limit of detection among laboratories, datasets were filtered to the average intensity of the 8th Agilent spike across all spots and arrays [(+)E1A\_r60\_a107 (spike 8)]. The rationale for using this Spike-in was that, upon viewing all quality control reports for all the laboratories, it consistently marked the final point in linearity of the dynamic range. The limits of detection for intensity, based on 560 array targets, were the following: laboratory 1 = 4.34, laboratory 2 = 5.25, laboratory 3 = 5.26, laboratory 4 = 4.87 (one array was removed due to poor hybridization), laboratory 5 = 6.5 (one duplicate array was removed from the final analysis), and laboratory 6 = 5.01. Prior to an ANOVA on normalized data (JMP Genomics, V6), all control spots were removed from each experimental dataset, leaving 61,057 targets. An FDR was set at 0.05.

**Results**

***17-alpha ethinylestradiol in the water***

The mean concentration (±SEM) of EE2 in the tanks was 15.67 ± 4.71 ng/L. There was no detectable EE2 in the control water samples.

### *Morphological and Physiological endpoints*

 There were no significant differences in body weight, body length, GSI or LSI between the control and EE2-treated group (Table 2). Plasma vitellogenin levels of EE2-treated males were 4.0 ± 12.7 µg/ml for control and 21,666 ± 1821 µg/ml for EE2 treated fish after 96h, and FHM responded as expected to the exposure.

### *Expression results across laboratories*

All laboratories submitted data from RNA extraction/purification methods, RNA quality assessment, and stated the method that was used to normalize the microarray intensity data (Table 1; Supplemental Table 1). Laboratory 5 had to omit two control samples due to low RNA quality and instead, two replicate samples were hybridized again as technical replicates to confirm that there was >98% reproducibility in signal intensity per target on the microarray between technical replicates (data not shown). Only one technical replicate was chosen at random to represent the single biological replicate. All labelled samples used by each laboratory met the minimum criteria for hybridization per Agilent guidelines (yields > 0.825 µg and specific activity > 6.0 pmol Cy3 per µg cRNA). Each laboratory submitted data containing the number, ID, fold change relative to controls, and p-value of all gene probes to laboratory 1.

### *Characterization of probe variability across laboratories*

In order to standardize the analysis approach and to determine how much variability was due to the hybridization and scanning protocols, technical variability for both mean target probe CV (a measure of inter-assay variability) and intensity was determined for both control and treatment samples for each laboratory based upon the common analysis workflow performed by laboratory 1. Each laboratory showed a comparable level of variability (Figure 1a). Overall, in each group, the liver transcriptome showed a CV (expressed as percent) that ranged from 6-8%. Likewise, the mean log2 spot intensity of all targets on the microarray ranged from 6.0-8.0 (Figure 1b).

 Each laboratory, using the statistical methods of their choice, submitted a list of differentially expressed probes (FDR of 0.05). A total of 641 differentially expressed probes (duplicates removed) were identified as being in common by all six laboratories following their own analyses. This number was largely influenced by laboratory 3, whose post-hoc corrected p-values resulted in a significant reduction in the number of probes identified (only 165 were reported); therefore, non-FDR corrected data (2375 probes) were used for laboratory 3 in subsequent comparisons. The decision to include data from laboratory 3 in the study, despite its influence on data variability and congruency across laboratories, was due to the goal of this study; namely, to assess microarray variability across established ecotoxicology laboratories regularly employing this technology. Despite the variability in methods used, 50.6% of non-duplicate probes (using an FDR of 0.05) were shared on average between each laboratory and laboratory 1, ranging from 30 - 77% overlap (Table 3). When laboratory 3 is excluded, the number of non-duplicate probes shared between each laboratory and laboratory 1 increases by 5% to 55.6%. Unsurprisingly, the number of probes any given laboratory shared in common with laboratory 1 is positively correlated with the number of probes identified. That is to say, the more probes identified, the more likely that some of those probes will be identified by another laboratory.

### *Congruence in the identification of identical differentially expressed genes*

Across the six laboratories, 12,491 differentially expressed genes (duplicates removed) were identified as responding to the EE2 treatment, out of a total of 22,010 genes on the array. There was a wide range in how many DEGs each laboratory detected (2196 - 6361; Table 4A), based on their own analysis using post-hoc corrected cut-off of ±1.5-fold (except laboratory 3 where uncorrected p-values were used). For example, laboratory 2 detected 6152 unique genes (Table 4A), of which 80% were shared with laboratory 1 (Table 4B). However, there was a wide range in the percentage of DEGs any given laboratory shared with another (Table 4B). Laboratories 1, 2 and 6 were most similar in their DEG profiles, sharing an average of at least 60% of the same DEGs with the other five laboratories. Between laboratories, laboratory 1 shared the highest percentage of DEGs with laboratory 2 (4040 genes in common; 79.8% of laboratory 2’s DEGs). Supplemental table 2 also presents the same analysis but when the data are analyzed by standard approach. The influence of using a fold change cut-off of 3.0 on the number of genes both unique and shared across laboratories is presented as Supplemental Table 3.

Each laboratory identified a subset of DEGs that no other laboratory detected (“unique DEGs”). This number varied widely, for example, laboratory 4 identified the lowest number of unique DEGs at 48, whereas laboratory 3 identified the highest number of unique DEGs at 2741 (Table 5). Across all laboratories, the number of unique DEGs detected totaled 5364, which represents almost 43% (5364/12,491) of all DEGs responding to the estrogenic treatment. When examined as percentages, only 2.2% of the DEGs reported by Laboratory 4 were unique genes that no other laboratory identified. Thus, one or more of the other laboratories identified the significant majority of transcripts in their list. Similarly, laboratories 1, 2, 3, 5 and 6 reported the following percentages of unique genes; 6.7%, 10.6%, 43.1%, 15.1% and 16.1%. All six laboratories identified the same 587 DEGs (which was 4.7% of the total 12,491 DEGs detected in one or more laboratory), and 43% (5364) of the expressed genes were uniquely detected by only one laboratory. The proportion of DEGs that were shared by all six laboratories ranged from 9.2% for laboratory 3, to 26.7% for laboratory 4. Laboratory 3 contributed 21.9% unique DEGs to the 12,491 DEGs reported and thus contributed the most to “DEG noise”. It is unclear whether DEG “noise” is reflected by the number of unique genes detected, or the increased detection represents increased sensitivity. Resolving the issue of sensitivity versus noise may require a comparative pathway analysis, or validation with other methods such as qPCR, to more clearly evaluate whether increased detection improves the analysis or confuses the issue, but in either case, it is clear that individual gene responses must be interpreted with caution.

***Congruence in the magnitude of response of DEGs across laboratories***

The mean fold change of DEGs for both increasing (Figure 2A) and decreasing transcripts (Figure 2B) relative to the control group demonstrate significant differences across laboratories in overall results. However, when data are reported as median fold changes (Figure 2C and 2D) distributions of average DEG fold changes across laboratories were not more similar. Using unadjusted p-values for laboratory 3 and post-hoc corrected, cut-off of ±1.5-fold for all other laboratories, the fold change distribution of DEGs for both increasing (Figure 2E) and decreasing (Figure 2F) transcripts demonstrates that the majority of genes fall within the 1.5-3.0 fold change range.

***Congruence in the identification of a suite of known E2-responsive transcripts***

In order to assess the performance of using microarrays to identify known biomarkers of estrogenic exposure in aquatic species, ten estrogen responsive genes or “E2-responsive standards” were selected. These E2-responsive genes were identified using published literature on the expression profiles in the liver of zebrafish following estrogenic treatments [36, 43] as well as input from the Comparative Toxicogenomics Database [44] following searches for transcripts responsive to estrogens. We targeted ten genes from this list and searched for these within the datasets collected from the six laboratories, with the intention of investigating transcripts that showed a broad dynamic range for both fold change and direction of change. These genes included (1) those may show relatively low response to estrogens, for example APOA1 (apolipoprotein A-I-1), APOE (apolipoprotein Eb), c-FOS, and BOP1 (Block Of Proliferation 1), (2) a relatively intermediate response to estrogens such as IGF1 (insulin-like growth factor 1), XBP1 (x-box binding protein 1), and ESR1 (estrogen receptor 1) and (3) a relatively high response to estrogens such as RTN1A (reticulon 1a), VTG1 and VTG3. There was strong congruence across laboratories for estimating the fold changes of these known E2-responsive transcripts, regardless of whether a common analysis for DEGs was used (Figure 3a) or whether each laboratory used an independent analysis of their choice (Figure 3b). Four laboratories were successful at identifying all ten E2-responsive genes as differentially expressed. Laboratory 3 missed three of these genes, and laboratory 4 missed one, and this corresponded to the laboratories with the least number of DEGs identified after a post-hoc correction of the data (i.e. may reflect the low number of DEGs identified initially).

**Discussion**

This study aimed to assess the reproducibility of microarray data across six independent laboratories, following a 96 h EE2 exposure to fathead minnow. The value of microarrays in environmental monitoring and risk assessment depends on the ability of independent laboratories to reliably distinguish subtle changes in gene expression data (differentially expressed genes relative to control) despite both biological and technical variability. Notwithstanding those probes exhibiting expected high biological variability, the mean probe variability across the transcriptome reported here is consistent with that reported by others (5-10% CV) [45, 46] for a single tissue in fish. The overall mean signal intensity on the arrays after normalization (i.e. range of intensity across all 60K probes) was 6-8 depending upon the laboratory. This is encouraging, given some flexibility in microarray sample preparation (i.e. different reagents for RNA extraction) afforded to the participating laboratories.

Data analysis pipelines contributed to data variability, and there was an increase in DEG overlap across laboratories by ~9% when a single pipeline was used (an increase to ~59% from ~50). Variability in the analysis also arose, in part, from differences in background intensity of the microarrays. Other interlaboratory microarray studies have substantiated the numerous sources of data variation, and have highlighted the role of using standardized protocols in producing highly correlated and reproducible data [18, 20-22]. It is important to recognize that a high level of concordance of differentially expressed genes (DEGs) across platforms and laboratories depends mainly on the statistical criteria used to rank and select them; using a fold change or ratio (versus a p value) generated the most robust results [20]. Using the ranked log ratios of fold changes, we confirmed that all laboratories were detecting similar overall transcriptome responses that were statistically and highly correlated. In the case of laboratory 3, post-hoc corrected p-values resulted in a significant reduction in the number of DEGs (and the number of differentially expressed probes) identified, and influenced the results. Thus, laboratory 3 appeared to differ most in the bioinformatics approach used to process the data. However, all laboratories used different strategies (Supplemental Table 1) for data processing, editing, averaging of probes, and statistical identification of DEGs which influences the p-value threshold for a decision on whether or not a transcript is differentially expressed.

Analysis of probe variability in each laboratory revealed that there were differences in the overall magnitude of response when mean fold-change is plotted. These values can be heavily influenced by transcripts that have large expression changes relative to the control, as in the case of VTG. Thus, the sensitivity of the arrays can vary from laboratory to laboratory in the context of fold change. Plotting the median values instead of the mean revealed that the laboratories do not differ substantially, and the median fold change for the large majority of transcripts is ~2-fold.

The absolute number of DEGs detected varied considerably across laboratories. Of the 22,010 genes on the array, 12,491 DEGs responded to the EE2 treatment. Of these, 587 (4.7%) common DEGs were detected by all six laboratories and 43% (5364) were uniquely detected by only 1 laboratory. It is difficult to examine the differences and similarity in detail, but we did examine the ability of each laboratory to detect a suite of core transcripts known to be impacted by an EE2 exposure (in lieu of number of DEGs identified) by assessing the ability of independent laboratories to identify common “gold standard” estrogen-responsive genes. It is noteworthy that, despite flexibility in the SOPs for much of the experimental design, four laboratories successfully identified significant changes in all ten gold standards along a continuum of relative fold changes (decreasing to increasing). This was true irrespective of whether each laboratory independently analyzed the data or whether a single laboratory applied a common analysis. The laboratories could very precisely converge on fold change estimates for some transcripts, for example APOE and XBP1, whereas other transcripts showed wider variability in fold change estimates across laboratories (e.g. c-FOS and VTG). Although all six laboratories correctly identified VTG1 and VTG3 as up-regulated following EE2 treatment, the magnitude of induction differed significantly between laboratories. Moving forward, a recommendation is to incorporate a ranking system of transcript response, as laboratories will report fold changes differently. Lastly, BOP1 was identified as an estrogen responsive gene that was lowly expressed in every dataset (>2 fold) when analyzed by laboratory 1. This suggests that the practice of removing genes based on a fold change cut-off may result in removing genes that are responsive to the treatment, although this is a trade-off as transcripts that show lower fold changes are perhaps more likely to be false positives. Arbitrary fold change cut-offs for expression data may not be the best approach moving forward and will result in discrepancies of DEGs reported among laboratories [47].

The ability to also identify core gene targets with fold changes close to background levels suggests that microarrays are capable of measuring a suite of responsive genes with a high degree of precision when sensitive transcriptional networks are identified and assessed *a priori*. In this example, all six laboratories would have almost perfect consensus in their ranking of these ten E2-responsive transcripts. For new chemicals, observational transcriptome data would need to be collected first in order to identify responsive transcripts. Databases such as the CTD: The Comparative Toxicogenomics Database could then be used to hone the list of responsive transcripts.

Expression profiling generates important information on the chemical mode of action [48, 49]. Under Canada’s EEM program, confirmed positive responses can enter into an “Investigation of Cause” phase, and pathway information can provide valuable insight into potential modes of action to help focus subsequent studies. However, to use transcriptomics in a meaningful and legislatively robust way, it must be able to provide thresholds for specific parameters that can be tested across a range of environments, conditions, and species. At this point, it may be important for risk assessment and environmental monitoring to identify a common subset of transcripts responsive to a chemical or mixture, as opposed to leveraging the entire molecular dataset. Due to variability in responses, a recommendation would be to first agree on a suite of biomarker transcripts for a particular chemical, followed by a ranking system based on fold change [50].

To incorporate microarray technology into monitoring programs and risk assessment, there needs to be a definition on what constitutes a significant change in expression (1.5-fold versus 2). Since different normalization techniques can significantly alter gene lists created, a decision on a common normalization technique is warranted as this may influence the estimate of fold change. Until then, we recommend using fold change (versus a p-value) as the ranking criterion for gene selection, with a non-stringent *P* value cut-off based on analyses using human data sets subjected to cross-laboratory and inter-platform comparisons (Shi et al., 2006). Reproducibility among laboratories is also reported to be highest when analysis was based on biological themes defined by enriched Gene Ontology (GO) categories [22, 51], an area of research that warrants further examination in an ecotoxicology context.

**Conclusions**

The standardization of transcriptomics and the generation of comparable data via an interlaboratory study are a necessary step prior to their inclusion in regulatory or monitoring programs. Agreement on what constitutes a significant change in expression levels must therefore be defined. Proper study design will require the identification of a pre-determined “magnitude of change” (or “critical effect size”) at the molecular level (i.e. transcripts) which could be the level of change required to have an effect at the individual and/or population level. Using a common suite of genes known to be estrogen-responsive to various degrees, we demonstrated that there can be high success rate across laboratories. Further evidence of reproducible and expanded DEG lists (or subsets) across laboratories at both low and high expression levels is required. We strongly recommend to researchers in ecotoxicology that they present data as fold-change ranking and to place less emphasis on p-values to generate DEGs, with the caveat that a single individual is not disproportionally driving a biological response. Additionally, the generation of biological pathways and/or GO terms would be beneficial as both anchoring points for molecular data and the reduction of stochastic noise among individual genes, though this was not done in the current study. The role for omics is still at this point best suited for finding mechanisms and screening and not for regulatory or monitoring programs due to the difficulty in their use for pass/fail decision-making. Currently, no molecular endpoints are used in Canada’s EEM, and no established methodology exists for using these endpoints in risk assessment. We suggest that we move towards a ranking system for triggering action or regulation based upon transcriptomic responses.

**Acknowledgements**

We would like to thank Dr. Graham Van Aggelen for his assistance with the manuscript. NSERC Discovery Grant 386275-2010 to CJM and 238621-2011 to KRM, and a Canada Research Chair to CJM and KRM. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Figure Captions

## **Figure 1:** Mean coefficient of variation (±SD) of all spots over all microarrays (n=8 biological replicates) within an experimental group. (A) Coefficient of variation (CV) is plotted separately in the control group (denoted Cntl) and 17α-ethinylestradiol (EE2) for each laboratory. (B) Mean probe intensity (±SD) in each experimental group.

## **Figure 2:** Metrics of variability and spread of fold changes in differentially expressed genes (DEGs) across laboratories. Mean fold change of DEGs with 95% confidence internals (vertical bars) shown for increasing (A) and decreasing (B) transcripts, relative to the control group. Median fold change of DEGs with interquartile range (vertical bars) shown for increasing (C) and decreasing (D) transcripts respectively, relative to the control group. Fold change distribution of differentially expressed genes (DEGs) for increasing (E) and decreasing (F) transcripts respectively, relative to the control group. Uncorrected p-value for transcripts was used for laboratory 3 instead of corrected p-values as the number of DEGs identified following a post-hoc correction was small relative to the other laboratories.

## **Figure 3:** Ten estrogen-responsive genes arranged from decreasing to increasing order by fold change. Each point represents a laboratory. Only those genes that were identified as differentially expressed are included in these graphs. The horizontal line represents the median expression. Zero represents no fold change relative to control. Abbreviations are presented in the text.

|  |  |  |
| --- | --- | --- |
|  | Method used to assess RNA quality | Integrity value |
|  |  | Mean ± SD | Range |
| Lab 1 | Agilent 2100 Bioanalyzer | 9.5 ±0.1 | Not given |
| Lab 2 | Agilent 2100 Bioanalyzer | 9.7 ± 0.3 | 9.1 - 10.0 |
| Lab 3 | Agilent 2100 Bioanalyzer | 8.7 ± 0.7 | 7.7 - 9.7 |
| Lab 4 | 2:1 ratio of 28S:16S bands on a 1.5% agarose gel  | 346.2 ng | 16.8 - 363.1 ng |
| Lab 5 | Bio-Rad Experion™ Electrophoresis Station | 8.5 ± 1.1 | Not given |
| Lab 6 | Agilent 2100 Bioanalyzer | 9.2 ± 0.3 | 8.8 - 9.9 |

## **Table 1:** RNA integrity values and method used by each laboratory. The Agilent 2100 gives a RIN value (RNA Integrity Number), whereas the Bio-Rad Experion Electrophoresis System gives an RQI value (RNA Quality Index). Other metrics used to assess sample quality as well as details on the methods for analysis are provided in Supplemental Table 1.

## **Table 2:** Morphometrics (±SEM) in FHM exposed to control (*n* = 8) and 25 ng/L EE2 (*n* = 8) Measurements include length, weight, gonadosomatic index (GSI) and liver somatic index (LSI). There were no significant differences among treatments in any endpoints (*α* = 0.05).

|  |  |  |
| --- | --- | --- |
|  | Control fish | EE2 exposed fish |
| Length (mm) | 78.63 ± 2.02 | 81.63 ± 1.80 |
| Weight (g) | 6.07 ± 0.35 | 6.16 ± 0.36 |
| GSI | 1.42 ± 0.20 | 0.98 ± 0.16 |
| LSI | 1.41 ± 0.18 | 1.62 ± 0.16 |

## **Table 3:** The number of differentially expressed probes (based on spot ID) identified by each laboratory using their own method of normalization (A) and identified using the same analysis pipeline performed by the benchmark laboratory, laboratory 1 (B). Number of probes in common with laboratory 1 (C), given also as a percentage (D). The targets on the FHM microarray represent both duplicate and non-duplicated genes (thus, “probe ID” is the value here). All data are based on FDR corrected p-values.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Lab 1 | Lab 2 | Lab 3 | Lab 4 | Lab 5 | Lab 6 |
| A | 9534 | 11846 | 9889 | 3488 | 6868 | 9743 |
| B | 9534 | 12244 | 6598 | 2498 | 5927 | 7044 |
| C |  | 6048 | 2375\* | 2886 | 3934 | 4632 |
| D  |  | 77% | 30%\* | 37% | 50% | 59% |

\*Non-FDR data used

## **Table 4:** The number of DEGs (duplicate probe IDs removed) shared between each laboratory using their own method of normalization and analysis methods (A) for all genes that were identified as different (post-hoc corrected), (B) demonstrating the percentage of genes each laboratory shared with another. For example, laboratory 1 and 2 found 4040 genes in common; this number is first divided by the total number of DEGs identified by laboratory 1 (5064) to give a percentage in table B of 79.8, alternatively 4040 is also divided by the total number of DEGs identified by laboratory 2 (6152) to give a percentage in table B of 65.7. The columns and the rows represent two different ways of comparing DEG lists between laboratories. Supplemental table 2 shows the exact same analysis but when the data are analyzed with the same analysis pipeline.

|  |  |  |
| --- | --- | --- |
| A |  | Shared DEGs |
| Laboratory | 1 | 2 | 3\* | 4 | 5 | 6 |
| Total number of DEGs (given in brackets) | 1 (5064) | X |  |  |  |  |  |
| 2 (6152) | 4040 | X |  |  |  |  |
| 3\* (6361) | 2027 | 2426 | X |  |  |  |
| 4 (2196) | 1863 | 1999 | 927 | X |  |  |
| 5 (4324) | 2551 | 2822 | 1918 | 1305 | X |  |
| 6 (5744) | 3443 | 3830 | 2172 | 1852 | 2378 | X |

\*Non-FDR data used

|  |  |  |
| --- | --- | --- |
|  B  |  | Percentage of shared DEGs |
| Laboratory | 1 | 2 | 3\* | 4 | 5 | 6 | avg |
| Total number of DEGs (given in brackets) | 1 (5064) | X | 65.7 | 31.9 | 84.8 | 59.0 | 59.9 | 60.3 |
| 2 (6152) | 79.8 | X | 38.1 | 91.0 | 65.3 | 66.7 | 68.2 |
| 3\* (6361) | 40.0 | 39.4 | X | 42.2 | 44.4 | 37.8 | 40.8 |
| 4 (2196) | 36.8 | 32.5 | 14.6 | X | 30.2 | 32.2 | 29.3 |
| 5 (4324) | 50.4 | 45.9 | 30.2 | 59.4 | X | 41.4 | 45.4 |
| 6 (5744) | 68.0 | 62.3 | 34.1 | 84.3 | 55.0 | X | 60.7 |

\*Non-FDR data used

**Table 5:**  The number of unique DEGs, out of the 12,491 total DEGs detected, that any given laboratory shared with one or more other laboratories. Columns represent individual laboratories, whereas rows denote the number of laboratories involved. For example, in row one, the values given represent how many laboratories reported the same unique DEGs, therefore row one denotes unique DEGs found by each laboratory. In row two, the values given represent how many unique DEGs were shared by any specific laboratory (column) and at least one other laboratory, for a total of two laboratories sharing DEGs. Row six reflects that all six laboratories shared 587 unique DEGs.

|  |  |  |  |
| --- | --- | --- | --- |
| A |  | Laboratory ID |  |
|  | 1 | 2 | 3\* | 4 | 5 | 6 | Total DEGs |
| Number of laboratories sharing DEGs  | one | 341 | 653 | 2741 | 48 | 653 | 920 | 5364 (42.9%) |
| two | 676 | 1101 | 1168 | 97 | 682 | 954 | 2339 (18.7%) |
| three | 1082 | 1345 | 688 | 174 | 652 | 1048 | 1664 (13.3%) |
| four  | 1274 | 1334 | 591 | 522 | 784 | 1143 | 1383 (11.1%) |
| five | 1104 | 1132 | 586 | 768 | 966 | 1092 | 1127 (9.0%) |
| six  | 587 | 587 | 587 | 587 | 587 | 587 | 587 (4.7%) |
|  | Total | 5064 | 6152 | 6361 | 2196 | 4324 | 5744 |  |

\*Non-FDR data used

## **References**

[1] McMaster ME, Jardine JJ, Ankley GT, Benson WH, Greeley MS, Gross TS, Guillette LJ, MacLatchy DL, Orlando EF, Van Der Kraak GJ. 2001. An interlaboratory study on the use of steroid hormones in examining endocrine disruption. *Environmental toxicology and chemistry* 20:2081-2087.

[2] Feswick A, Ankley GT, Denslow N, Ellestad LE, Fuzzen M, Jensen KM, Kroll K, Lister A, MacLatchy DL, McMaster ME, Orlando EF, Servos MR, Tetreault GR, Van Den Heuvel MR, Munkittrick KR. 2014. An inter-laboratory study on the variability in measured concentrations of 17beta-estradiol, testosterone, and 11-ketotestosterone in white sucker: implications and recommendations. *Environ Toxicol Chem* 33:847-857.

[3] Munkittrick K, van den Heuvel MR, Metner D, Lockhart W, Stegeman J. 1993. Interlaboratory comparison and optimization of hepatic microsomal ethoxyresorufin O‐deethylase activity in white sucker (Catostomus commersoni) exposed to bleached kraft pulp mill effluent. *Environmental toxicology and chemistry* 12:1273-1282.

[4] Van Den Heuvel MR, Dixon DG, Munkittrick KR, Stegeman JJ. 1995. Second‐round interlaboratory comparison of hepatic ethoxyresorufin‐O‐deethylase activity in white sucker (Catostomus commersoni) exposed to bleached‐kraft pulp mill effluent. *Environmental toxicology and chemistry* 14:1513-1520.

[5] EPA US. 2003. Comparative evaluation of vitellogenin methods. *(68-W-01-023)*.

[6] EPA U. 2001. Final Report: Interlaboratory Variability Study of EPA Short-term Chronic and Acute Whole Effluent Toxicity Test Methods. *(821-B-01-004)*.

[7] Canada E. 2010. Pulp and paper Environmental Effects Monitoring (EEM) technical guidance document.

[8] Barrett TJ, Munkittrick KR. 2010. Seasonal reproductive patterns and recommended sampling times for sentinel fish species used in environmental effects monitoring programs in Canada. *Environmental Reviews* 18:115-135.

[9] Munkittrick K, Kraak G, McMaster M, Portt C. 1992. Reproductive dysfunction and MFO activity in three species of fish exposed to bleached kraft mill effluent at Jackfish Bay, Lake Superior. *Water quality research journal of Canada* 27:439-446.

[10] Munkittrick KR, McGeachy SA, McMaster ME, Courtenay SC. 2002. Overview of freshwater fish studies from the pulp and paper environmental effects monitoring program. *Water Quality Research Journal of Canada* 37:49-77.

[11] Munkittrick KR, Arens CJ, Lowell RB, Kaminski GP. 2009. A review of potential methods of determining critical effect size for designing environmental monitoring programs. *Environ Toxicol Chem* 28:1361-1371.

[12] Larsson Å, Förlin L, Grahn O, Landner L, Lindesjöö E, Sandström O. 2000. Guidelines for interpretation and biological evaluation of biochemical, physiological and pathological alterations in fish exposed to industrial effluents, Stockholm.

[13] Hutchinson TH, Ankley GT, Segner H, Tyler CR. 2006. Screening and testing for endocrine disruption in fish-biomarkers as" signposts," not" traffic lights," in risk assessment. *Environmental Health Perspectives* 114:106.

[14] Miller DH, Jensen KM, Villeneuve DL, Kahl MD, Makynen EA, Durhan EJ, Ankley GT. 2007. Linkage of biochemical responses to population‐level effects: A case study with vitellogenin in the fathead minnow (Pimephales promelas). *Environmental Toxicology and Chemistry* 26:521-527.

[15] Van Aggelen G, Ankley GT, Baldwin WS, Bearden DW, Benson WH, Chipman JK, Collette TW, Craft JA, Denslow ND, Embry MR. 2010. Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: hurdles, achievements, and future outlook. *Environmental health perspectives*:1-5.

[16] Hook S. 2010. Promise and progress in environmental genomics: a status report on the applications of gene expression‐based microarray studies in ecologically relevant fish species. *Journal of fish biology* 77:1999-2022.

[17] Bahamonde PA, Feswick A, Isaacs MA, Munkittrick KR, Martyniuk CJ. 2016. Defining the role of omics in assessing ecosystem health: Perspectives from the Canadian environmental monitoring program. *Environ Toxicol Chem* 35:20-35.

[18] Yauk CL, Berndt ML. 2007. Review of the literature examining the correlation among DNA microarray technologies. *Environ Mol Mutagen* 48:380-394.

[19] Muller C, Schillert A, Rothemeier C, Tregouet DA, Proust C, Binder H, Pfeiffer N, Beutel M, Lackner KJ, Schnabel RB, Tiret L, Wild PS, Blankenberg S, Zeller T, Ziegler A. 2016. Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. *PLoS One* 11:e0156594.

[20] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, De Longueville F, Kawasaki ES, Lee KY. 2006. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology* 24:1151-1161.

[21] Scott D, Devonshire A, Adeleye Y, Schutte M, Rodrigues M, Wilkes T, Sacco M, Gribaldo L, Fabbri M, Coecke S. 2011. Inter-and intra-laboratory study to determine the reproducibility of toxicogenomics datasets. *Toxicology* 290:50-58.

[22] Weis B. 2005. Standardizing global gene expression analysis between laboratories and across platforms. *Nature methods* 2:351-356.

[23] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29:365-371.

[24] Hockley SL, Mathijs K, Staal YC, Brewer D, Giddings I, van Delft JH, Phillips DH. 2009. Interlaboratory and interplatform comparison of microarray gene expression analysis of HepG2 cells exposed to benzo (a) pyrene. *OMICS A Journal of Integrative Biology* 13:115-125.

[25] Vidal-Dorsch DE, Bay SM, Moore S, Layton B, Mehinto AC, Vulpe CD, Brown-Augustine M, Loguinov A, Poynton H, Garcia-Reyero N, Perkins EJ, Escalon L, Denslow ND, Cristina CR, Doan T, Shukradas S, Bruno J, Brown L, Van Agglen G, Jackman P, Bauer M. 2015. Ecotoxicogenomics: Microarray interlaboratory comparability. *Chemosphere* 144:193-200.

[26] Ankley GT, Villeneuve DL. 2006. The fathead minnow in aquatic toxicology: past, present and future. *Aquatic Toxicology* 78:91-102.

[27] EPA. 2002. Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms. Fifth Edition., p. 275.

[28] Barel-Cohen K, Shore LS, Shemesh M, Wenzel A, Mueller J, Kronfeld-Schor N. 2006. Monitoring of natural and synthetic hormones in a polluted river. *Journal of Environmental Management* 78:16-23.

[29] Fernandez MP, Ikonomou MG, Buchanan I. 2007. An assessment of estrogenic organic contaminants in Canadian wastewaters. *Science of the Total Environment* 373:250-269.

[30] Atkinson SK, Marlatt VL, Kimpe LE, Lean DR, Trudeau VL, Blais JM. 2012. The occurrence of steroidal estrogens in south-eastern Ontario wastewater treatment plants. *Science of the Total Environment* 430:119-125.

[31] Pawlowski S, Van Aerle R, Tyler C, Braunbeck T. 2004. Effects of 17α-ethinylestradiol in a fathead minnow (Pimephales promelas) gonadal recrudescence assay. *Ecotoxicology and environmental safety* 57:330-345.

[32] Parrott JL, Blunt BR. 2005. Life‐cycle exposure of fathead minnows (Pimephales promelas) to an ethinylestradiol concentration below 1 ng/L reduces egg fertilization success and demasculinizes males. *Environmental toxicology* 20:131-141.

[33] Martyniuk CJ, Kroll KJ, Doperalski NJ, Barber DS, Denslow ND. 2010. Environmentally relevant exposure to 17alpha-ethinylestradiol affects the telencephalic proteome of male fathead minnows. *Aquat Toxicol* 98:344-353.

[34] Salierno JD, Kane AS. 2009. 17α‐Ethinylestradiol alters reproductive behaviors, circulating hormones, and sexual morphology in male fathead minnows (Pimephales promelas). *Environmental Toxicology and Chemistry* 28:953-961.

[35] Hoffmann J, Torontali S, Thomason R, Lee D, Brill J, Price B, Carr G, Versteeg D. 2006. Hepatic gene expression profiling using Genechips in zebrafish exposed to 17α-ethynylestradiol. *Aquatic Toxicology* 79:233-246.

[36] Martyniuk CJ, Gerrie ER, Popesku JT, Ekker M, Trudeau VL. 2007. Microarray analysis in the zebrafish (Danio rerio) liver and telencephalon after exposure to low concentration of 17alpha-ethinylestradiol. *Aquat Toxicol* 84:38-49.

[37] Doyle MA, Bosker T, Martyniuk CJ, Maclatchy DL, Munkittrick KR. 2013. The effects of 17-alpha-ethinylestradiol (EE2) on molecular signaling cascades in mummichog (Fundulus heteroclitus). *Aquat Toxicol* 134-135:34-46.

[38] Skillman AD, Nagler JJ, Hook SE, Small JA, Schultz IR. 2006. Dynamics of 17alpha-ethynylestradiol exposure in rainbow trout (Oncorhynchus mykiss): absorption, tissue distribution, and hepatic gene expression pattern. *Environ Toxicol Chem* 25:2997-3005.

[39] MacLatchy DL, Courtenay SC, Rice CD, Van der Kraak GJ. 2003. Development of a short-term reproductive endocrine bioassay using steroid hormone and vitellogenin end points in the estuarine mummichog (Fundulus heteroclitus). *Environ Toxicol Chem* 22:996-1008.

[40] Runnalls TJ, Beresford N, Kugathas S, Margiotta-Casaluci L, Scholze M, Scott AP, Sumpter JP. 2015. From single chemicals to mixtures-Reproductive effects of levonorgestrel and ethinylestradiol on the fathead minnow. *Aquat Toxicol* 169:152-167.

[41] Kidd KA, Blanchfield PJ, Mills KH, Palace VP, Evans RE, Lazorchak JM, Flick RW. 2007. Collapse of a fish population after exposure to a synthetic estrogen. *Proceedings of the National Academy of Sciences* 104:8897-8901.

[42] Garcia-Reyero N, Kennedy AJ, Escalon BL, Habib T, Laird JG, Rawat A, Wiseman S, Hecker M, Denslow N, Steevens JA, Perkins EJ. 2014. Differential effects and potential adverse outcomes of ionic silver and silver nanoparticles in vivo and in vitro. *Environ Sci Technol* 48:4546-4555.

[43] Levi L, Pekarski I, Gutman E, Fortina P, Hyslop T, Biran J, Levavi-Sivan B, Lubzens E. 2009. Revealing genes associated with vitellogenesis in the liver of the zebrafish (Danio rerio) by transcriptome profiling. *BMC Genomics* 10:141.

[44] Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, Mattingly CJ. 2013. The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res* 41:D1104-1114.

[45] Dreier DA, Loughery JR, Denslow ND, Martyniuk CJ. 2016. The influence of breeding strategy, reproductive stage, and tissue type on transcript variability in fish. *Comp Biochem Physiol Part D Genomics Proteomics*.

[46] Wang RL, Bencic DC, Garcia-Reyero N, Perkins EJ, Villeneuve DL, Ankley GT, Biales AD. 2014. Natural Variation in Fish Transcriptomes: Comparative Analysis of the Fathead Minnow (Pimephales promelas) and Zebrafish (Danio rerio). *PLoS One* 9:e114178.

[47] Dalman MR, Deeter A, Nimishakavi G, Duan ZH. 2012. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics* 13 Suppl 2:S11.

[48] Thomas RS, Rank DR, Penn SG, Zastrow GM, Hayes KR, Pande K, Glover E, Silander T, Craven MW, Reddy JK, Jovanovich SB, Bradfield CA. 2001. Identification of toxicologically predictive gene sets using cDNA microarrays. *Mol Pharmacol* 60:1189-1194.

[49] Mancia A, Abelli L, Kucklick JR, Rowles TK, Wells RS, Balmer BC, Hohn AA, Baatz JE, Ryan JC. 2015. Microarray applications to understand the impact of exposure to environmental contaminants in wild dolphins (Tursiops truncatus). *Mar Genomics* 19:47-57.

[50] Dembele D, Kastner P. 2014. Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinformatics* 15:14.

[51] Zhang L, Zhang J, Yang G, Wu D, Jiang L, Wen Z, Li M. 2013. Investigating the concordance of Gene Ontology terms reveals the intra- and inter-platform reproducibility of enrichment analysis. *BMC Bioinformatics* 14:143.