


Spring 3-13-2017

# Transcriptome Approach for Identifying Potential Biomarkers for Endocrine Disruption due to Crude Oil Exposure using Killifish (*Fundulus Heteroclitus*).

Frank J. Zadlock IV  
frank.zadlock@student.shu.edu

Follow this and additional works at: <https://scholarship.shu.edu/dissertations>

 Part of the [Aquaculture and Fisheries Commons](#), [Bioinformatics Commons](#), and the [Computational Biology Commons](#)

---

## Recommended Citation

Zadlock, Frank J. IV, "Transcriptome Approach for Identifying Potential Biomarkers for Endocrine Disruption due to Crude Oil Exposure using Killifish (*Fundulus Heteroclitus*).\" (2017). *Seton Hall University Dissertations and Theses (ETDs)*. 2263.  
<https://scholarship.shu.edu/dissertations/2263>

**Transcriptome approach for identifying potential biomarkers  
for endocrine disruption due to crude oil exposure using  
killifish (*Fundulus heteroclitus*).**

**By**


**Frank J Zadlock IV**


Submitted in partial fulfillment of the requirements for the degree of Doctor of  
Philosophy in Molecular Bioscience from the Department of Biological Sciences of  
Seton Hall University

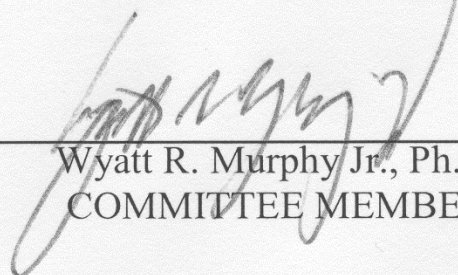
May, 2017

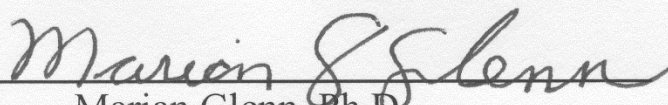
© 2017 Frank J Zadlock IV

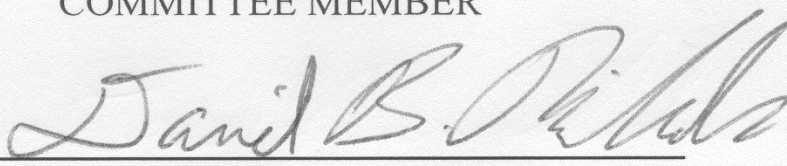
APPROVED BY

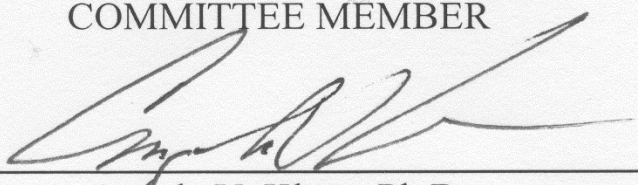
  
Carolyn S. Bentivegna, Ph.D.  
MENTOR

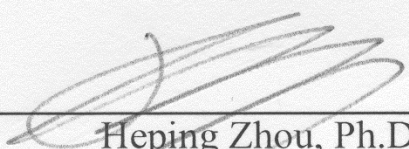
  
Jane, L. Ko, Ph.D.  
COMMITTEE MEMBER

  
Wyatt R. Murphy Jr., Ph.D.  
COMMITTEE MEMBER

  
Marian Glenn, Ph.D.  
COMMITTEE MEMBER

  
Daniel B. Nichols, Ph.D.  
COMMITTEE MEMBER

  
Angela V. Klaus, Ph.D.  
DIRECTOR OF GRADUATE STUDIES

  
Heping Zhou, Ph.D.  
CHAIRPERSON, DEPARTMENT OF BIOLOGY DEPARMENT

## **Acknowledgements**

Completion of this doctoral dissertation was possible with the support of several people. I would like to extend a sincere thank you to my co-mentor Dr. Ziping Zhang for exposing me to the world of Next Generation Sequencing and for providing continued guidance every step of the way. A profound gratitude goes to my co-mentor, Dr. Carolyn S. Bentivegna, for providing me the opportunity to join her lab. My Ph.D. project would not exist today without her wide range of creative ideas, knowledge, and willingness to explore a new field of science. I am forever indebted to Dr. B for her guidance, patience, and encouragement during my time in her lab. I am forever grateful to both the Biology and Chemistry departments for funding the development of the bioinformatics facility used during this project. A special mention goes to Dr. Murphy for his enthusiastic support in the creation of the workstation and server. One of these days we will come up with a creative name for it.

I would like to express special thanks and warm gratitude to my dissertation committee members: Dr. Jane Ko, Dr. Wyatt R. Murphy Jr., Dr. Marian Glenn, and Dr. Daniel B Nichols for taking their valuable time reviewing this project. Furthermore, I would like to acknowledge the entire Biology Department at Seton Hall University for the knowledge that I acquired during the time I spent obtaining my Masters and Ph.D. degrees.

I thank my fellow lab mates for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had.

A special thanks to my family. Words cannot express how grateful I am to my mother-in law, father-in-law, my mother, and father for all of the sacrifices that you've made on my behalf. I would like express appreciation to my beloved wife, Giovanna, who I am indebted for her personal support, strength, and great patience at all times. I am so grateful to have someone like you in my life.

Financial support for this study was provided by the Department of Biological Sciences of Seton Hall University through teaching assistantship and funding was provided by the NJ Environmental Protection Agency, PI-Carolyn S. Bentivegna, and Louisiana Department of Wildlife and Fisheries, PI- Ralph Portier (LSU), Co-PI, Carolyn S. Bentivegna.

## Table of Contents

<b>Title</b>	<b>Page</b>
Acknowledgement	iv
List of Tables	vii
List of Figures	viii
Abstract	x
Introduction	1
Materials and Methods	20
Results	39
Discussion	117
References	142
Appendix Publisher's Approval	156

## Tables

<u>Title</u>	<u>Page</u>
<b>Table 1.</b> Experimental outline establishing exposure times of crude oil and sacrificing schedule.	21
<b>Table 2.</b> Semi-quantitative PCR primer sequences and expected product size.	25
<b>Table 3.</b> qPCR primer sequences and expected product size.	37
<b>Table 4.</b> Average morphometric measurements consisting of weight and length for male control and exposure groups	41
<b>Table 5.</b> Statistics of the raw reads after Illumina sequencing and processing	68
<b>Table 6.</b> Statistics of the Assemblies	70
<b>Table 7.</b> BLASTX alignments from the six different assemblies against the southern platyfish and Amazon molly databases.	74
<b>Table 8.</b> BLASTX alignments of the six different assemblies to the CEGMA dataset.	76
<b>Table 9.</b> The Detonate's RSEM-EVAL scores	84
<b>Table 10.</b> Summary of the top three performers for each evaluation metric category.	86
<b>Table 11.</b> Statistics of Reference Derived Transcriptome Assemblies	89
<b>Table 12.</b> Cloned sequences aligned percent identities with sequences in GenBank	114
<b>Table 13.</b> Trinotate expression relationships with qPCR analysis	117



## List of Figures

<u>Title</u>	<u>Page</u>
<b>Fig. 1</b> Transcriptome analysis pipeline employed in this study.	32
<b>Fig. 2</b> Presented is the PCR for CYP19a and $\beta$ -actin in control killifish at (a) Day 0, (b) Day 1, (c) Day 3, and (d) Day 7 along with the corresponding densitometry analysis for (e) Day 0, (f) Day 1, (g) Day 3, and (h) Day 7 control gonads normalized to $\beta$ -actin.	43
<b>Fig. 3</b> Presented is the PCR for CYP19a and $\beta$ -actin in exposed killifish at (a) Day 1, (b) Day 3, and (c) Day 7 along with the corresponding densitometry analysis for (d) Day 1, (e) Day 3, and (f) Day 7 exposed gonads normalized to $\beta$ -actin.	45
<b>Fig. 4</b> Presented is the PCR for VTG and $\beta$ -actin in control killifish at (a) Day 0, (b) Day 1, (c) Day 3, and (d) Day 7 along with the corresponding densitometry analysis for (e) Day 0, (f) Day 1, (g) Day 3, and (h) Day 7 control livers normalized to $\beta$ -actin.	48
<b>Fig. 5</b> Presented is the PCR for VTG and $\beta$ -actin in exposed killifish at (a) Day 1, (b) Day 3, and (c) Day 7 along with the corresponding densitometry analysis for (d) Day 1, (e) Day 3, and (f) Day 7 exposed livers normalized to $\beta$ -actin.	50
<b>Fig. 6</b> Presented is the PCR for CYP1A and $\beta$ -actin in control killifish at (a) Day 0, (b) Day 1, (c) Day 3, and (d) Day 7 along with the corresponding densitometry analysis for (e) Day 0, (f) Day 1, (g) Day 3, and (h) Day 7 control livers normalized to $\beta$ -actin.	52
<b>Fig. 7</b> Presented is the PCR for CYP1A and $\beta$ -actin in exposed killifish at (a) Day 1, (b) Day 3, and (c) Day 7 along with the corresponding densitometry analysis for (d) Day 1, (e) Day 3, and (f) Day 7 exposed livers normalized to $\beta$ -actin.	54
<b>Fig. 8</b> Presented is the PCR for (a) CYP19, (b) CYP1A, (c) VTG, and $\beta$ -actin in non-sexually active (NSA) killifish along with the corresponding densitometry analysis for (d) CYP19a, (e) CYP1A, and (f) VTG normalized to $\beta$ -actin.	56
<b>Fig. 9</b> Expression levels for genes used to detect endocrine disruption at Day 7 in the crude exposure experiment and NSA.	59
<b>Fig. 10</b> Day 7 EXP, Day 7 CON, and NSA fish weights normalized to the relative mRNA expression for CYP19a, VTG, and CYP1a.	61
<b>Fig. 11</b> Representative EEMs for gall bladders of crude oil exposed or control killifish collected on Day 3 or 7.	64
<b>Fig. 12</b> Levels of fluorescence (CPS/ $\mu$ A) in gall bladders after Day 7 exposure to crude oil or control.	65

<b>Fig. 13</b> Phylogenetic tree analysis. A phylogenetic tree analysis of the 11 publically available fish genomes and killifish testis.	72
<b>Fig. 14.</b> BUSCO Analysis.	77
<b>Fig. 15</b> Full Length Transcript Analysis.	79
<b>Fig. 16.</b> Open Reading Frames Analysis.	81
<b>Fig. 17</b> BUSCO Analysis of the reference killifish transcriptome	91
<b>Fig. 18.</b> Volcano plot showing log fold-change and FDR for the comparisons of the sexually active exposed testis vs the sexually active control testis.	94
<b>Fig. 19.</b> Volcano plot showing log fold-change and FDR for the comparisons of the non-sexually active exposed testis vs the sexually active control testis.	96
<b>Fig. 20.</b> Cytoscape Network: Steroid Hormone Receptor Signaling Pathway.	98
<b>Fig. 21</b> Apoptosis heatmap.	101
<b>Fig. 22</b> Response to Heat heatmap.	104
<b>Fig. 23</b> Steroid Hormone Receptor Signaling Pathway heatmap.	107
<b>Fig. 24</b> Concentration Volcano Plot.	110
<b>Fig. 25</b> Changes in EGLN, GSTA4, and MMP14 gene expression measured by qPCR.	114
<b>Fig. 26</b> Changes in AK7, TDRD7b, and PDK1 gene expression measured by qPCR.	115
<b>Fig. 27</b> Correlation between transcriptome expression and qPCR (difference).	117
<b>Fig. 28</b> Candidate biomarkers.	142

## Abstract

The killifish, *Fundulus heteroclitus*, is a common fish model in aquatic toxicology. However, little is known in this organism about how endocrine disrupting compounds, (EDCs) including polycyclic aromatic hydrocarbons (PAHs), impact the reproductive status from a molecular standpoint. The objective of this project is to apply high throughput sequencing to *F. heteroclitus* testes to examine the molecular mechanisms that impair reproduction when exposed to crude oil. First, a crude oil exposure experiment was conducted. Following exposure, semi-quantitative PCR was performed to detect changes in gene expression of the following genes known to respond to EDCs: gonadal aromatase (CYP19a), vitellogenin (VTG), and cytochrome P450 1A (CYP1A). Excitation-Emission matrix (EEM) spectroscopy was performed to verify PAH exposure in the gonads. Killifish that demonstrated endocrine disruption along with verified PAH exposure were selected for sequencing. Illumina NextSeq 500 technology was applied to three experimental groups to determine genes and pathways turned on during sexual activation and disrupted by crude oil exposure: 1) an exposed spawning male gavaged with crude oil collected from the DeepWater Horizon oil rig, 2) a control spawning male gavaged with fish oil, and 3) a control non-spawning male. The *de novo* assembler, Bridger, was used to assemble the sequence reads. The Trinotate pipeline was used to annotate the resulting genes and determine differential gene expression among the three transcriptomes. The Trinotate annotation and differential expression analysis was validated by qPCR of select genes. Heatmaps displayed genes that were turned on, off, or had at minimum a two-fold change in expression. The Cytoscape plug-in ClueGO created

a functionally organized GO/pathway term network to visualize gene interactions. Genes found within the “Regulation of Androgen Receptor Signaling Pathway” node was shown to be impacted by GO terms associated with “Response to Heat” and “Apoptosis”.

Candidate biomarkers were found associated with apoptosis (EP300; histone acetyltransferase p300, SIRT1; Sirtuin 1 and SMARCA4; Transcription activator BRG1), impaired spermatogenesis (SMARCA4 and DNAJA1; heat shock protein family (Hsp40) member A1,) and suppressed androgen receptor transcriptional activation (HDAC6; (Histone deacetylase 6 and PTGES3; Prostaglandin E Synthase 3). Overall, this research is the first to assign functional annotations to the testes transcriptomes in killifish, resulting in the most comprehensive reproductive information available for the species to date. This transcriptomic data can provide the ground work for studying the killifish’s population dynamics, biomonitoring, and reproductive health. Additionally, this data can be utilized in comparative studies in other fish models to further enhance breeding programs and evolutionary studies.

## **Introduction**

Crude oil is one of the major pollutants in the marine environment. With the advancing global economy, the increased demand for crude oil has promoted the fast growth of offshore oil exploration and ocean transportation. The political will to make the U.S. more energy independent is likely to intensify crude oil production and thereby enlarge the number of impacted areas and intensify existing levels of contaminants. Accidental oil spills have been known to devastate the environment and affect the local economy as recently exemplified by the DeepWater Horizon oil spill in the Gulf of Mexico in 2010. Since spilled oil can go anywhere because of the tides and wind, increasing incidents of oil spills from oil exploration, production, and transportation activities can result in oil pollution all over the world. This oil pollution may influence population levels of marine species by negatively affecting their development, growth and reproduction. It is noteworthy that the accumulation of oil contaminants in various aquatic species not only debases the quality of the commercial aquatic products but also could affect human health if consumed.

In general, spills of large quantities of crude oil have the potential to cause severe short and long-term damage to marine ecosystems. The most visible and immediate short-term impact of oil on affected wildlife is its ability to adhere to the organisms, especially after the sludge washes ashore. The animals can die due to suffocation, ingesting the oil, or inhaling the toxic fumes (Peterson et al., 2001). In contrast to the short-term effects of crude oil spills, long-term biological impacts are more complicated

and significant at both the population and ecosystem level. This is because some components of the oil do not degrade quickly and therefore act through various toxic mechanisms that do not result in immediate or obvious mortality (Peterson et al., 2001; Peterson et al., 2003). There is evidence that 7–10 years after an oil spill, hydrocarbon exposure is correlated with reduced survival of some near shore vertebrate species including sea otters (Monson et al., 2000, Bodkin et al., 2002), sea ducks (Trust et al., 2000, Esler et al., 2000, Esler et al., 2002), and pigeon guillemots (Golet et al., 2002, Seiser et al., 2000). In 2001, a survey of formerly heavily/moderately oiled beaches within Prince William Sound estimates that approximately 60,000 L of Exxon Valdez crude oil remained in the intertidal zone since the 1989 oil spill, including deposits a few centimeters below the surface that were often still fluid and toxic (Short et al., 2003). The initial oil from the spill could be underneath rocks and boulders which allows it to persist for over ten years at study sites on Gulf of Alaska shores distant from the spill's origin (Irvine et al., 2006). Four years after the Prestige oil spill off the coast of Galicia in 2002, toxic compounds present in the oil such as polycyclic aromatic hydrocarbons (PAHs) are still being detected in the marine food chain (Laffon et al., 2006, Morales-Caselles et al., 2006, Ordás et al., 2007). The difficulty of easily containing and cleaning up oil spills leads to long exposure periods that may have many adverse genetic ramifications on aquatic wildlife within an ecosystem.

Exposure to crude oil and its components can potentially damage fishery resources. This occurs by affecting population densities of commercial fish species as well as species on which they feed. Population densities can be impacted when survival

and reproduction of oil-exposed fish are compromised. Several laboratory studies have documented oil-related declines in reproductive parameters in marine teleosts such as alterations in levels of reproductive hormones, inhibited gonadal development, and reduced egg and larval viability (Idler et al., 1995, Thomas et al., 1995, Truscott et al., 1992). Both crude oil and weathered oil byproducts are highly toxic to fish eggs and larvae (Incardona et al., 2004). Oil contamination may cause increased mortality of eggs and larvae even at low concentrations (McGurk et al. 1996). Exposure to oil and oil byproducts also leads to a range of sub-lethal effects on fish eggs and larvae, including premature hatching (Carls et al., 1999), morphological malformations (Hose et al., 1996) and genetic damage (Norcross et al., 1996). Low levels of dissolved oil hydrocarbons may also slow larval growth rates and affect their swimming ability and feeding behaviors (Tilseth et al., 1984). Mortality rates on malformed, premature or slow-growing larvae are likely to be extremely high (Carls et al., 1999).

PAHs are primary toxic constituents in crude oil (Whitehead et al. 2011). Research shows that crude oil exposure can be detected by measuring PAHs in fish tissues (Kreitsberg et al., 2010, Murawski et al., 2014) and bile (Lin et al., 1994, Fuentes-Rios et al., 2005). PAHs in contaminated aquatic environments enter the gastrointestinal tract through the diet. Here they are absorbed into the blood stream and passed into the hepatic portal vein of the liver where they are metabolized by the cytochrome P450 enzyme systems and excreted into bile and urine (Bentivegna et al., 2016). Fish under laboratory conditions have previously been shown capable of metabolizing and

eliminating PAHs within a few days following a single dose exposure (Varanasi et al., 1989).

PAHs are fluorescent aromatic compounds (FACs) that can be detected using synchronous fluorescence scanning (Lin et al., 1994), fixed wavelength fluorescence (Aas et al., 2000b), high performance liquid chromatography with fluorescence detection (HPLC-F) (Kreitsberg et al., 2010, Rey-Salgueiro et al., 2009), and Excitation-Emission Matrix(EEM) spectroscopy (Elcoro et al., 2014, Ferretto et al., 2014). However, these fluorescent methods are unable to differentiate the existence of one specific PAH from another due to the overlapping excitation/emission wavelengths used for detecting most FACs (Bentivegna et al., 2016). Therefore, “PAH-like” compounds including PAH metabolites typically found in natural samples are detected using the aforementioned FACs methods (Ariese et al., 1993). This is an advantage over the more reference driven method of gas chromatography-mass spectroscopy (GCMS) that detects only limited types of PAHs (Bentivegna et al., 2016).

PAHs are known to be endocrine disrupting compounds (EDC) (Whitehead et al., 2011). Chemicals that are able to cause endocrine disruption have one of three characteristics: they are persistent and bioaccumulative, are present at high concentrations, or they are constantly entering the environment (Tyler et al., 1998). EDCs have been found in freshwater, estuarine, and marine environments, raising the possibilities that EDCs can threaten population sustainability over time and disrupt those ecosystems (Mills et al., 2005). EDCs in aquatic environments are readily available to



fish through various routes that include aquatic respiration, osmoregulation, and maternal transfer of contaminants in lipid reserves of eggs (Van Der Kraack et al., 2001). Ingestion of contaminated food and skin contact with contaminated sediments are also exposure routes (Mills et al., 2005).

EDCs exert their effects in many ways by disrupting the natural hormone/receptor relationships involved in reproduction. For instance, they can mimic endogenous hormones by binding to hormone receptors and influencing cell signaling pathways, they can antagonize hormonal binding to hormone receptors, they also can alter the production and breakdown of natural hormones, and EDCs can modify hormone receptor levels (Sonnenschein and Soto, 1998; Matozzo et al., 2008). EDCs in particular may interfere negatively with the hypothalamus-pituitary-gonad-liver (HPGL) reproductive axis of organisms (Neubert et al., 1997).

In the hypothalamus, the KiSS (Kisspeptin) system is believed to be the mediating link to the response between environmental cues and metabolic signals that initiates the HPGL reproductive axis to induce oogenesis in female fish (Sempere et al., 2006). This system consists of the G-coupled protein receptor 54 (GPR54) that resides on neurons and produces gonadotrophin-releasing hormone (GnRH) along with its neuropeptide ligand encoded by the Kiss1 gene, kisspeptin (Zohar et al., 2010). Kiss1 is regulated by photoperiods, seasons in seasonally breeding animals, and metabolic factors that relay metabolic signals to the reproductive axis (Popa et al., 2008). When positively regulated by these factors, kisspeptin excites GnRH neurons by acting on GPR54 and causes the

release of GnRH. This results in stimulating the pituitary gland to secrete gonadotrophic hormones (GTH) (Popa et al., 2008).

The pituitary gland is known as the master gland because it controls the activities of the other endocrine glands (Davidovici et al., 2008). GTH, the main hormone that the pituitary gland releases to control the gonads, can be broken down into the follicle-stimulating hormone (FSH) and the luteinizing hormone (LH) (Mateos et al., 2002). In male fish, LH regulates steroidogenesis by activating LH receptors expressed on Leydig cells to stimulate androgen production and FSH regulates spermatogenesis by activating FSH receptors on Sertoli cells, respectively (Shulz et al., 2001).

In female fish, both FSH and LH ensure normal functioning of the ovaries with FSH being involved in the initiation of gametogenesis and regulation of gonadal growth and LH mainly regulating gonadal maturation and ovulation (Mateos et al., 2002). Once oogenesis is triggered, the pituitary gland will secrete these gonadotrophic hormones into the blood stream where they are carried to the ovaries to induce oocyte growth and ovulation (Nicolas et al., 1999).

GTH released from the pituitary glands also initiates vitellogenesis, the process of yolk deposition in oocytes that will provide energy reserves for the embryonic development of the offspring (Matozzo et al., 2008). This occurs when GTH released from the pituitary glands signals the follicle cells to synthesize estrogens (primarily 17 $\beta$ -estradiol (E2)) (Bermanian et al., 2004). Another source of 17 $\beta$ -estradiol for the induction of vitellogenesis is derived from cytochrome P450 aromatase (CYP19), which converts

androgens into estrogens (Cheshenko et al., 2008). The 17 $\beta$ -estradiol is released into the blood and transported into the liver where it enters the hepatocytes by diffusion and binds with high affinity to the estrogen receptor alpha (ER $\alpha$ ) (Bermanian et al., 2004). The activated ER $\alpha$  triggers the expression of vitellogenin (VTG) along with its own gene expression (Bermanian et al., 2004). The VTG is transported via the vascular system to the ovary and incorporated by receptor-mediated endocytosis into yolk platelets under gonadotrophin promotion (Nicolas et al., 1998). The VTG is then cleaved to form the phosphorus containing phosvitin protein and two lipid containing proteins, lipovitellins I and II (Davail et al., 1998).

Any factor causing an impairment of the reproductive HPGL axis, vitellogenic cycle, or steroidogenesis could have detrimental influences on oogenesis, spermatogenesis, fecundity, viable embryos, hatching rate, and larval survival (Anderson et al., 1996; Nicolas et al., 1998). Negative effects in any of these circumstances would lead to reproductive deficits in fish populations. Because of these actions, increasing attention has been given to evaluating adverse effects on reproduction and development caused by EDCs in aquatic environments.

Many studies have shown that PAHs can have a negative effect on vitellogenesis in fish (Nicholas, 1999). A biomarker to assess a population's vitellogenic reproductive health is the induction of cytochrome P4501A (CYP1A) in the liver. Aryl hydrocarbon receptor (AHR) ligands, such as halogenated aromatic hydrocarbons and PAHs, are capable of inducing CYP1A expression while disrupting 17 $\beta$ -estradiol-induced

expression of the VTG and reducing ER $\alpha$  levels (Bemanian et al., 2004). This is supported by a study by Bugel et al (2010) that showed an inverse relationship between hepatic CYP1A protein and hepatic VTG mRNA expression in *F. heteroclitus*. This suggests a possible link between AhR agonist exposure and vitellogenesis. It is also known that CYP1A induction can increase the metabolism of estrogens and the lack of estrogen levels correlated to the lack of VTG (Navas et al, 2000). Research in rainbow trout (*Oncorhynchus mykiss*) shows that PAHs are able to exert their antiestrogenic activity by binding to AhR in cultured hepatocytes and inhibiting estradiol-regulated VTG synthesis (Navas et al., 2000). Another study that showed that PAHs can induce CYP1A expression was done by Meyer (2002). This study revealed that CYP1A protein expression was induced by the PAH-type, CYP1A inducer  $\beta$ -naphthoflavone by utilizing killifish, *F. heteroclitus*. Debunsky (2013) found that killifish exposed to the oil from the Deepwater Horizon oil spill had divergent gene expression in the liver and gill tissue coincident with the arrival of contaminating oil. Changes in expression included the up-regulation of cytochrome P4501A (CYP1A) protein in gill, liver, intestine and head kidney for over one year following peak landfall of oil (August, 2011) compared to fish collected from reference sites. Their results suggest that crude oil exposure could alter vitellogenesis due to the presence of CYP1A agonists such as PAHs.

EDCs can also exert negative effects on cytochrome P450 19a (CYP19a) expression. Inhibiting expression of CYP19a limits the conversion of androgens into 17 $\beta$ -estradiol that is required for the induction of vitellogenesis. The study by Patel (2006) shows that female killifish exposed to the PAH, benzo(a)pyrene (BaP), in a water-borne

study for 15 days experienced inhibited ovarian aromatase (CYP19a) activity (Patel et al., 2006).

A great laboratory model organism for evaluating effects of oil exposure on reproductive health is the Atlantic killifish (*Fundulus heteroclitus*). Killifish are a promising biosensor for aquatic pollution because they are a commonly studied sentinel teleost species (Burnett et al., 2007), have a narrow home range and high site fidelity (Teo et al., 2003), are sensitive to organic pollutants (Van Veld et al., 2008), and can be maintained in a laboratory setting. There is evidence that PAH exposure results in reproductive and developmental deficits in killifish collected from PAH-impacted sites (reviewed in Nicolas, 1999). The reported effects reviewed by Nicolas include the reduction in circulating hormones and plasma VTG, estrogenic and antiestrogenic effects, retardation of oocyte maturation and reduction of reproductive success. (Nicolas, 1999). Gulf killifish (*Fundulus grandis*) were used as a model species to investigate the consequences of the Deepwater Horizon oil spill because they are among the most abundant vertebrate animals in the Gulf of Mexico-exposed marshes (Whitehead et al., 2011). Sub-lethal effects detected at the molecular level were used to predict long term population-level impacts of oil pollution. Genome expression profiles, using microarrays and RNAseq, were used to characterize the killifish liver, which is the primary tissue for metabolism of toxic oil constituents (Whitehead et al. 2011). Tissue morphology and expression of CYP1A protein were characterized for gills along with exposing developing embryos to field-collected water samples to document bioavailability and bioactivity of oil contaminants for the early life stages (Whitehead et al. 2011). The study

found that biologically relevant sub-lethal exposures to crude oil can cause alterations in genome expression and tissue morphology suggestive of physiological impairment lasting for over two months after initial exposure (Whitehead et al. 2011).

To date, no research has been presented on the network analysis of the killifish testes transcriptomes to identify gene modules and candidate genes associated with testis-derived reproductive disruption from the mechanistic point of view. Key genes associated with disrupted physiological function of gonads need to be identified by comparing transcriptomes of oil exposed groups versus control groups under laboratory conditions. Modern biomarkers need to be developed in order to protect our natural resources and to assess natural resource damage caused by oil spills. Those biomarkers need to link molecular responses to population level effects. In doing so, impaired reproductive responses will be associated with detectable molecular responses.

Over the past decade, significant progress has been made in genome wide gene expression profiling by the development and application of differential display (Liang et al, 1992; Shen et al., 2009), RNA fingerprinting (McClelland et al., 1995), serial analysis of gene expression (SAGE) (Velculescu et al., 1995), suppression subtraction hybridization (Brulle et al., 2012), cDNA AFLP (Breyne et al., 2003), and others. These technologies have identified and characterized different gene functions, different signal pathways, and have validated drug target interactions. However, each of the above techniques has disadvantages. For instance, high false positive rates, being time and labor intensive, and most importantly these techniques identify genes with short fragment sizes

that severely limit downstream efforts of reliable annotation (Debouck et al., 1995). Microarrays are another popular gene expression profiling technique that has several limitations, such as: dependence upon existing knowledge about genome sequences; high background levels owing to cross-hybridization; and a limited dynamic range of detection owing to both background and saturation of signals (Wang et al., 2010). Moreover, comparing expression levels across different experiments is often difficult and can require complicated normalization methods (Wang et al., 2010). In gene-expression studies, microarrays are now being replaced by Next Generation Sequencing (NGS) based methods (Metzker et al., 2010). In contrast to microarray methods, NGS has clear advantages and is expected to revolutionize the manner in which eukaryotic transcriptomes are analyzed (Cloonan et al., 2008; Morozova et al., 2008; Ozsolak et al., 2010; Wang et al., 2009). This project will utilize NGS to detect key genes involved in mechanisms of gonadal disruption in killifish.

NGS techniques are new “transcriptome” methods of gene expression analysis that provide general representation of all transcripts (i.e., mRNAs) expressed in particular cells or organs under particular conditions and exposure times. NGS reduces the cost of DNA sequencing by over two orders of magnitude, making global transcriptome analysis inexpensive, routine and widespread (Shendure et al., 2008). NGS allows researchers to accurately probe the current state of a transcriptome and assess many genetically important issues, such as; gene expression levels, differential splicing events, and allele-specific gene expression (Shen et al., 2011). NGS has a clear advantage over older gene expression technologies (e.g., microarrays, etc.) in that it is not limited to experimental

systems having well characterized genomes or transcript sequence libraries (Shen et al., 2011). This positions NGS as an important technique that can provide new opportunities for better characterization of experimental systems and species for which a whole genome sequence is lacking or unavailable (Feldmeyer et al., 2011; Wang et al., 2010; Xiang et al., 2010).

Starting in 2005, a variety of massively parallel sequencing instruments such as the Roche/454, the Life Technologies SOLiD, and the Illumina platforms were used to sequence human and model organism genomes. This method has already been applied to a number of model organisms, such as yeast, Arabidopsis, Drosophila, mouse, and human (Graveley et al., 2010; Marioni et al., 2008; Morin et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008). Only a few studies have used NGS technologies to research the impact of environmental contaminants in aquatic organisms (Debunsky et al., 2013 and Yednock et al., 2015).

NGS technologies have offered unprecedented opportunities to obtain genetic information for non-model organisms with little or no molecular information available (Gordo et al., 2012). This increasingly accessible technology provides an efficient and cost-effective approach for analyzing the transcriptome of non-model organisms that lack a fully-sequenced genome (Fan et al., 2013; Garg et al., 2011; Sharma et al., 2014). It has been employed to identify novel transcriptome sequences, single nucleotide polymorphisms (SNPs), simple sequence repeats (SSRs), splicing variants, transcript isoforms, new large intergenic noncoding RNAs, and relative levels of transcript



expressions (Gordo et al., 2012; Fan et al., 2013; Sharma et al., 2014; Nandety et al., 2013). With the arrival of NGS technologies, the number of publications characterizing *de novo* assemblies for non-model organisms have steadily been on the rise (Ashrafi et al., 2012).

*De novo* transcriptome assembly is performed by taking the enormous amount of short read sequences produced by NGS and overlapping them to form contiguous sequences (contigs) (Haznedarogulu et al., 2012). The quality of the assembly output is reliant on the user designated *k-mer* value defined as the sequence overlap between two reads forming the contig (Haznedarogulu et al., 2012; Moreton et al., 2014; Surget-Groba et al., 2010; Robertson et al., 2010). Low *k-mer* values have a tendency to recover less abundant transcripts, while producing a large amount of contigs, with a number of them highly fragmented due to sequencing errors and lack of overlap (Surget-Groba et al., 2010; Chopra et al., 2014). High *k-mer* values will produce a more contiguous assembly consisting of high coverage transcripts and splice variants. However, the assembly will contain fewer contigs leading to lower transcript representation (Surget-Groba et al., 2010; Chopra et al., 2014). Therefore, utilizing a single *k-mer* approach when performing a *de novo* assembly can result in loss of relevant biological information due to the lack of transcript diversity (Chopra et al., 2014). A logical approach to resolve this dilemma is to cluster multiple single *k-mer* assemblies together in order to take advantage of the characteristics of both the low and high *k-mer* values and thereby improve the accuracy of the assembly (Haznedarogulu et al., 2012; Moreton et al., 2014; Surget-Groba et al., 2010).

The number of *de novo* transcriptome programs developed for assembly of short sequence reads has increased within the past few years. In 2010, Trans-ABYSS was reported to have the ability to merge multiple individual *k-mer* assemblies, allowing the transcriptome to be represented by wide levels of transcript expression (Robertson et al., 2010). In 2011, Trinity was reported to be able to fully reconstruct a large fraction of transcripts with low base error rates and have the ability to report alternative splice isoforms (Grabherr et al., 2011). Trinity is currently regarded to be the best single *k-mer* assembler (He et al., 2015). In 2012, Oases was reported to improve significantly on the Trans-ABYSS and Trinity assemblers by merging the use of multiple *k*-mers presented in Trans-ABYSS with a topological analysis similar to that presented by Trinity (Shulz et al., 2012). In 2014, SOAPdenovo-Trans was reported to be able to perform multiple individual *k-mer* assemblies and provide higher contiguity, lower redundancy and faster execution when compared to Trinity and Oases (Xie et al., 2014). All of these assemblers are founded on the *de Bruijn* graph-based assembly method to which programmers add their own algorithms (Robertson et al., 2010; Grabherr et al., 2011; Shulz et al., 2012; Xie et al., 2014). In 2015, Bridger, which employs a new *de novo* assembly method that does not construct *de Bruijn* graphs, was created (Chang et al., 2015). This assembler uses a rigorous mathematical model, called the minimum path cover, to construct splice graphs that are used to build compatibility graphs for transcriptome reconstruction from short RNA-seq reads (Chang et al., 2015). This multiple *k-mer* assembler aims to build a bridge between the key concepts of two popular assemblers, the reference-based assembler, Cufflinks (Trapnell et al., 2010), and the *de novo* assembler, Trinity (Chang et

al., 2015). Given the number of programs available, there is a need for more definitive information on what assemblers and parameters work best for constructing a *de novo* transcriptome (Moreton et al., 2014; Zhao et al., 2011).

Much effort has been dedicated towards improving the capabilities of *de novo* assembly software. However, methods to evaluate the assembler's performance are still a few steps behind. For example, a common approach to assess genome assemblies is to evaluate statistics such as the number of contigs, the amount of contigs over 1,000 bps, and the N50 value. The N50 value is defined as the length of the largest contig from all the contigs ranked smallest to largest that represents 50% of the assembly length (Li et al., 2014; Baker 2012). However, these metrics are also routinely used to evaluate the quality of *de novo* transcriptome assemblies even though they may be misleading regarding their accuracy (Li et al., 2014; Baker 2012; O'Neil et al., 2013). Evaluating mRNA characteristics such as the percentage of assembled full length transcripts and the number of long open reading frames (ORFs) are other common metrics for evaluating transcriptome assemblies (Chopra et al., 2014; He et al., 2015; Nakasugi et al., 2014). A novel reference-free evaluation method to assess the quality of transcriptomes is Detonate RSEM-EVAL (Li et al., 2014). This program produces a statistically principled evaluation score using multiple factors, such as the compactness of the assembly and its support from the RNA-Seq reads used to create it (Li et al., 2014).

Annotation-based metrics describe the percentage of sequences within an assembly that match protein sequences found in a related species or curated database

(Schliesky et al., 2012). An accurate assembly should contain a high percentage of these conserved proteins while a low percentage reflects mis-assemblies. These types of metrics can be challenging for non-model species that do not have an annotated phylogenetically related species to which it can be aligned. If the evolutionary distance between the two species is too great, then orthologs may have undergone nucleotide changes making alignments less likely to occur (O'Neil et al., 2013). A database to gauge the performance of an assembly strategy in the absence of a well annotated phylogenetically related species is the CEGMA (Core Eukaryotic Genes Mapping Approach) database (Parra et al., 2007). This is a manually curated database that contains 248 core proteins present in a wide range of taxa (Parra et al., 2007). The same approach can be achieved by utilizing the manually curated BUSCO (Benchmarking Universal Single-Copy Orthologs) protein set for the quantitative assessment of transcriptome assembly and annotation completeness (Fan et al., 2013). Using this database, assemblies are matched to 3,023 vertebrate genes and results return the percentage of unaligned sequences. Annotation metrics are useful for assessing both genome and transcriptome assemblies.

Evaluation metrics are important for assessing the quality of genome and transcriptome assemblies. Unfortunately, there is a lack of consensus as to which evaluation metrics work best let alone how many of them to use. For example, Chropra *et al.* performed a comparison study using peanut (*Arachis* spp.) RNA-Seq data. Assemblies were evaluated based on the N50 length, average contig length, number of contigs, the novelty of each assembly using the Mummer tool, the accuracy determined by RMBT,

and the continuity by estimating the number of full length transcripts (Chopra et al., 2014). Moreton *et al.* relied on the RMBT and CEGMA percentages as well as the N50 length, the number of transcripts, and the number of transcripts >1kb when evaluating different assemblies of the duck (*Anas platyrhynchos*) (Moreton et al., 2014). He *et al.* assessed the assemblies of sweet potato fungus (*Trametes gallica*) and wild rice (*Oryza meyeriana*) based on N50 length, average contig length, number of contigs > 1,000 bps, as well as their ORFs and percent annotation using BLASTX results against phylogenetically related species (He et al., 2015). More information on which evaluation metrics best predict the quality of *de novo* transcriptome assemblies would help establish “best practices” particularly for less experienced users. Therefore, this study presented compared assembly programs, *k-mer* strategies, and various metrics for determining *de novo* transcriptome assembly quality. Based on the eleven evaluation metrics, it was found that the product of those assemblies was more influenced by the assembler itself than the *k-mer* strategy. Overall, Bridger performed more often within the top three of each evaluation metric than the *de Bruijn* graph-based programs for the *de novo* transcriptome assembly of the killifish RNA-Seq reads.

The specific hypothesis of the proposed research is that crude oil, which contains PAHs, is responsible for gonadal steroidogenesis disruption leading to decreased reproductive health and success in killifish. This proposed research will focus on utilizing NGS to profile the global expression patterns in the testis transcriptome of *F. heteroclitus* in response to crude oil exposure in a laboratory setting. The outcome of this investigation will identify gene modules and candidate genes associated with disrupted

steriodogenesis that causes reproductive failure at the functional genomic level. The NGS data will be analyzed to investigate the molecular mechanisms of PAH-mediated inhibition of endocrine and stress related responses and to generate potential biomarkers of reproductive stress from crude oil exposure. The NGS data will also be used to study the differences in gene expression of sexually active and sexually non-active testis.

The first aim involved determining exposure times at which the reproductive system of *F. heteroclitus* responds to crude oil. Wild fish were exposed by using the gavage method and endocrine disruption was detected over time using quantitative PCR and established biomarkers: CYP1A, CYP19a, and VTG. Exposure to PAHs was verified using EEMs using bile. Based on these outcomes, fish were selected for transcriptome analysis.

The second aim was to generate transcriptomes of the selected gonads of male and female fish. These groups consisted of 1) an exposed spawning male and exposed vitellogenic female gavaged with crude oil collected from the DeepWater Horizon oil rig prior to the accident; 2) a control spawning male and vitellogenic female gavaged with fish oil; 3) a control non-sexually active male and female. The Illumina NextSeq 500 sequencing was performed by the Waksman Genomics Facility.

The third aim was to determine the assembly method for the transcriptomes. A *de novo* assembly comparison study was performed to investigate which assembly program (Trinity, Bridger, Oases/Velvet, or SOAPdenovo-Trans) and *k-mer* strategy would be best for the *de novo* transcriptome assembly of the killifish reads based on eleven

evaluation metrics. This analysis was performed on a non-sexually active male killifish. Based on the evaluation metrics performed, it was determined that the Bridger assembly was able to construct the best assembly of the testis transcriptome in *F. heteroclitus*. Therefore, Bridger was used to construct a reference transcriptome with a multi *k-mer* approach. The paired end raw data files from each of the eight killifish gonads were interlaced to form the files used for assembly and mapping.

The fourth aim was to use various bioinformatics tools to determine differential gene expression levels in the individual transcriptomes and to visualize Gene Ontology (GO) interactions associated with crude oil exposure. Genes associated with endocrine disruption and stress responses were of particular interest. Genes from the GO categories with a log fold change of at least  $\pm 2$  were incorporated into a Cytoscape network to visualize the complex pathways and integrated activities of the genes. Heatmaps were generated to display the individual genes within the Cytoscape pathways of interest.

The fifth aim was to validate the annotated transcriptome and expression analysis. This was accomplished by Sanger sequencing and qPCR of selected PCR products. Overall, this project utilizing a transcriptome approach to identify new candidate biomarkers related to endocrine disruption caused by crude oil exposure. This data will enhance the capabilities for studying the killifish's population dynamics, biomonitoring, and reproductive health to provide benchmarks for comparative studies for other fish models.

## Materials and Methods

### *Fish collection*

Fish identified as Atlantic killifish (*Fundulus heteroclitus*) were collected from Tuckerton, NJ (Little Sheepshead Creek) using baited minnow traps on July 12<sup>th</sup>, 2013. The authority who issued the permission for capturing the killifish was The New Jersey Department of Environmental Protection; Division of Fish and Wildlife (Permit #1125). The vertebrate work done in this study was approved by The Rutgers University's Institutional Animal Care and Use Committee (IACUC) (Protocol #08-025). In all instances the fish were alive when captured, and capture methods followed approved animal handling protocols. The fish were transported immediately back to Rutgers University in aerated containers containing water from the collection site to reduce stress.

### *The Gavage Method*

The fish were acclimated to laboratory conditions for two weeks. The control and exposure group were maintained in two different tanks. The gavage method was employed to expose the control group with 25  $\mu$ L of 100% fish oil while the treatment group received 25  $\mu$ L of 50% fish oil and 50% crude oil. The fish oil was obtained directly from DayBrook Fisheries (<http://www.daybrook.com/>) and was produced commercially from Gulf menhaden collected in 2009. The crude oil was Macondo 252 (MC252) collected from the DeepWater Horizon oil rig prior to the accident and other oil rigs tapped into the same crude oil source. Both the control and exposure groups were dosed for three consecutive days, Day 0 through Day 2 as seen in Table 1.



**Table 1.** Experimental outline establishing exposure times of crude oil and sacrificing schedule.

<b>Schedule</b>	<b>Day 0</b>	<b>Day 1</b>	<b>Day 2</b>	<b>Day 3</b>	<b>Day 4</b>	<b>Day 5</b>	<b>Day 6</b>	<b>Day 7</b>
<b>Start Time</b>	9:00 AM	9:00 AM	9:00 AM	9:00 AM				9:00 AM
<b>Sacrifice time</b>	Time Zero	Day 1		Day 3				Day 7
<b>Dose time</b>	Dose	Dose	Dose					

All images, figures, and tables in this dissertation are generated by the author unless otherwise cited.

### *Concentration-Response Experiment followed by Euthanization of Sexually Active Fish*

Sexually active killifish were sacrificed on Day 0, Day 1, Day 3, and Day 7 by euthanizing them with an overdose of MS-222 (tricaine methanesulphonate) followed by spinal cord dislocation as seen in Table 1. Following the sacrificing, gonads, liver, and gall bladder were collected from each individual fish. Gonads and livers were divided into two parts. One part was used for PAH detection using EEMs. The second part was stored in RNAlater (Qiagen) at -20°C prior to RNA extraction. The bile from the gall bladder was subjected to EEMs analysis only to determine PAH exposure. Measurements included sex, weight (g) and length (cm) fish.

### *Non-Sexually Active Fish*

Non-sexually active killifish were housed under laboratory conditions at Rutgers University for five months before they were sacrificed on December 6<sup>th</sup>, 2013 with an overdose of MS-222 followed by spinal cord dislocation. Following the sacrificing, gonads, liver, and gall bladder were collected from each individual fish. As above, gonads and livers were divided into two parts, one for EEMs and one for RNA, and bile from the gall bladders was subjected to EEMs. Measurements included sex, weight (g) and length (cm) fish.

### *Extraction of PAHs from gonads and gall bladders*

Intact gall bladders were used for extraction of PAHs. The gall bladders were individually homogenized mechanically in 500 µL of 75% ethanol. After homogenizing,

500  $\mu$ L of 75% ethanol was added to the homogenate. The samples were vortexed for 1 minute continuously in order to extract PAHs then centrifuged for 20 minutes at 13,000 rpms to remove the tissues. After centrifugation, the supernatant was pipetted into another eppendorf tube for fluorescence analysis.

### *Fluorescence Analysis*

Gall bladder supernatants were analyzed using a Fluorolog-3 Spectrofluorometer (Horiba Jobin Yvon, Inc., Edison, NJ), equipped with single excitation and emission monochromators and non-ozone producing 450 W xenon arc lamp source. A Hamamatsu R928 side on photomultiplier tube was used to collect the emitted photons, and yielded a signal in photon counts per second (CPS). Samples were analyzed in 1 mL fused silica cuvettes with 1 cm excitation pathlengths. The excitation scans were from 260 to 400 nm and emission scans were from 320 to 480 nm. Lamp intensity variations were corrected using a photodiode reference signal ( $R1/\mu A$ ). The fluorescence intensity values (color on the contour maps) were represented as photon labeled counts per second per microamps ( $CPS/\mu A$ ). The Fluorolog 3 provided 3D contour maps showing excitation scans for multiple emission wavelengths as heatmaps with the red color representing high intensity and blue color representing low intensity fluorescence. The contour map separated fluorescent compounds based on their optimal excitation and emission wavelengths. Figures of contour maps were generated by SigmaPlot version 13.

### *RNA isolation*

Total RNA was extracted from the testes using TRIzol® Kit (Invitrogen™) following the manufacturer's instructions. Potential genomic DNA contamination in the RNA sample was removed by DNase I digestion (Ambion, Inc.). Using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Inc.), the RNA was quantified by measuring the absorbance at 260 nm. The purity of the RNA sample was assessed at an absorbance ratio of 260/280 and the integrity of the total RNA was determined by gel electrophoresis separating the samples on a 0.7% agarose gel stained with ethidium bromide (data not shown).

#### *RT-PCR and PCR*

Reverse Transcription-PCR was performed using the DNase free RNA as a template and oligo-dt primers to synthesize the cDNA (Applied Biosystems, Foster City, CA). The cDNA was used as a template to amplify genes of interest with PCR. Genes targeted for quantification included Beta Actin, CYP1A, CYP19a and VTG. PCR products were produced using a Multi Gene II (Labnet International, NC). Primer sets for these genes have been established along with their expected product size (Bugel et al, 2010) as seen in Table 2. The PCR reaction for all four genes was initiated with denaturation at 94 °C for 3 minutes, followed by 35 amplification cycles at 94 °C for 15 seconds for denaturing, 60 °C for 30 seconds for annealing, and 72 °C for 1 minute for elongation. The PCR product size was confirmed by separating them on a 2% agarose gel stained with ethidium bromide. To calculate the relative concentrations for each gene, densitometry analysis was performed using SoftMax Pro 5.3.

**Table 2:** Semi-quantitative PCR primer sequences and expected product size.

<b>Gene</b>	<b>Forward Primer (5'-3')</b>	<b>Reverse Primer (5'-3')</b>	<b>Product Size</b>
<b>CYP1A</b>	TGTTGCCAATGTGATCTGTG	CGGATGTTGTCCTTGTCAA A	258 bp
<b>VTG</b>	AGGATTCGTCCGAACAACAC'	TTTCAGACGGCACTCAGAT G	416 bp
<b>CYP19a</b>	ACGAGAAAGAGCTGCTGCTGA AGA	TGATGTCCAGCTTATCTGC CTGCT	198 bp
<b>β-actin</b>	GCTCTGTGCAGAACAACCACA CAT	TAACGCCTCCTTCATCGTT CCAGT	136 bp

### *Next Generation Sequencing*

Killifish gonads were selected for transcriptome sequencing based on the results from the EEMs data, PCR analysis, and RNA availability. Total RNA (10 µg) was used for library construction and subjected to Illumina's NextSeq 500 sequencing at the Waksman Genomics Facility (<http://www.waksman.rutgers.edu/genomics/home>). This resulted in three male and three female transcriptomes: 1) an exposed spawning male and vitellogenic female gavaged with crude oil collected from the DeepWater Horizon oil rig prior to the accident; 2) a control spawning male and vitellogenic female gavaged with fish oil; 3) a control non-spawning male and non- vitellogenic female.

### *Illumina short-read library construction and sequencing*

Ribosomal depletion and mRNA selection was performed with MicroPolyA purist (Ambion). The mRNA was quantified and ribosomal RNA fractions under 2% were verified using a BioAnalyzer mRNA Nano Kit (Agilent). The dUTP-strand specific cDNA library was made using chemical hydrolysis and the Illumina Ultra Directional RNA-seq kit (NEB). Libraries were barcoded with Tru-Seq adaptors and amplified with 12-15 cycles of PCR (Illumina). Completed RNAseq libraries were quantified using Qubit DNA HS, BioAnalyzer High Sensitivity DNA (Invitrogen, Agilent, KAPA). The libraries were sequenced using the NextSeq 500 High Output 300 cycles kit, reading 155 x 155 bp Paired End Sequencing.

### *Sequence Data Processing*

The quality of the raw Illumina sequence reads was initially assessed using FastQC v0.10.1 (Andrew 2010). Based on the analysis report, Trimmomatic v0.32 was used to remove all the low quality reads with a Phred score below 20 as well as the Illumina adapters (Bolger et al., 2014). Contaminating sequences were removed from the reads by using Deconseq with the parameters set to 90% of the contig length with an identity of 94% (Schmieder et al., 2011). FastQC was performed again to verify the integrity of the remaining raw Illumina sequence reads. Upon completion, the quality assessed reads were then ready to be used as the input for the various assembly strategies.

#### *De novo transcriptome assembly*

To validate which transcriptome assemblers work the best, RNA from the sexually active male killifish was used. Six different assembly strategies were created from four different assemblers: Trinity (v2.0.6), Velvet (v1.2.07) and Oases (v0.2.08), SOAPdenovo-Trans (v1.03), and Bridger (He et al., 2015; Schulz et al., 2012; Xie et al., 2014; Chang et al., 2015; Zerbino et al., 2008). The Trinity assembly was the only single *k-mer* assembly. It was run with its default *k-mer* value of 25. The Bridger, Oases, and SOAPdenovo-Trans assemblies were performed with multiple *k-mer* strategies. Bridger, Oases, and SOAPdenovo-Trans used a small multiple *k-mer* (SMK) strategy consisting of the *k-mer* lengths of 21, 25, 27, 29, 31, and 33. The SMK strategy was based on the limitations of the Bridger assembler, which can only use *k-mers* values up to 33. Additionally, Oases and SOAPdenovo-Trans were performed using a large multiple *k-mer* (LMK) strategy consisting of *k-mer* lengths of 25, 35, 45, 55, 65, 75, and 85. All six

assembly strategies incorporated the *k-mer* value of 25 to better compare their performances amongst each other. For the multiple *k-mer* strategies used in the Oases, Bridger, and SOAPdenovo-Trans assemblies, all seven individual *k-mer* assemblies from each group were concatenated followed by CD-HIT-EST (v 4.6.1) to further remove the redundancy and to cluster the contigs for annotation (Li et al., 2006)

### *Statistics of Assemblies*

The six different assembly strategies were assessed using typical statistics for the evaluation of *de novo* genome assemblies. These included the total number of contigs produced, each assemblies N50 length, and the amount of contigs over 1,000 bps long. These statistics were determined using Transrate (v1.0.0 beta3)

<http://hibberdlab.com/transrate/>.

### *RMBT Analysis*

The accuracy of each assembly was assessed by determining the percentage of raw reads that could be mapped back to transcripts (RMBT). First, indexes were generated using Bowtie2-build (Langmead et al., 2012) Then Bowtie2 (v2.2.5) was used to map the reads against each assembly and provide the metric of accuracy, which is the percentage of raw reads that align (Langmead et al., 2012)

### *Phylogenetic Tree Alignments*



Another common assessment tool to evaluate the quality of *de novo* transcriptome assemblies is to align the assembled contigs to a well-annotated phylogenetically related species. Quality is based on the percentage of contigs that match the protein sequences of the related species. To determine which closely related fish genome to use as a reference, PhyloT (<http://phylot.biobyte.de/contact.html>) was used. This program generated a phylogenetic tree between killifish and the eleven publically available fish genomes on Ensembl: Amazon molly (*Poecilia Formosa*), Mexican tetra (*Astyanax mexicanus*), Atlantic cod (*Gadus morhua*), Japanese pufferfish (*Takifugu rubripes*), medaka (*Oryzias latipes*), southern platyfish (*Xiphophorus maculatus*), spotted gar (*Lepisosteus oculatus*), stickleback (*Gasterosteus aculeatus*), green spotted pufferfish (*Tetraodon nigroviridis*), Nile tilapia (*Oreochromis niloticus*), and zebrafish (*Danio rerio*). This program creates trees based on the NCBI taxonomy database, and it was visualized by the web based tool, Interactive Tree of Life (v2) (Letunic et al., 2011). Based on the created phylogenetic tree, southern platyfish (*X. maculatus*) and Amazon molly (*P. formosa*) were shown to be the closest relative to killifish. Therefore, the six different assemblies were aligned to the Ensembl proteins of southern platyfish and Amazon molly using BLASTX with an E-value cut-off of 1e-3 to quantify the percentage of previously annotated genes found in each assembly.

#### *CEGMA and BUSCO Alignments*

CEGMA (v.2.5) and BUSCO (v1.1) are other reference-based tools for assessing the degree of annotation. They were individually employed to quantitate assembly

completion based on the percentage of contigs that do or do not (therefore mis-assemble) align to highly conserved proteins (Simão et al., 2015).

#### *Full Length Transcript Analysis*

The number of full length transcripts was quantified to further evaluate the performance of each assembly by following scripts provided by the Trinity software package (<http://trinityrnaseq.sourceforge.net/>). A modified ‘BLASTX’ script was used to calculate each assembly’s alignment coverage to the curator-evaluated database, SwissProt. Full-length transcripts were defined in this study by having > 70% alignment coverage and > 90% alignment coverage to SwissProt proteins.

#### *Open Reading Frames Analysis*

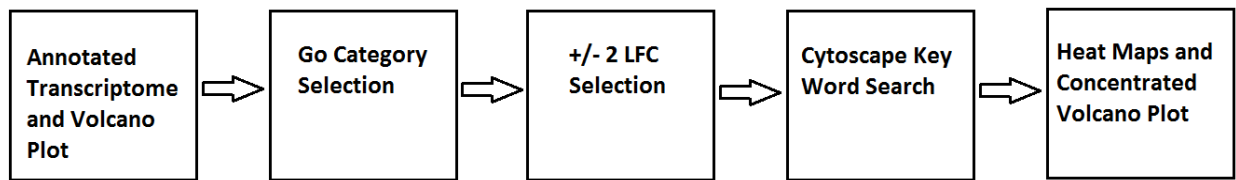
The presence of long open reading frames (ORFs) was analyzed to determine the quality of each assembly by using scripts provided by TransDecoder (<http://transdecoder.github.io/>). BlastP (v 2.2.30+) was used to search the protein database SwissProt with an E-value cut-off of 1e-3. ORFs ranging from >799 bps, >999 bps, and >1,199 bps were determined using gawk to filter the Fasta files by length.

#### *Detonate RSEM-EVAL Score*

DETONATE’s RSEM-EVAL was used to evaluate the quality of the six different assemblies. This program offers a reference-free evaluation method that relies only on the assembly and the reads used to create it (Baker et al., 2012).

### *Visualization and Analysis of Transcriptome Data*

The transcriptomes were analyzed using the pipeline displayed in Fig. 1, and each step is further described in detail below. Briefly, the Trianotate pipeline was utilized to annotate each *de novo* assembled transcriptome and to determine the differential expressed genes between the EXP vs CON groups and the NSA vs CON groups. One of the outputs from the differential expression analysis was the global profile of the total amount of differentially expressed genes between the EXP vs CON groups and the NSA vs CON groups in the form of a volcano plot. GO categories related to crude oil exposure and endocrine disruption were used to filter out matching annotated GO terms from each transcriptome. The resulting list of genes was further filtered by selecting the differentially expressed genes with a log-fold change (LFC) of  $\pm 2$ . This list was then imported into Cytoscape to visualize molecular interaction networks and biological pathways related to endocrine disruption due to the crude oil exposure. Key search terms such as “Androgen” were used to search through the large Cytoscape network to find interacting clusters of nodes of interest. Genes in these nodes were further analyzed with heatmaps to visualize the LFC between treatment groups. Key genes of interest were further highlighted with the creation of a concentrated volcano plot.



**Fig. 1** Transcriptome analysis pipeline employed in this study.

### *Reference Transcriptome Creation*

A reference transcriptome was created using the assembler Bridger with a multi *k-mer* approach. The paired end raw data files from each of the eight killifish were individually interlaced to form the files used for assembly and mapping. Bridger was used to assemble six different individual transcriptomes for each killifish using the following six *k-mer* lengths: 21, 25, 27, 29, 31, and 33. Each transcriptome derived by a single *k-mer* was concatenated into a representative transcriptome for that particular fish. Each of the eight killifish transcriptomes were concatenated into one “raw” Reference Transcriptome where CD-HIT-EST and USEARCH were used to cluster together similarly constructed contigs with a score of 94% in order to remove the redundancy.

### *Differential Gene Expression Analysis*

A modified Trinity DGE pipeline was used for expression profiles. Individualized raw reads from each transcriptome were aligned to the reference transcriptome using STAR (Dobin et al., 2012). The estimated transcript abundance for each transcriptome was performed by eXpress (<http://bio.math.berkeley.edu/eXpress/overview.html>). Trinity’s R scripts for edgeR were used to determine DGE using the blind option because of the lack of replicates. The EXP group was compared to the CON group and the NSA group was compared to the CON group.

### *Trinotate Database Creation*

The Trinotate (<https://trinotate.github.io/>) pipeline was used to annotate the Reference Transcriptome. Local BLAST databases were created containing SwissProt and UniProt90 curated entries. BLASTX and BLASTP were run on the Reference Transcriptome against both databases and the output was populated into a sqlite boilerplate database of annotated genes. Additional searches using PFAM, GO, SignalP, tmHMM, and HMMER were conducted and pulled into the boilerplate database. DGE results were populated into TrinotateWeb along with the sqlite boilerplate database, and this information was used to create Heat Maps, Volcano Plots, and MA plots.

### *GO Category Groups*

GO categories related to crude oil exposure and endocrine disruption were created using QuickGO (<https://www.ebi.ac.uk/QuickGO/>) (Garcia et al., 2012 and Yednock et al., 2015). Categories included Estrogen, Ovary, Ovulation, Reproduction, Sex, Spermatogenesis, Testis, Steroidogenesis, Xenobio, Hypoxia, and Heat Shock. Any killifish sequence that was annotated with a GO term related to one of the above categories was organized into that category for further bioinformatics analysis. This was performed by using a custom script to pull GO IDs matching the terms from the Trinotate database into an organized tab delimited text file.

### *Cytoscape*

Cytoscape v.3.4 with the ClueGo v 1.7 and the Golorize v.1.0.0beta1 plug-ins in combination was utilized to visualize complex networks and integrate activities of genes

related to testis-derived reproductive disruption at the functional transcriptomic level (Shannon et al. 2003, Bindea et al., 2009, and Garcia et al., 2007). Cytoscape is an open source bioinformatics program for visualizing molecular interaction networks and biological pathways. The protein–protein interactions visualized in Cytoscape are largely derived from the literature and suggest possible interactions (Shannon et al. 2003). The program displays shapes (nodes) that represent proteins and lines (edges) that represent direct interactions between the proteins. ClueGo visualizes non-redundant biological terms for large clusters of genes in functionally grouped networks (Bindea et al., 2009). Golorize highlights the nodes that belong to the same class using color-coding and then constructs a network using a class-directed layout algorithm (Garcia et al., 2007). Cytoscape, ClueGo and Golorize were all downloaded from [www.cytoscape.org](http://www.cytoscape.org).

#### *Cluster and Java TreeView*

Cluster v3.0 (<http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm>) and Java TreeView v1.1.5 (<http://jtreeview.sourceforge.net/>) were used to visually display the expression levels of all the genes within the Cytoscape network of interest.

#### *Cloning and Sanger Sequencing*

To experimentally verify the assembly results along with the expression analysis, primers were designed based on the assembled transcriptome and PCR was performed. From the isolated RNA, a cDNA library was constructed using oligo-dT primers (Applied Biosystems, Foster City, CA). The cDNA was used as a template to amplify the

genes of interest. The PCR amplification was performed at 94 °C for 30 sec, 60 °C for 30 sec, and 72 °C for 30 sec using the designed primer pairs for each gene. These primers were designed using Primer3Plus (<http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi>) and are shown in Table 3. The presence of a unique PCR product of appropriate size was verified by agarose gel electrophoresis. Each band was extracted from the 2% agarose gel and purified using QIAEX II Gel extraction Kit (Qiagen). The TOPO TA Cloning kit (Invitrogen) was used to clone each gene. The purified PCR product was ligated into TOPO® vector and transformed into Transform One Shot® TOP10F', chemically competent *E.coli* cells. The cells were spread on LB-plate containing 100 µg/mL Ampicillin (Invitrogen), 200 mg/mL IPTG (isopropyl-beta-D-thiogalactopyranoside) (Invitrogen), and 20 mg/mL X-gal (Invitrogen). Single, isolated white colonies were identified and labeled on the Ampicillin-LB plates. Half of the white colony was continuously streaked within a quadrant of a 100 µg/mL Ampicillin-LB patch plate, and the other half of the white colony was dissolved into 50 uL of UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen) for colony PCR. The insert of interest was amplified with the M13R vector primer and the gene-specific forward primer. Each individual PCR product was separated on a 2% agarose gel. The individual bands were excised and gel purified using the gel extraction kit. Each purified plasmid PCR product of interest was confirmed using commercially available Sanger sequencing (Genscript, Piscataway, NJ, USA). Each cloned sequence was aligned to sequences in the nr database of GenBank using BLASTX 2.6.0.



**Table 3:** qPCR primer sequences and expected product size.

<b>Gene</b>	<b>Forward Primer (5'-3')</b>	<b>Reverse Primer (5'-3')</b>	<b>Product Size</b>
<b>GSTA4</b>	AGACAAAAGCGATCCTGCAT	AGGTACACGGGTCCAGTCAG	232 bp
<b>PTEN</b>	TCTTTGTGAGCGTCAGGATG	GGCGACATCAAAGTGGAGTT	250 bp
<b>EGLN2</b>	CTGTGATGGCATAACCGAGTG	CGTGCGTCATGTTGATAACC	233 bp
<b>DNAJB1</b>	GGGGGTATGGAAGAGGACAT	CCTCCTCTAGCGACACCTTG	202 bp
<b>ERCC2</b>	CATCAGAGGCAAGACGGACT	AGCTGGTCCTCCTGTCTGAA	201 bp
<b>PDK1</b>	TGGGAAGGTCAAGGTGAATC	GGCGTTGAATTCCTCCAGTA	163 bp
<b>PPARD</b>	TTCCAGAAGTGCCTTTCGTT	GGTAGGCCGTGTTCACTTGT	172 bp
<b>MMP14</b>	GGTCACTTTTGGAGGGGATT	GCTCGGAAGAAAAACGTCTG	224 bp
<b>AK7</b>	TGGGCTGATGTTCGATTACA	CAGTACGGAATGGGAGAGGA	163 bp
<b>TDRD7b</b>	GTGGTGCTGATGGGAGAAAT	AGCTCGTGTTTGCCGTAGTC	182 bp

### *Quantitative Polymerase Chain Reaction (qPCR)*

The qPCR reactions were carried out using the StepOnePlus System 96-well PCR instrument (Applied Biosystems). The primer sets were verified by TA cloning followed by Sanger sequencing. Of the eleven cloned genes, six were analyzed by qPCR including GSTA4, EGLN2, PDK1, MMP14, AK7, and TDRD7b. The comparative  $C_T$  method was employed to calculate the relative gene expression, and  $\beta$ -actin was used as the housekeeping gene. Each gene was analyzed in quadruplets. Single product formation was verified by performing a melt curve after each qPCR run along with verifying by agarose gel electrophoresis.

### *Statistical analyses*

Differences between the lengths and weights of the sacrificed killifish were investigated using One Way ANOVA, with Tukey posthoc test, and Independent samples T-Test, equal variance assumed, 2-tailed significance,  $p \leq 0.05$ . ANOVA was used to determine statistical differences in weight and length of CON and EXP fish between treatment days, 0-7. T-Test was used to determine statistical differences between the CON vs EXP group at Day 1, Day 3, and Day 7. It was also used to determine statistical differences between NSA vs Day 7 EXP and NSA vs Day 7 CON. The Pearson Correlation, 2-tailed significance test was performed to determine whether the differences in weight of fish had an impact on the expressions of CYP19a, VTG, and CYP1A.

## Results

### *2.1 Fish morphometric characteristics*

Each fish was euthanized with an overdose of MS-222 (tricaine methanesulphonate) and measurements of body weight and length were recorded as seen in Table 4. There were no statistical differences for weight or length between CON (control group) fish (n=10-11) or EXP (exposure group) fish (n=10-14) used at Day 0 through Day 7 as determined by One Way ANOVA analysis, with Tukey posthoc test,  $p < 0.05$ . An Independent samples T-Test, equal variance assumed, 2-tailed significance,  $p \leq 0.05$  showed that there was a statistical difference between the weight of the Day 1 CON (n=10) and EXP (n=10) group,  $p=0.001$ , no difference for Day 3 CON (n=10) and EXP (n=10) group,  $p>0.05$ , and a trend for the Day 7 CON (n=11) and EXP (n=14) group,  $p=0.071$ . The same T-Test was performed for the lengths of the sample groups and it was determined that there was statistical differences between Day 1 CON (n=10) and EXP (n=10),  $p=0.001$ , Day 3 CON (n=10) and EXP (n=10),  $p=0.044$ , and Day 7 CON (n=11) and EXP (n=14),  $p=0.033$ . The same T-Test was performed to determine if there were any significant differences in weight and length between the NSA (non-sexually active group) (n=8) group compared to the Day 7 CON (n=11) and the Day 7 EXP (n=14) groups. Statistical difference was observed between the NSA and CON groups for both weight ( $p=0.001$ ) and length ( $p=0.022$ ). However, there were no significant differences between the NSA group and EXP group for both weight and length,  $p > 0.05$ . Overall, the EXP and NSA groups were significantly different than the CON group. However, correlation graphs and statistical data presented below show that the relative gene expression derived from the semi-

quantitative PCR data did not correlate with the fish weight for CYP19a, VTG, and CYP1a.

**Table 4:** Average morphometric measurements consisting of weight and length for male and female CON, EXP and NSA group. SD (95 %) is in parenthesis.

Group	Day 0 (mm)	Day 0 (g)	Day 1 (mm)	Day 1 (g)	Day 3 (mm)	Day 3 (g)	Day 7 (mm)	Day 7 (g)
CON	68 (9.0 )	5.00 (2.1 )	67.3 (4.4)*	3.94 (0.9)*	69.1 (2.8)*	4.74 (0.8)	64.9 (3.9)*	3.48 (0.7)* $\Delta$
EXP			75.6 (4.4)*	5.75 (1.0)*	74.9 (7.9)*	5.6 (1.9)	72.2 (10.1)*	5.05 (2.6)*
NSA							70.1 (5.0)	5.4 (1.3) $\Delta$

\* T-Test significant difference between CON vs EXP groups,  $p \leq 0.05$

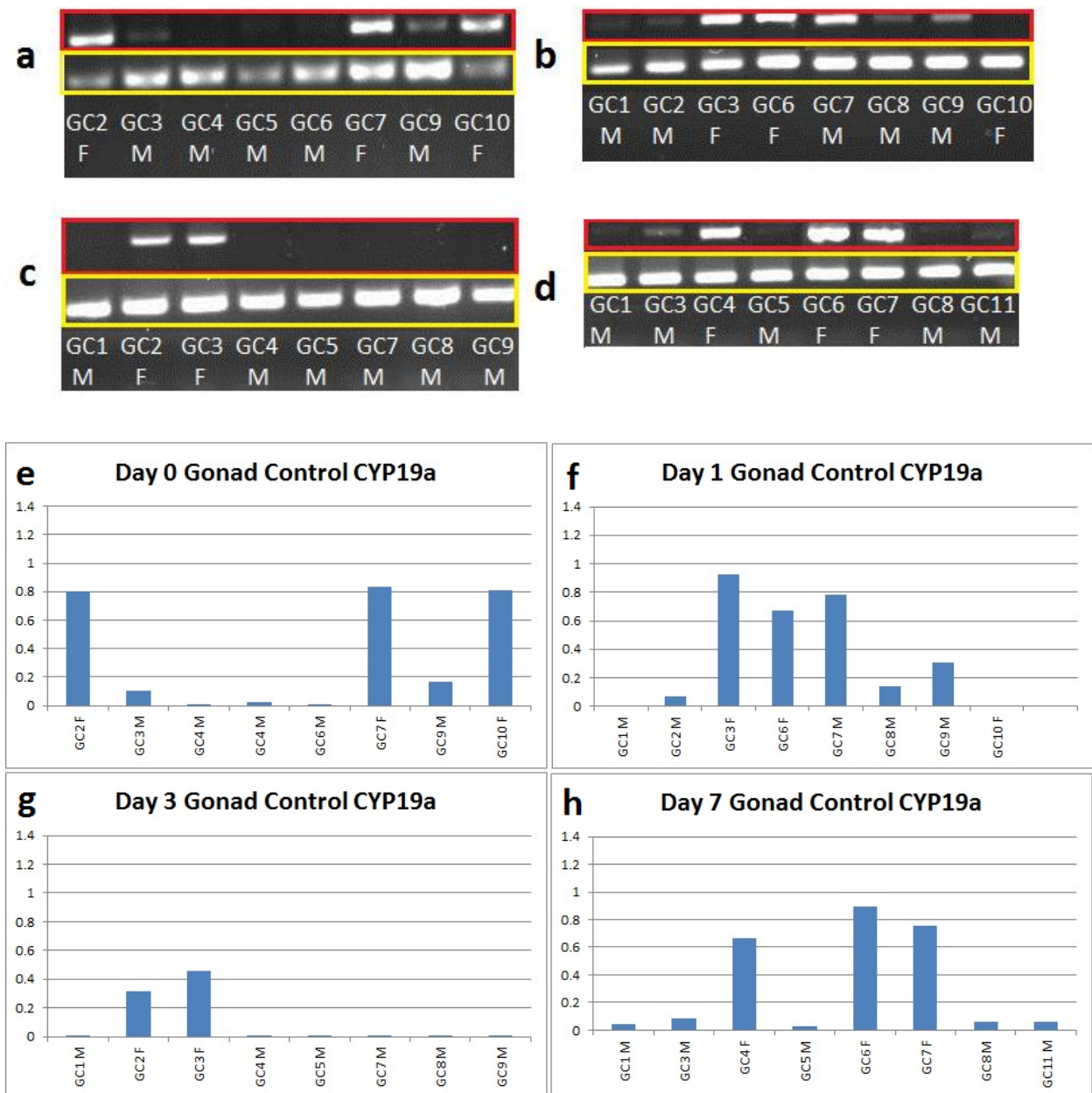
$\Delta$  T-Test significant difference between NSA vs CON at Day 7,  $p \leq 0.05$

## *2.2. Effects of crude oil exposure on gene expression*

The responses of CYP19a in gonad and of VTG and CYP1A in liver were used to assess the degree of endocrine disruption. Sexually active fish treated with crude oil were (exposed) compared to those treated with fish oil only (control) at Day 0, 1, 3 and 7. Based on levels and patterns of gene expression along with EEMS data, certain fish were selected for generating transcriptomes.

The spawning season in the summer is the peak time when CYP19a expression in females would be expected to be intense. It should be noted that male fish do normally convert small amounts of androgens into estrogens but at a level less than what is expected from female fish. With that said, changes in CYP19a expression in gonads exposed to estrogenic endocrine disrupting compounds are associated with the feminization of males and decreased 17 beta-estradiol production in females (Cheshenko et al., 2008).

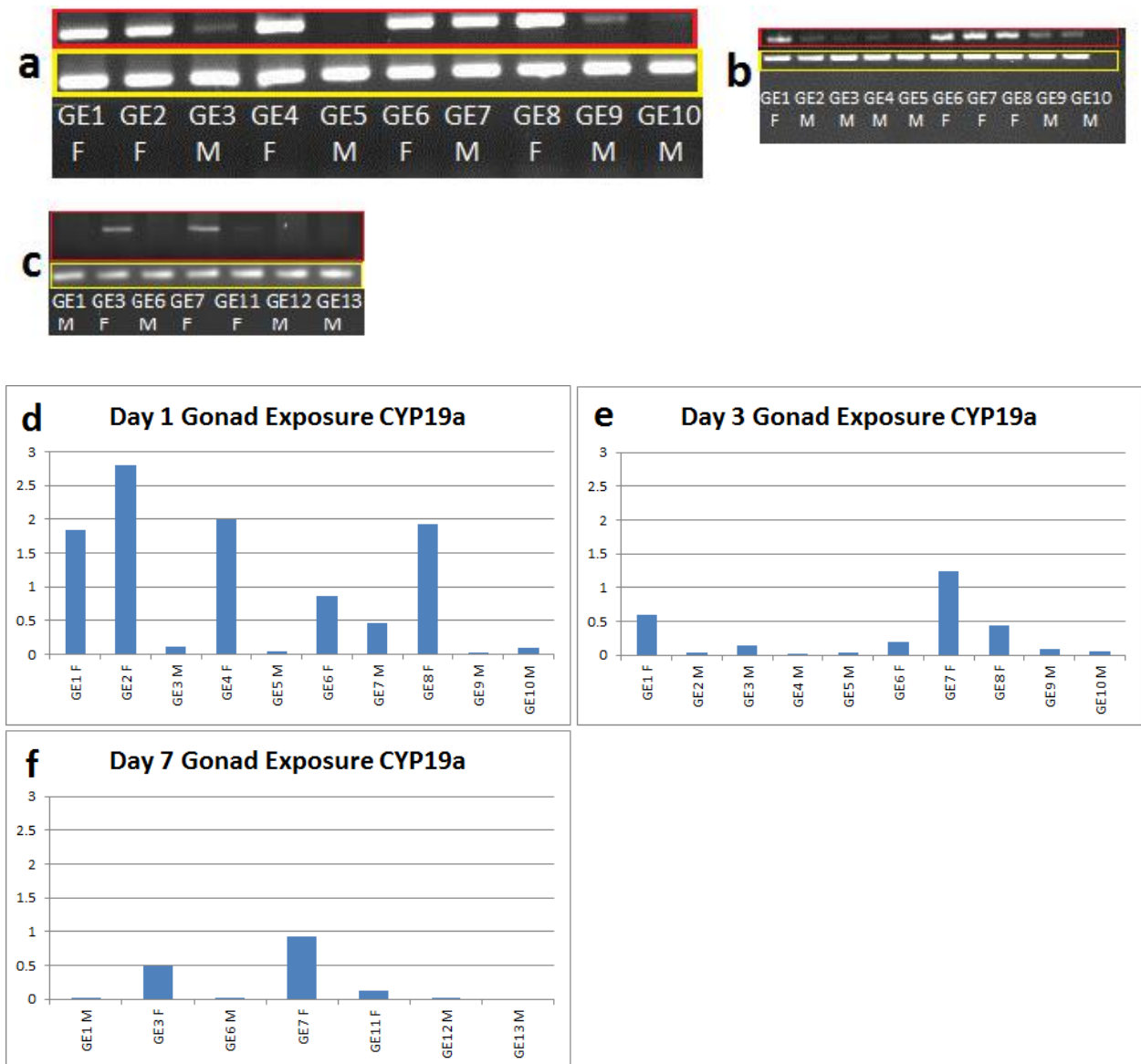
Results for the CYP19a expression in the Day 0, 1, 3, and 7 control gonads can be seen in Fig.2a-d. The top part of the gel represents the CYP19a expression for the control gonads that correspond to their  $\beta$ -Actin expression below. The  $\beta$ -Actin expression was consistent even though the CYP19a expression was not. Fig. 2e-h shows the normalized densitometry analysis for PCR expressions representing the controls at the four time points. As expected, the sexually active females generally were producing CYP19a at levels higher than the males. Based on the overall biomarker and EEMS selection criteria for transcriptome analysis, the females Day7 GC6 and GC7 and males Day7 GC3 and GC11 were chosen partially based on their expected CYP19a expressions.



**Fig. 2** PCR for CYP19a and  $\beta$ -actin in control killifish. Time points were (a) Day 0, (b) Day 1, (c) Day 3, and (d) Day 7 along with the corresponding densitometry analysis for (e) Day 0, (f) Day 1, (g) Day 3, and (h) Day 7 control gonads normalized to  $\beta$ -actin. GC1-GC11 represents gonad control from fish 1, gonad control from fish 2, and so forth. “M” stands for male and “F” stands for female.

The results for the CYP19a expression in the exposed gonads can be seen in Fig. 3a-c. The top part of the gel represents the CYP19a expression for the exposed gonads that correspond to their  $\beta$ -Actin below. The  $\beta$ -Actin expression was consistent and was used as an internal control for the three time points. Fig. 3d-f shows the normalized densitometry analysis for PCR expressions representing the treatment groups. As expected, the sexually active females generally were producing more CYP19a than males at all three time points. This verified that the females were sexually active. The down-regulation seen at Day 3 and Day 7 compared to Day 1 can be interpreted as a sign of crude oil endocrine disruption in females. The Day 7 female, GE11F, was chosen for transcriptome analysis due to its down-regulation of CYP19a compared to control as well as other biomarkers provided below. The males at Day 7 did not show much expression of CYP19a. Ideally, males showing signs of endocrine disruption would have expressed CYP19a at higher levels than Day 0. Instead, male Day 7 GE1 and GE6 were chosen for transcriptome analysis based on other biomarkers and EEMS data (see below).



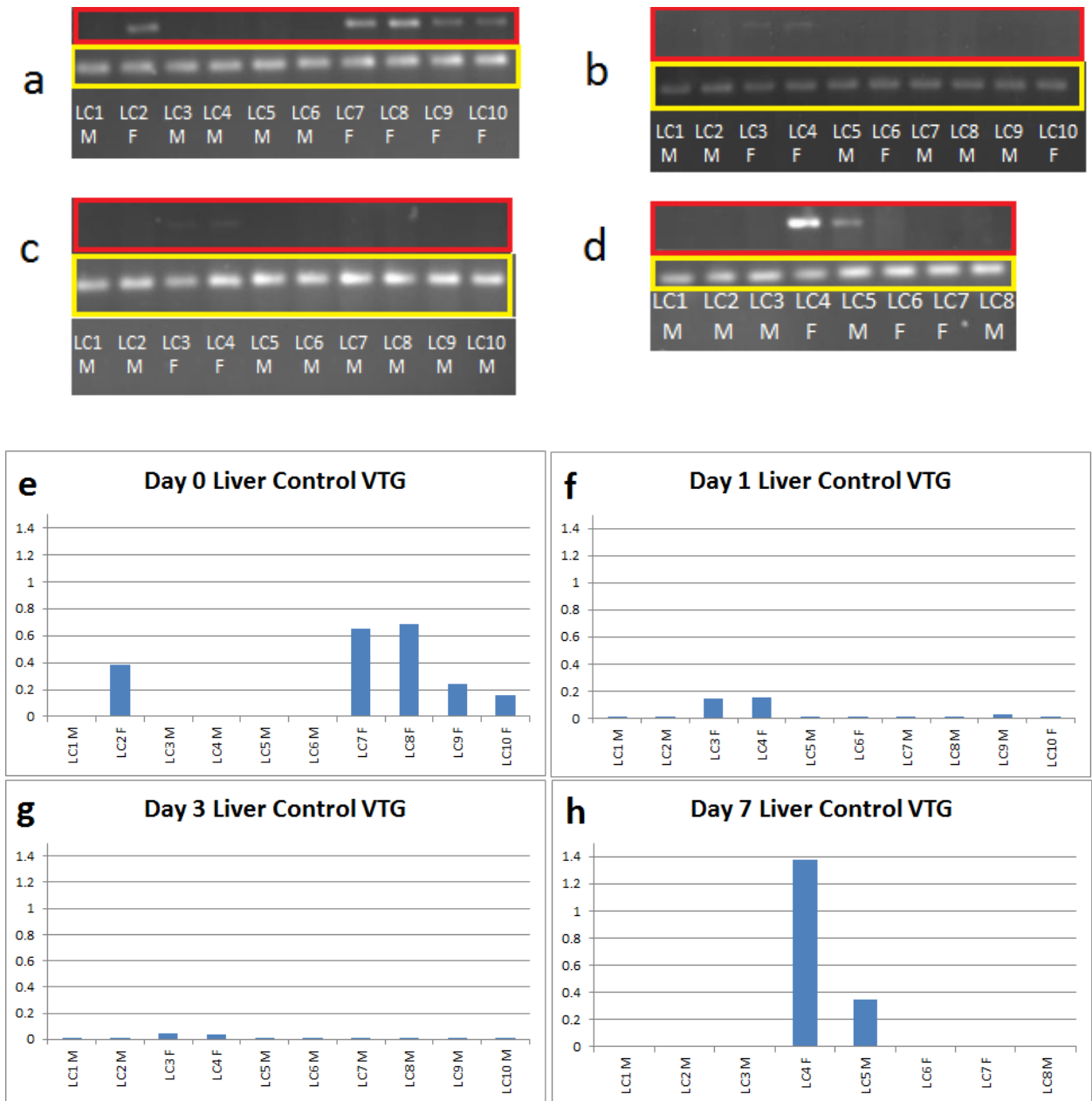


**Fig. 3** PCR for CYP19a and  $\beta$ -actin in exposed killifish. Time points were (a) Day 1, (b) Day 3, and (c) Day 7 along with the corresponding densitometry analysis for (d) Day 1, (e) Day 3, and (f) Day 7 exposed gonads normalized to  $\beta$ -actin. GE1-GE10 represents gonad exposed from fish 1, gonad exposed from fish 2, and so forth. “M” stands for male and “F” stands for female.

VTG is the precursor protein of egg yolk and sexually active females are expected to be producing this gene at high levels during the spawning season. VTG induction, mainly in males and immature females, has been proposed as a biomarker to assess the exposure of estrogenic endocrine disrupting compounds in aquatic environments (Matozzo et al., 2008). Although typically silent, the VTG gene is present in male fish (Matozzo et al., 2008) along with the hepatic estrogen receptor (Pait et al., 2003). If they are exposed to estrogen or estrogenic endocrine disrupting compounds, they can produce vitellogenin which could lead to the feminization of males within a population (Pait et al., 2003).

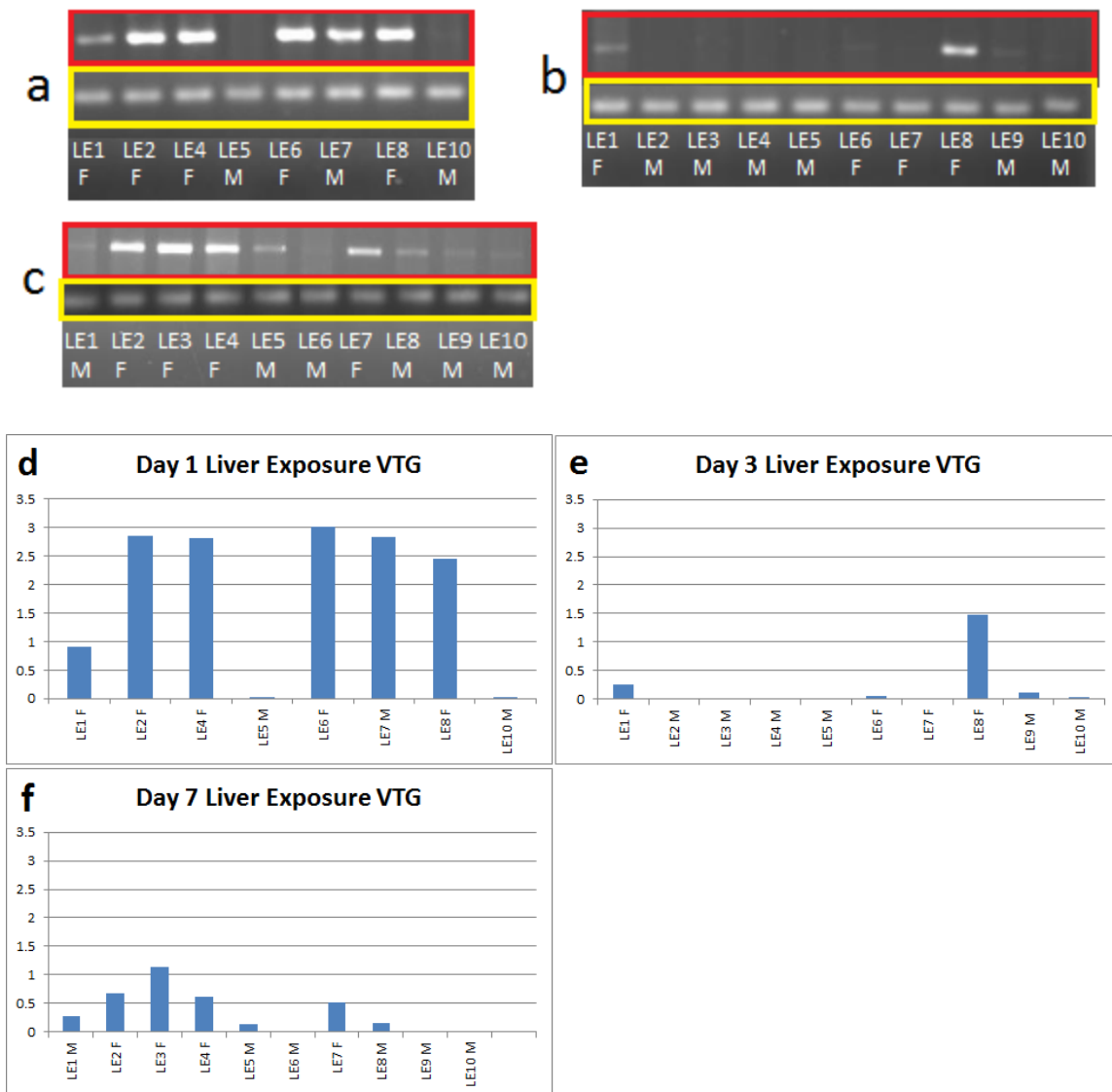
Results for the VTG expression in the control livers can be seen in Fig 4a-d. The top part of the gel represents the VTG expression for the control livers that correspond to their  $\beta$ -Actin expression below. The  $\beta$ -Actin expression was consistent and used as an internal control. To sum up the VTG data in a graphical view, Fig. 4e-h shows the normalized densitometry analysis for PCR expressions representing VTG and  $\beta$ -Actin. With a few exceptions, the majority of the sexually active females are expressing VTG while the males are not. The expression of VTG in female control was reduced over time for unknown reasons. The loss of VTG expression was not consistent with the high level of CYP19a expression in female control gonad at Day 7. Females LC6 and LC7 were slightly producing VTG at Day 7; and for this reason together with other biomarkers, their gonads were chosen for transcriptome analyses. Female LC4 expressed the highest level of VTG at Day 7. Unfortunately, not enough RNA was recovered from the gonad of this fish for transcriptome analysis. Male control was expected to lack VTG expression

and most did. For this reason in addition to other biomarkers, gonads from LC3 were selected for analyses. Gonads from male LC11 were also chosen based on other biomarkers. Unfortunately, the liver sample from this male was lost and VTG expression not measured.



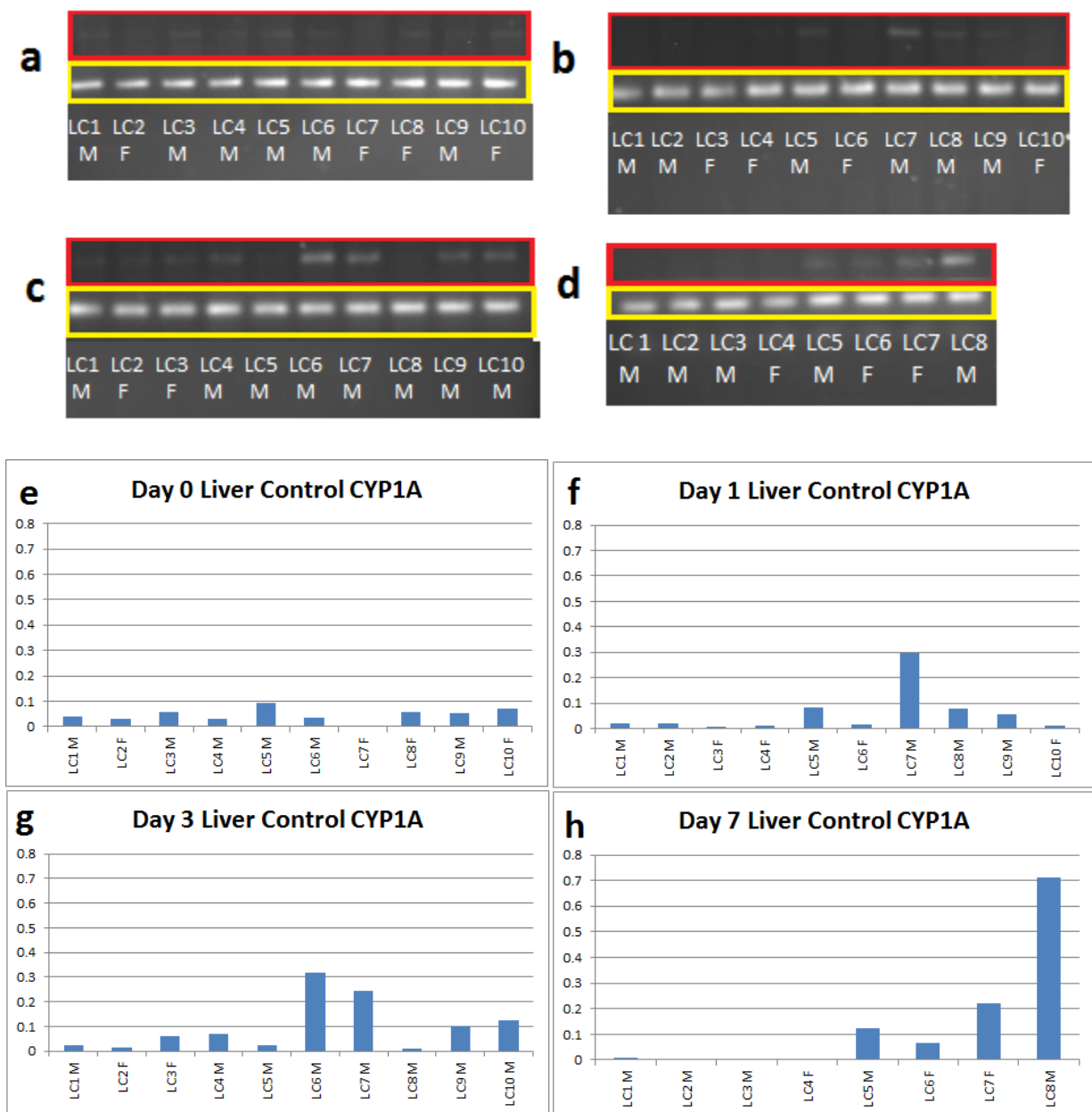
**Fig. 4** PCR for VTG and  $\beta$ -actin in control killifish. Time points were (a) Day 0, (b) Day 1, (c) Day 3, and (d) Day 7 along with the corresponding densitometry analysis for (e) Day 0, (f) Day 1, (g) Day 3, and (h) Day 7 control livers normalized to  $\beta$ -actin. LC1-LC10 represents liver control from fish 1, liver control from fish 2, and so forth. “M” stands for male and “F” stands for female.

The results for the VTG expression in exposed livers can be seen in Fig. 5a-c. The top part of the gel represents the VTG expression for the exposed livers that correspond to their  $\beta$ -Actin below. The  $\beta$ -Actin expression was consistent and was used as an internal control for the three time points. To sum up the VTG data in a graphical view, Fig. 5d-f shows the normalized densitometry analysis for PCR expressions representing the exposure groups at the 3 time points. Overall, the exposure groups unlike the control groups had males expressing VTG at each time point. Five of the six males on day 7 showed signs of feminization by expressing VTG, the most among any time point in this study. Day 7 GE1 and GE6 were chosen for transcriptome analysis partially because they were showing signs of endocrine disruption by expressing VTG. As in control females, the expression of VTG appeared reduced over time. The female Day 7 GE11 was chosen for transcriptome analysis based on its CYP19a expression in gonad and EEMS data even though a liver sample was not available for PCR analysis of VTG.



**Fig. 5** PCR for VTG and  $\beta$ -actin in exposed killifish. Time points were (a) Day 1, (b) Day 3, and (c) Day 7 along with the corresponding densitometry analysis for (d) Day 1, (e) Day 3, and (f) Day 7 exposed livers normalized to  $\beta$ -actin. LE1-LE10 represents liver exposure from fish 1, liver expose from fish 2, and so forth. “M” stands for male and “F” stands for female.

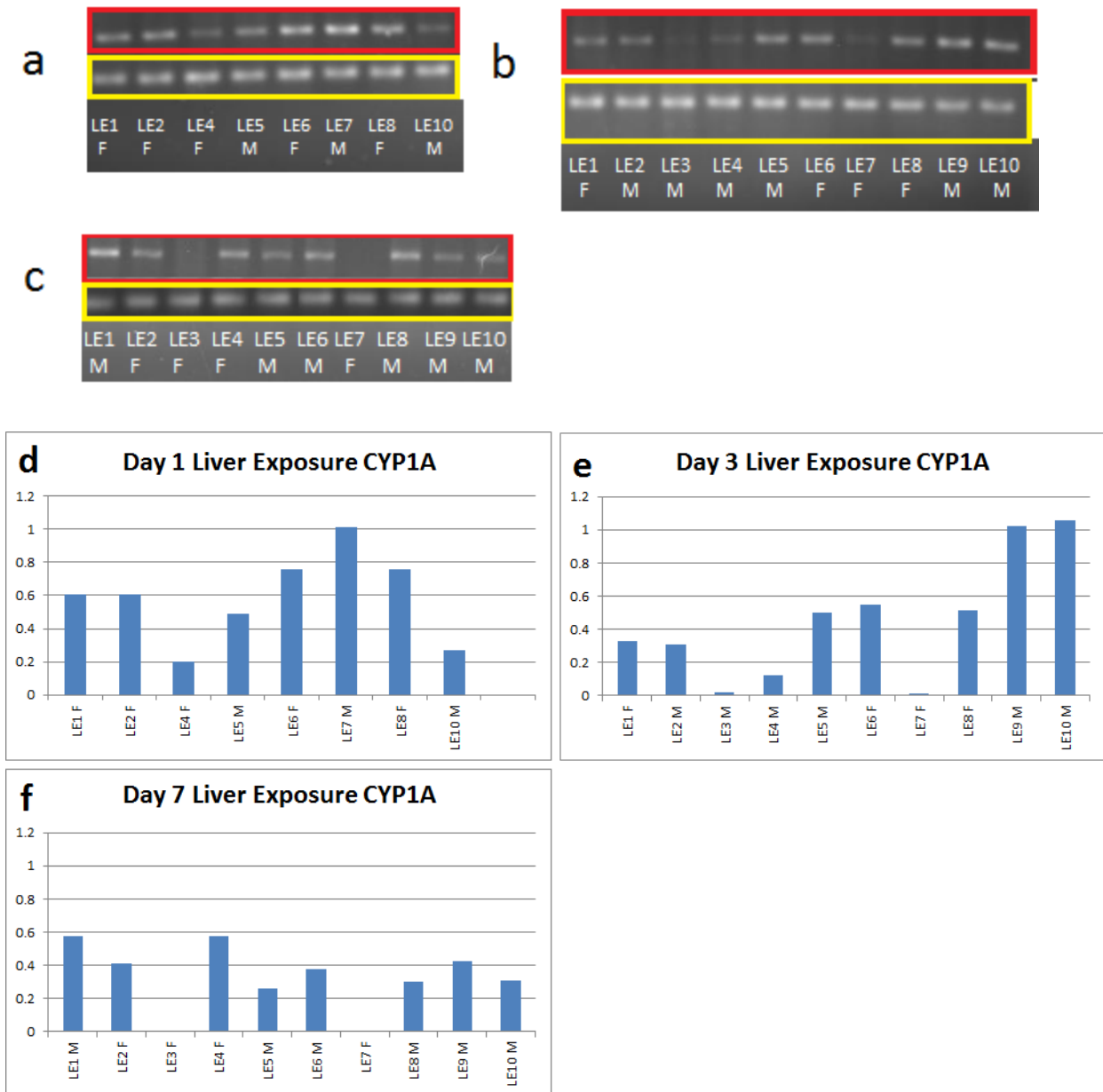
Results for the CYP1A expression in the control livers can be seen in Fig 6a-d. The top part of the gel represents the CYP1A expression for the control livers that correspond to their  $\beta$ -Actin expression below. The  $\beta$ -Actin expression was consistent and used as an internal control. To sum up the CYP1A data in a graphical view, Fig. 6e-h shows the normalized densitometry analysis for PCR expressions representing CYP1A and  $\beta$ -actin. The control fish showed a little variance with the expression of CYP1A. In most species of fish there are relatively low levels of CYP1A constitutively expressed; however, this gene can be highly induced by exposure to AhR ligands such as PAHs (Patel et al., 2006 and Bemanian et al., 2004). The presence of CYP1A in the control fish suggested previous exposure. This is very possible given that these were wild fish. The male Day7 GC3 and GC11 were chosen for transcriptome analysis based on their overall biomarker and EEMS performance. The Day 7 GC3 male showed the expected lack of CYP1A expression for control. No liver was available for Day 7 GC11, so liver CYP1A could not be measured. This male was chosen based on gonad CYP19a and EEMS data. The females Day 7 GC6 and GC7 were chosen for transcriptome analysis based on the overall biomarker and EEMS selection criteria even though they both were slightly expressing CYP1A signifying previous exposures.



**Fig. 6** PCR for CYP1A and  $\beta$ -actin in control killifish. Time points were (a) Day 0, (b) Day 1, (c) Day 3, and (d) Day 7 along with the corresponding densitometry analysis for (e) Day 0, (f) Day 1, (g) Day 3, and (h) Day 7 control livers normalized to  $\beta$ -actin. LC1-LC10 represents liver control from fish 1, liver control from fish 2, and so forth. “M” stands for male and “F” stands for female.



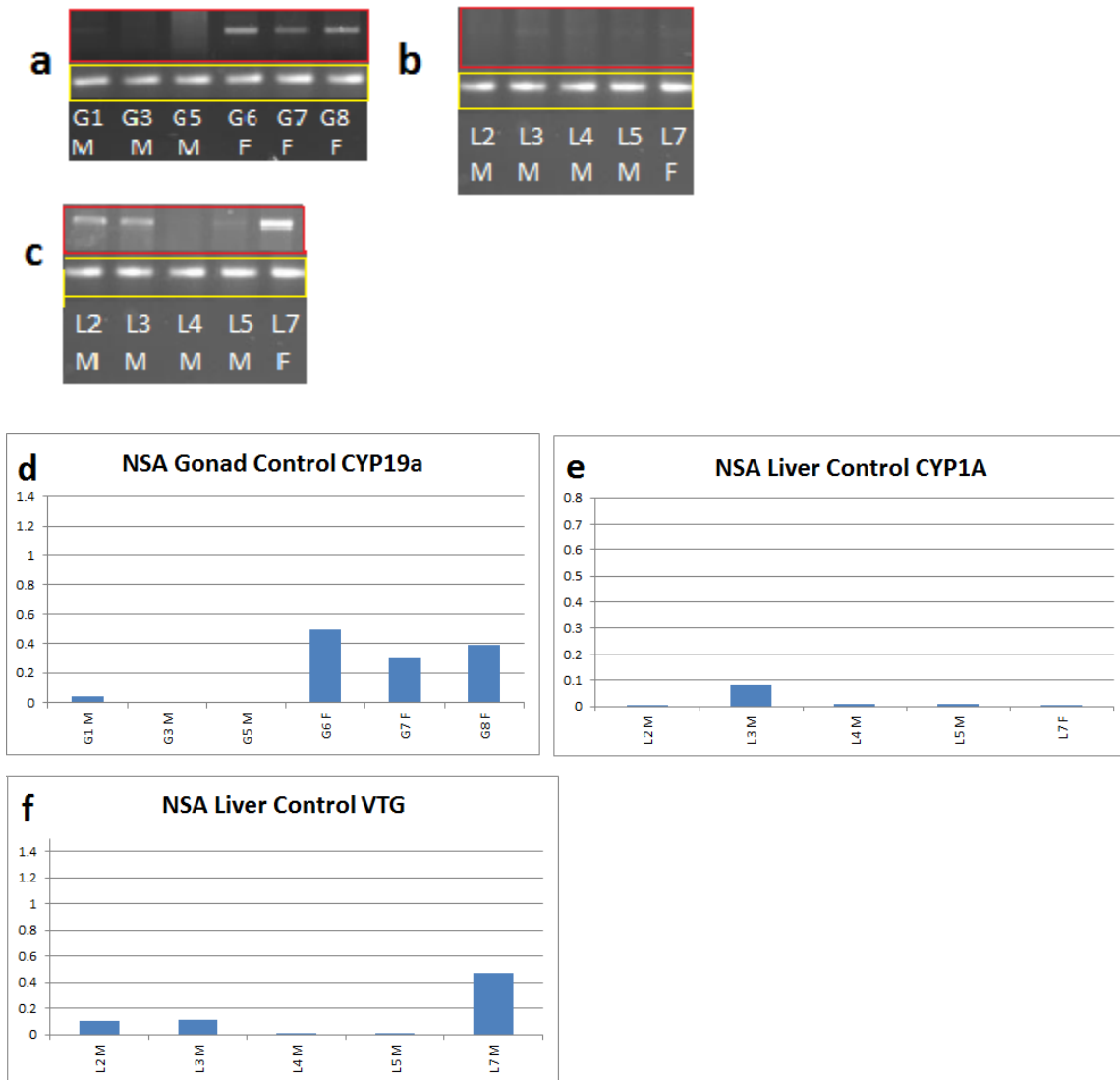
The results for the CYP1A expression in the exposed livers can be seen in Fig. 7a-c. The top part of the gel represents the CYP1A expression for the exposed livers that correspond to their  $\beta$ -Actin below. The  $\beta$ -Actin expression was consistent and was used as an internal control. To sum up the CYP1A data in a graphical view, Fig. 7d-f shows the normalized densitometry analysis for PCR expressions representing the exposure groups at the three time points. The CYP1A expression in the exposed livers was consistent with a few exceptions from Day 1 through Day 7. The levels of expression were higher than those seen at Day 0 (Fig. 6). Consistent expression of CYP1A suggested that the crude oil contained AhR ligands and that the gavage method of dosing was successfully delivering PAHs to the liver. Males Day 7 GE1 and GE6 were chosen for transcriptome analysis partially because they were showing signs of AhR ligands by expressing CYP1A. The female Day 7 GE11 was chosen for transcriptome analysis based on its CYP19a expression and EEMS data even though a liver sample was not available for PCR analysis.



**Fig. 7** PCR for CYP1A and  $\beta$ -actin in exposed killifish. Time points were (a) Day 1, (b) Day 3, and (c) Day 7 along with the corresponding densitometry analysis for (d) Day 1, (e) Day 3, and (f) Day 7 exposed livers normalized to  $\beta$ -actin. LE1-LE10 represents liver exposure from fish 1, liver exposure from fish 2, and so forth. “M” stands for male and “F” stands for female.

### *2.3. Non-sexually Active Killifish Gene Expression.*

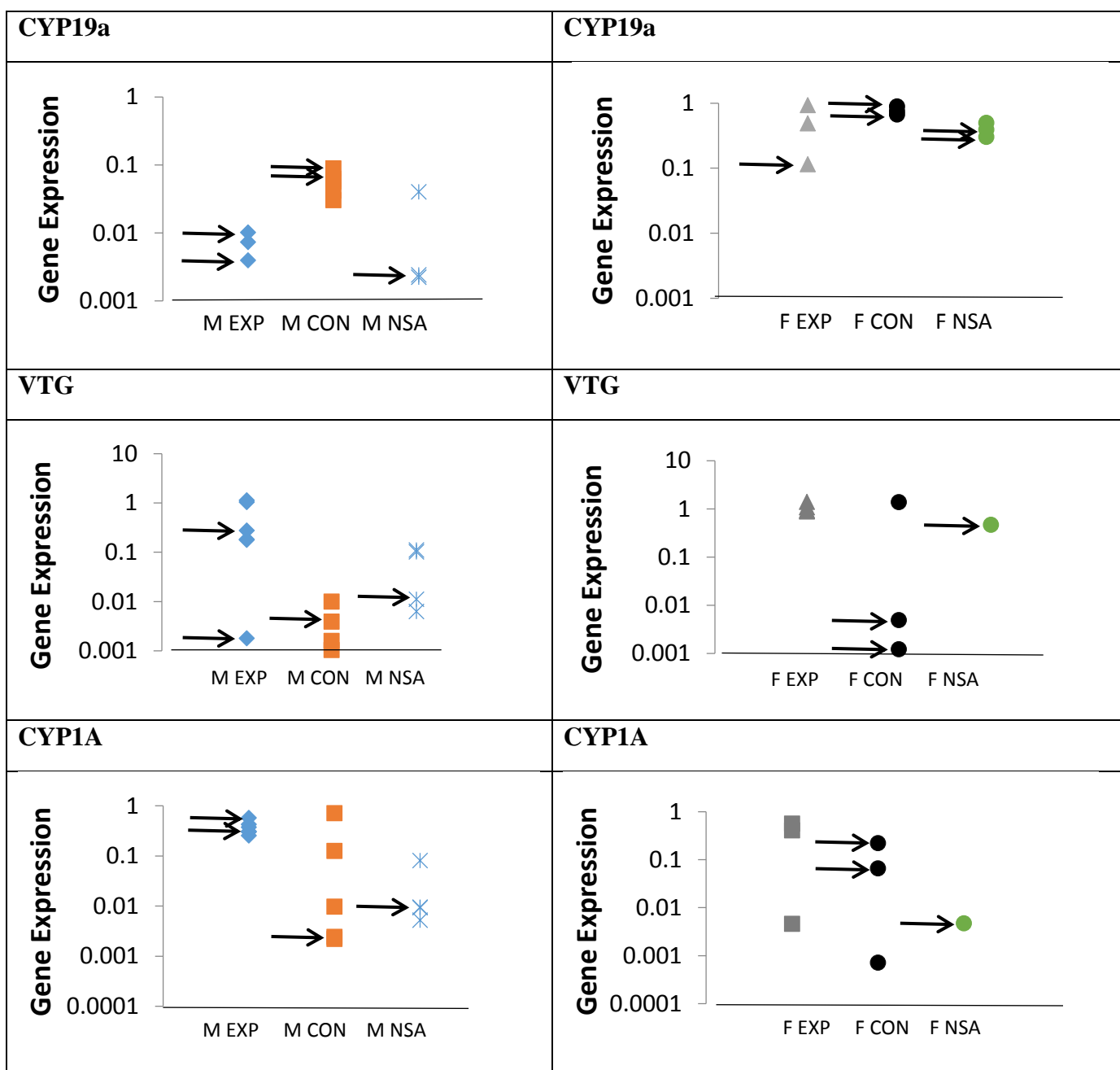
The results for the non-sexually active (NSA) expressions for CYP19a, CYP1A, and VTG can be seen in Fig. 8a-c. The top part of the gel represents the biomarker's expression for the NSA killifish and corresponds to their  $\beta$ -Actin below. The  $\beta$ -Actin expression was consistent and was used as an internal control. To sum up each biomarker's expression data in a graphical view, Fig. 8d-f shows the normalized densitometry analysis for PCR expressions representing the NSA killifish. The NSA Male G5 was chosen for transcriptome analysis because it expectedly expressed CYP19a slightly and had a very little CYP1A along with VTG induction. The NSA Females G7 was selected for transcriptome analysis because it was normally producing CYP19a and only slightly producing CYP1A and VTG. The NSA Female G8 was selected for transcriptome analysis because it was normally producing CYP19a. However, a liver sample was not available to analysis the VTG and CYP1A expression. Non-sexually active females should not be producing high levels of VTG because they are out of spawning season (Nicolas et al., 1999).



**Fig. 8** PCR for (a) CYP19, (b) CYP1A, (c) VTG, and  $\beta$ -actin in non-sexually active (NSA) killifish. Corresponding densitometry analysis is provided for (d) CYP19a, (e) CYP1A, and (f) VTG normalized to  $\beta$ -actin. L2-L7 represents livers and G1-G8 represents gonads from NSA fish 1, from fish NSA 2, and so forth. “M” stands for male and “F” stands for female.

To sum up in a graphical view, Fig. 9 shows the normalized densitometry analysis for the PCR expressions representing the Day 7 CON, Day 7 EXP, and NSA killifish. Those chosen for transcriptome analysis are highlighted by arrows. The results for CYP19a expression in the chosen control and exposed males were unexpected. Feminization should have caused up-regulation of CYP19a expression in exposed males, but both the control and exposed males had relatively low levels compared to females. Crude oil clearly had some effect as CYP19a was approximately 10x lower in exposed compared to control males. The exposed females showed signs of endocrine disruption with the down-regulation of CYP19a expression. The chosen control females were expressing CYP19a at expected levels for sexually active females. Both male and female NSA fish were expressing CYP19a as expected. The results for VTG expression in chosen exposed males indicated that the crude oil caused the expression of VTG to become up-regulated signifying feminization in Day 7 LE1 but not in Day 7 LE6. No liver sample was available for the exposed female chosen for sequencing. As expected, both control females were slightly expressing VTG while the control male was not. The female NSA fish selected for sequencing were expressing VTG normally. CYP1A was increased in crude oil exposed males. No liver sample was available for crude oil in the exposed female chosen for sequencing. There was variability among the female and male control fish possibly due to pre-exposure (they were wild fish) and in most species of fish there are relatively low levels of CYP1A constitutively expressed. However, both male and females chosen for sequencing were slightly expressing CYP1A but at levels lower than exposed fish. Both the male and female NSA fish chosen were expressing CYP1A at

lower levels than the exposed fish. Finally, other factors contributing to selection of sexually active fish samples were the EEMS data as well as the amount of gonad RNA available for transcriptome analysis.



**Fig. 9** Expression levels for genes used to detect endocrine disruption at Day 7 in the crude exposure experiment and NSA. Each symbol represents the expression of an individual fish. M = male, F= female, EXP= gavaged with crude oil, CON= gavaged with Daybrook fish oil, NSA=Non-Sexually Active. Arrows indicate fish selected for transcriptome analyses based on their expression levels for particular genes as well as their EEMS results and amount of RNA.

### *2.3. Correlation of Fish Weights vs Relative mRNA Expression*

All fish weights (g) for Day 7 EXP, Day 7 CON, and NSA were correlated to the semi-quantitative PCR expression for CYP19a, VTG, and CYP1a to determine if the differences in size had an impact on their expression. There was no correlation observed between VTG levels and weight of all fish (n=23); although a trend was observed,  $p=0.08$ . There was no correlation observed between VTG levels and weight for males (NSA, CON, and EXP, n=15),  $p=0.986$ . There was no correlation observed between CYP1A and CYP19a levels and weight of all fish (n=23),  $p = 0.862$  and  $p = 0.892$ , respectively. There was no correlation observed between CYP1A (n=15) and CYP19a (n=12) levels and weight for males (NSA, CON, and EXP),  $p = 0.236$  and  $p = 0.254$ , respectively. Overall, weight did not appear to influence the expression of genes used to select the fish for NGS.



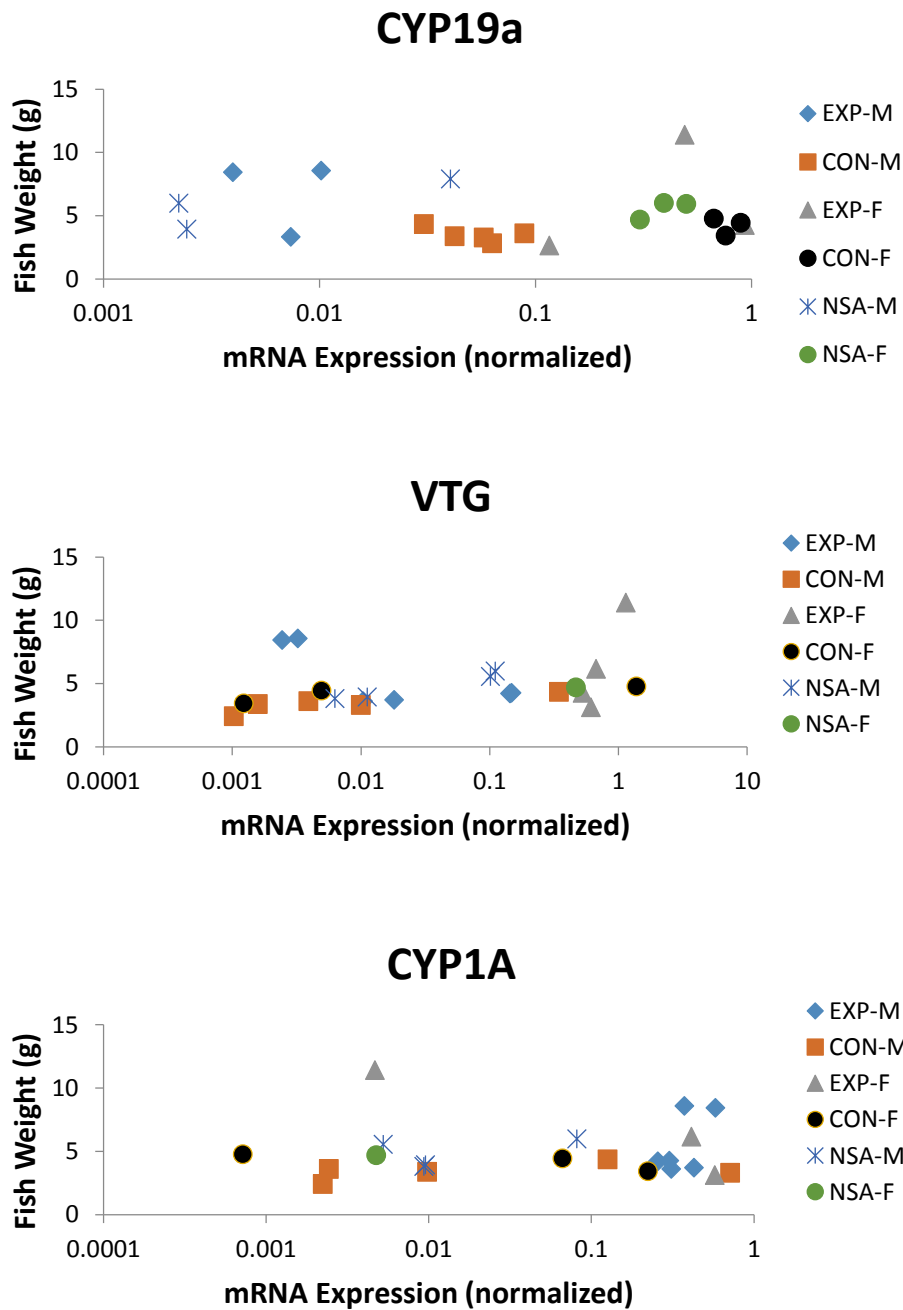


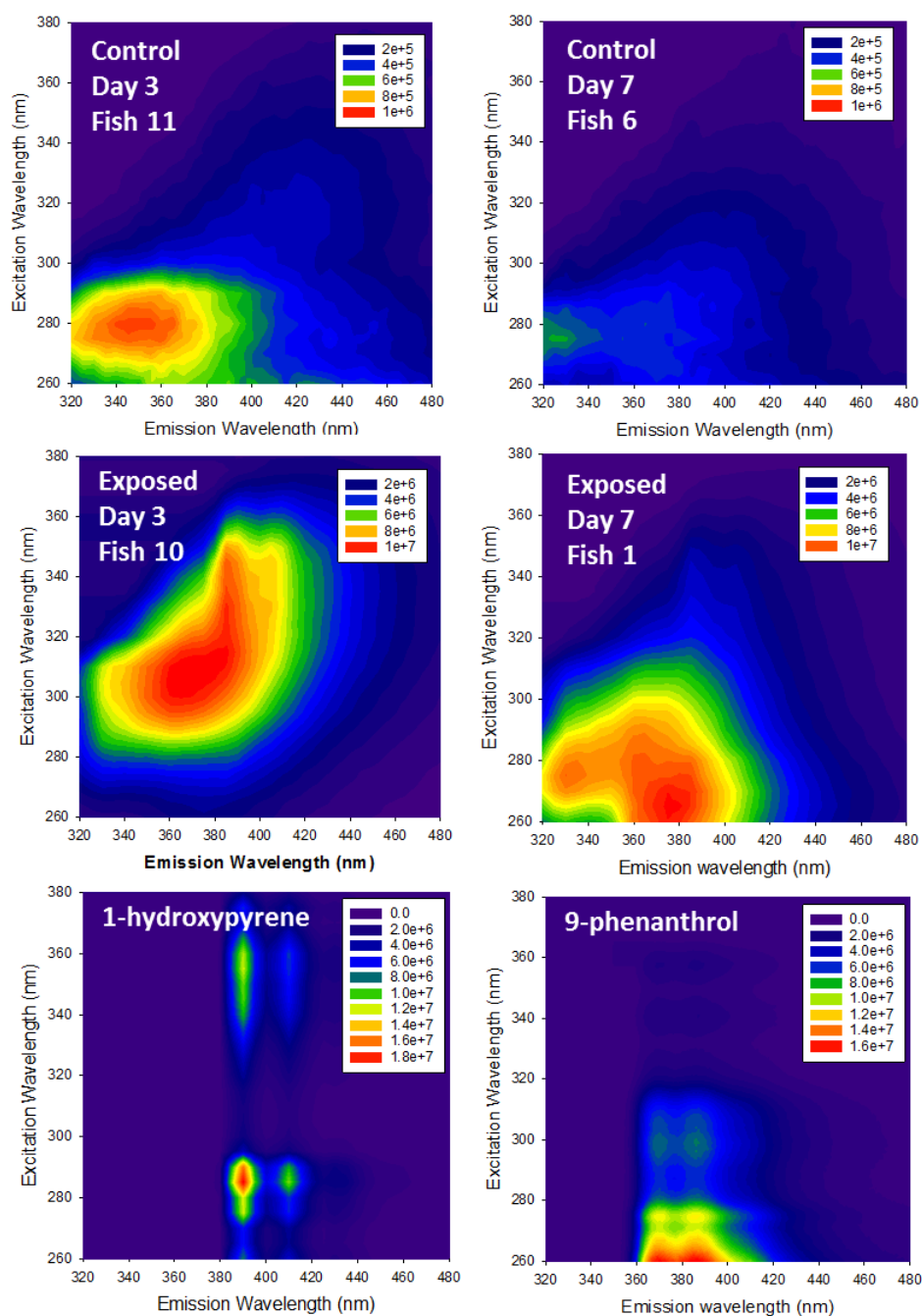
Fig. 10 Day 7 EXP, Day 7 CON, and NSA fish weights (g) normalized to the semi-quantitative PCR expression for CYP19a, VTG, and CYP1a. Weight did not influence the gene expression.

#### *2.4 Excitation-Emission Matrix Spectroscopy (EEMS)*

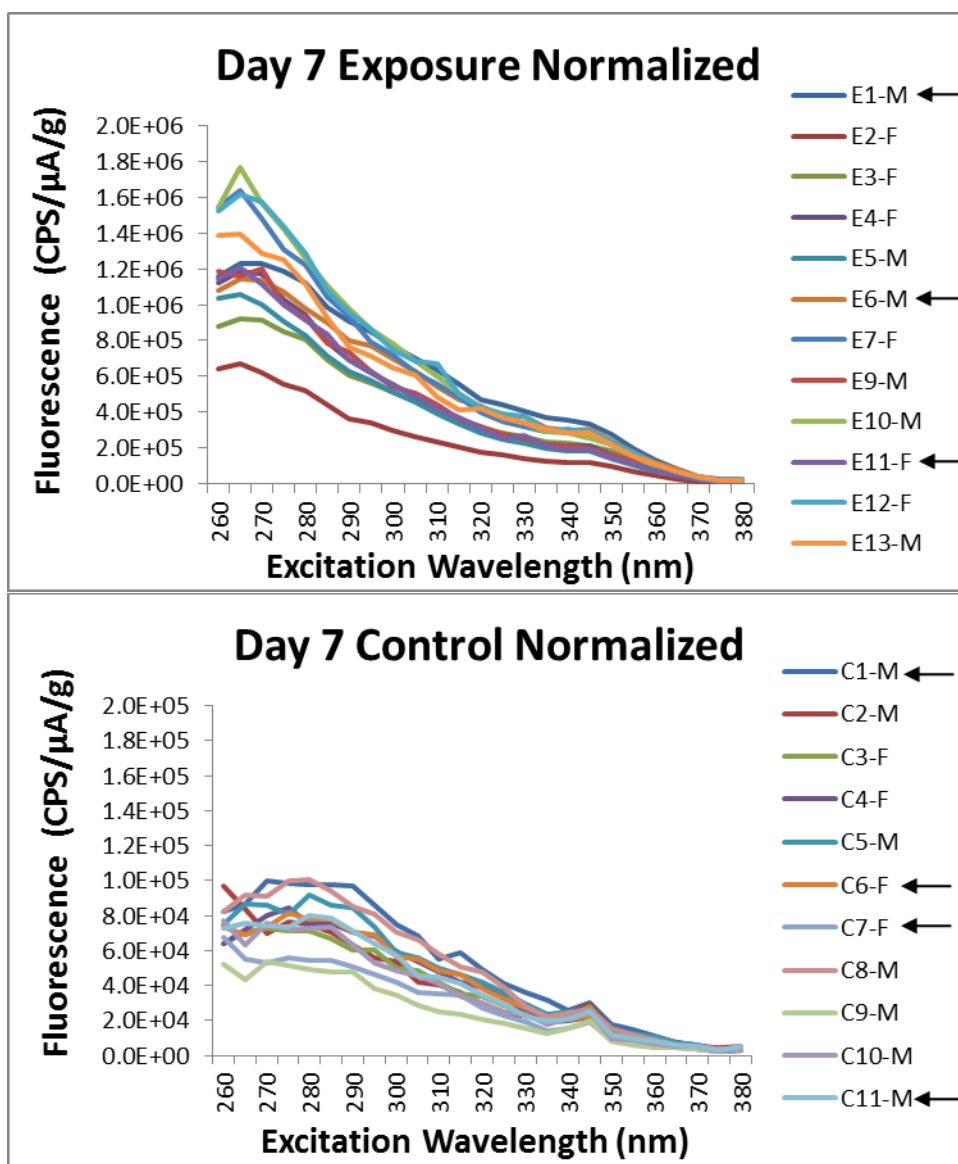
3D EEMS was used to confirm exposure to crude oil. Spectra for one fish from the Day 3 and Day 7 control and exposed groups are provided as well as PAH standards for 9-phenanthrol and 1-hdropyrene (Fig. 11). PAH-like compounds were extracted from fish gall bladders. Spectra for control fish primarily showed a protein signature with no apparent PAH-like compounds. Spectra of exposed, Day 3 fish showed maximum fluorescence at excitation 300 nm and emission 370-380 nm, which was consistent with the minor peak of phenanthrol, and at excitation 340 nm and emission 385, which was consistent with the minor peak of hydroxypyrene. By Day 7, the spectra had changed with the major peak at excitation 260 nm and emission 370-380 nm, which was consistent with the major peak of phenanthrol. The change in spectra over time indicated preferential metabolism of hydroxypyrene-like PAHs over phenanthrol-like PAHs. Similarity of spectra to PAH standards confirmed exposure to PAHs from the crude oil gavage.

2D graphs of EEMS data were generated to concisely visualize all fish (Fig. 12). Emission was set at 385 nm and excitation ranged from 260-380 nm. Note that the scale for the controls is 10x less than that for exposed fish. Overall, the figure illustrated similar and low levels of PAHs in all control fish and similar and higher levels of PAHs in exposed fish. The fluorescent intensity at Ex265/Em385 in the CON ranged from  $2 \times 10^5$  to  $4 \times 10^5$  and in the EXP group ranged from  $2.5 \times 10^6$  to  $1.2 \times 10^7$ . The fish chosen for transcriptome analyses, based on all the factors for selection, are indicated by arrows in Fig. 12. The EXP group representatives were Day 7 GE11F, Day 7 GE1M, and Day 7

GE6M, and the CON group representatives were Day 7 GC6F, Day 7 GC7F, Day 7 GC11M, and Day 7 GC3M.



**Fig. 11** Representative EEMS for gall bladders of crude oil exposed or control killifish collected on Day 3 or 7. PAH standards for 1-hydroxypyrene ( $0.2 \mu\text{g/mL}$ ) and 9-phenanthrol ( $3.0 \mu\text{g/mL}$ ) are included for comparison. Note that the fluorescence scale for control is 1,000x lower than exposed. The fluorescence spectra in controls resembled that of albumin.



**Fig. 12** Levels of fluorescence (CPS/μA) in gall bladders after Day 7 exposure to crude oil or control. Levels were normalized by dividing them by fish weight (g). Emission was at 385 nm and excitation ranged from 260-380 nm. Note that the scale for control is 10x less than exposure. NSA fish were not analyzed by EEMS.

## 2.5 Sequencing and *de novo* transcriptome assembly comparison study

A *de novo* transcriptome assembly comparison study was executed to evaluate the performance of four popular *de novo* assemblers (Bridger, Trinity, Velvet/Oases, and SOAPdenovo-Trans) along with various *k-mer* strategies to determine which process generated the highest quality transcriptome. The raw sequence reads from a non-sexually active male killifish (G5) was utilized throughout this experiment.

To globally profile the testis transcriptome of killifish, we employed Illumina NextSeq 500 technology to sequence the libraries generating 7,197,900 pair-end short reads encoding 1,033,683,143 bases (Table 5). All the raw sequencing reads were deposited into the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) and can be accessed under the accession number SRX1058750. To perform quality control of the raw sequence reads, they were processed to remove Illumina adaptor sequences, low quality reads with a Phred score value less than 20, and contaminating sequences. The processing of the raw sequence reads resulted in 6,349,606 (88.21%) clean reads that were assembled by the following assembly strategies; SOAP LMK, SOAP SMK, Oases LMK, Oases SMK, Bridger, and Trinity (Table 5).

**Table 5.** Statistics of the raw reads after Illumina sequencing and processing

	Killifish
Number of nucleotide bases	1,033,683,143
Number of raw reads	7,197,900
Number of clean reads for assembly	6,349,606
Percent of used reads for assembly	88.21%
Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE ( <a href="http://journals.plos.org/plosone/">http://journals.plos.org/plosone/</a> ).	

### *2.5.1 Statistics of Assembly*

The statistics of each assembly was initially used to evaluate the performance of each assembly strategy. Assemblies with the highest to lowest amount of contigs produced were: Bridger > SOAP LMK > SOAP SMK > Trinity > Oases SMK > Oases LMK (Table 6). Assemblies with the longest to shortest N50 length were: Oases SMK > Oases LMK > Bridger > Trinity > SOAP SMK > SOAP LMK (Table 6). Assemblies with the highest to lowest average contig length were: Oases SMK > Oases LMK > Bridger > Trinity > SOAP SMK > SOAP LMK (Table 6). Assemblies with the highest to lowest amount of contigs over 1kb were: Bridger > Oases SMK > SOAP SMK > Trinity > Oases LMK > SOAP LMK (Table 6). In summary, the Bridger assembly produced the most amount of contigs and contigs over 1kb. The Oases SMK assembly produced the longest N50 length and had the highest average contig length. The SOAP LMK assembly performed the worst for the N50 length, most amount of contigs over 1kb, and the average contig length. The Oases LMK assembly produced the least amount of contigs.



**Table 6.** Statistics of the Assemblies

	SOAP LMK	SOAP SMK	Oases LMK	Oases SMK	Bridger	Trinity
Contig Number	198,085	187,104	99,567	135,312	303,906	180,658
N50 Length	917	1,042	1,676	1,743	1,668	1,189
Minimum						
Contig Length	200	200	200	200	201	224
Largest Contig						
Length	11,658	11,658	16,019	21,489	15,151	11,023
Average contig						
length	622	672	1,021	1,035	879	711
Contigs Over 1k	33,326	36,136	33,847	45,992	81,769	35,527
RMBT	83.78%	80.77%	85.28%	84.18%	87.71%	82.96%

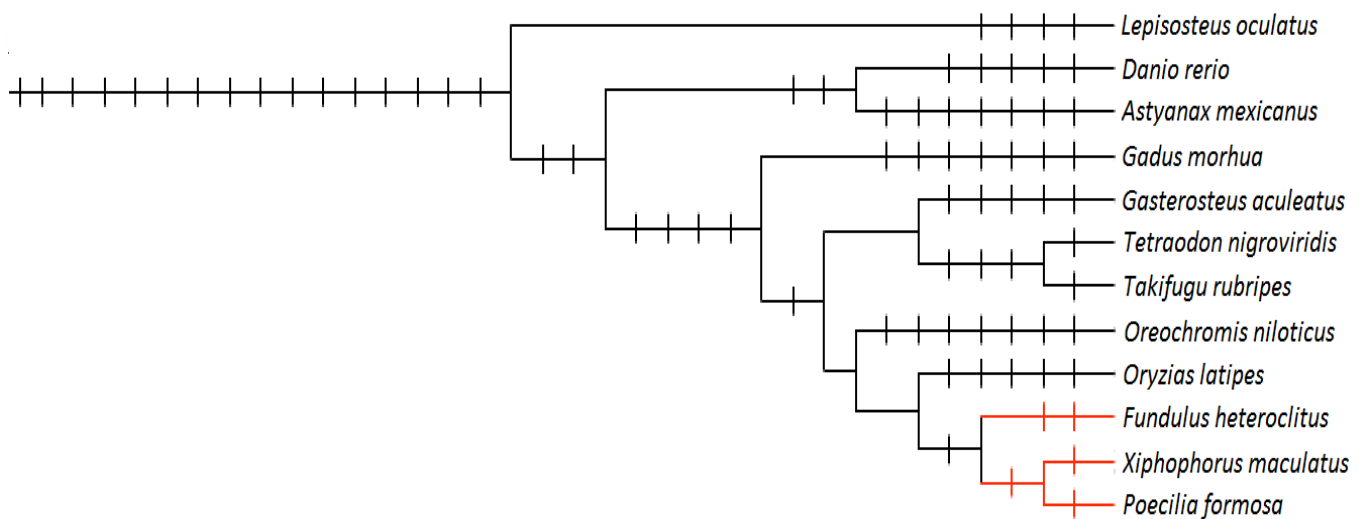
Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE  
(<http://journals.plos.org/plosone/>).

### 2.5.2 RMBT Analysis

A common method to evaluate the accuracy of a *de novo* assembly without a reference genome is to determine the percentage of reads that can be mapped back to transcripts (RMBT) constructed by the assembler. Based on this metric, the Bridger assembly had the highest RMBT percentage (87.71%) and the SOAP SMK assembly had the lowest (80.77%). Results for Trinity, SOAP LMK, Oases SMK, and Oases LMK were similar with RMBT percentages ranging from 82.96% to 85.28% (Table 6).

### 2.5.3 Phylogenetic Tree Alignments

A strategy to gauge the credibility of a *de novo* assembly is to see how well its annotated sequences compare to those of a related species. Therefore, a phylogenetic tree of killifish and eleven publically available fish genomes was constructed using Ensembl (Zadlock et al., 2017). Based on the results, the southern platyfish and Amazon molly were determined to be the closest related genera to killifish (Fig. 13). Therefore, the contigs produced from each assembly strategy were aligned to the southern platyfish and Amazon molly genomes independently using BLASTX with an E-value of  $<1e-3$ . Results showed that the Oases LMK assembly had the best percentage of alignments to both the southern platyfish (50.14%) and Amazon molly (51.82%) databases (Table 7). The Trinity assembly had the lowest alignment percentage to both the southern platyfish (32.68%) and Amazon molly (34.15%) databases. Results for SOAP LMK, SOAP SMK, Bridger and Trinity were similar with alignment percentages ranging from 32.68 to 37.88% for southern platyfish and 34.15 to 39.35% for Amazon molly (Table 7).



**Fig. 13** Phylogenetic tree analysis. A phylogenetic tree analysis of the 11 publically available fish genomes and killifish testis. As highlighted in red, the results of Ensembl showed that southern platyfish (*X. maculatus*) and Amazon molly (*P. Formosa*) are the closest relatives to killifish (*F. heteroclitus*). Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE (<http://journals.plos.org/plosone/>).

**Table 7.** BLASTX alignments from the six different assemblies against the southern platyfish and Amazon molly databases.

Database	SOAP LMK	SOAP SMK	Oases LMK	Oases SMK	Bridger	Trinity
southern platyfish	34.60%	35.22%	50.14%	47.00%	37.88%	32.68%
Amazon molly	36.46%	37.19%	51.82%	48.71%	39.35%	34.15%

Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE (<http://journals.plos.org/plosone/>).

#### *2.5.4 CEGMA Alignments*

A reference-based approach for assessing the quality of an assembly is to align the contigs to the 248 highly conserved proteins in the CEGMA dataset. All of the CEGMA proteins were present in the killifish testis transcriptome. However, none of the assembly strategies were able to incorporate all of them as seen in Table 8. Full-length CEGMA proteins, defined as having at least 70% of the protein length found in the CEGMA dataset, ranged from 97.58 to 98.79%. Partial CEGMA proteins ranged from 99.60 to 100%. The Oases LMK and Oases SMK assemblies contained the highest percentage of full length (98.79%) and partial length (100%) CEGMA proteins. The Trinity assembly contained the lowest percentage (97.58%) of full length CEGMA proteins as well as the lowest percentage (98.79%) of partial CEGMA proteins (Table 8). Both of the SMK and LMK strategies of SOAP and Oases were unable to assemble the same CEGMA proteins at full or partial length (Table 8).

**Table 8.** BLASTX alignments of the six different assemblies to the CEGMA dataset.

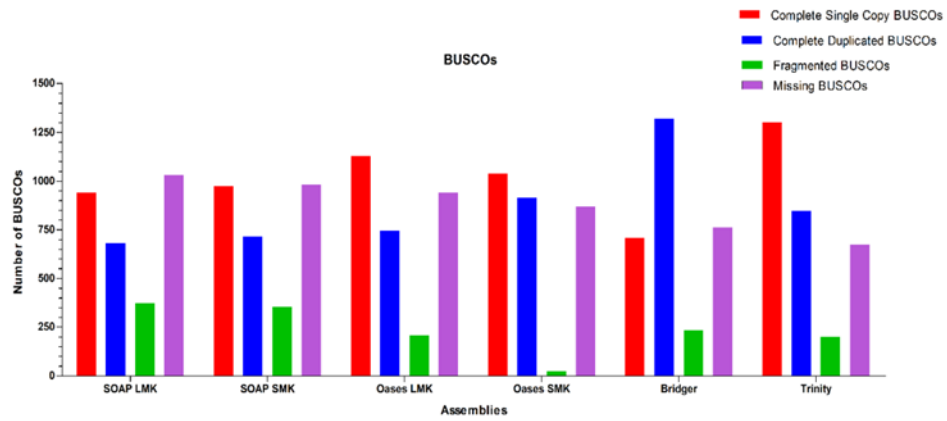
Assembly	CEGS %	CEGs Missing	Partials %	Partial Missing
SOAP LMK	98.39	KOG0261, KOG0209, KOG0462, KOG2311	99.60	KOG2311
SOAP SMK	98.39	KOG0261, KOG0209, KOG0462, KOG2311	99.60	KOG2311
Oases LMK	98.79	KOG0292, KOG2311, KOG4392	100	
Oases SMK	98.79	KOG0292, KOG2311, KOG4392	100	
Bridger	98.39	KOG0261, KOG0292, KOG0969, KOG2311	99.60	KOG0969
Trinity	97.58	KOG0292, KOG0209, KOG0434, KOG0469, KOG0481, KOG2623	98.79	KOG0209, KOG0292, KOG0434

Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE (<http://journals.plos.org/plosone/>).

### 2.5.5 BUSCO Alignments

BUSCO is another reference-based program for assessing quality of *de novo* assemblies. The program determined the percentage of mis-assembled transcripts by trying to align all transcripts to highly conserved proteins within the BUSCO dataset. None of the assembly strategies was able to incorporate all of the 3,023 vertebrate BUSCOs genes as seen in (Fig. 14.) Trinity performed best in terms of having the least amount of missing genes 675 (22.3%), followed by Bridger 762 (25.2%), Oases SMK 869 (28.7%), Oases LMK 940 (31%), SOAP SMK 980 (32.4%), and SOAP LMK 1,030 (34.1%).

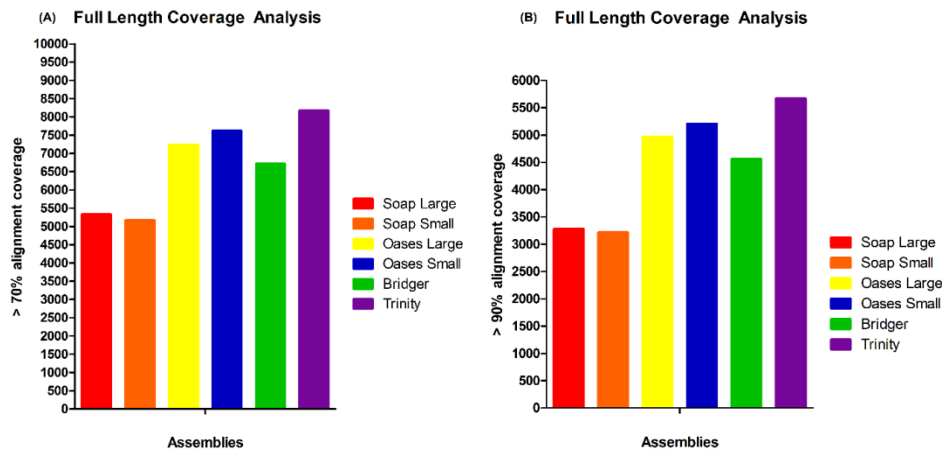




**Fig. 14.** BUSCO Analysis. The Trinity assembly performed the best by having the least amount of missing BUSCOS. Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE (<http://journals.plos.org/plosone/>).

### *2.5.6 Full Length Transcript Analysis*

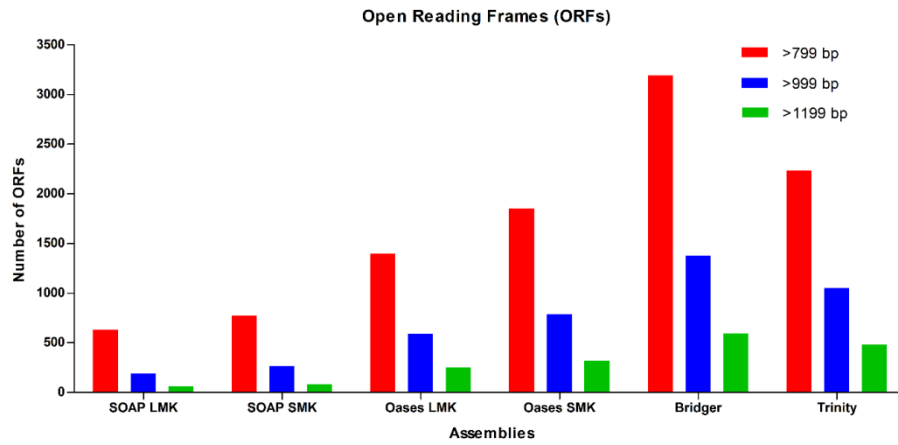
Another metric of assembler performance is quantifying the amount of transcripts that appear to be nearly full length. Based on the alignments with the manually annotated and reviewed SwissProt database, the Trinity assembly performed the best. This assembly produced 8,168 proteins that had greater than 70% alignment coverage and 5,664 proteins that had greater than 90% alignment coverage as seen in (Fig. 15a and Fig. 15b). Other assemblies with the highest to lowest amount of nearly full length proteins were: Oases SMK > Oases LMK > Bridger > SOAP LMK > SOAP SMK. This rank order was the same for 70% and 90% alignment analyses.



**Fig. 15** Full Length Transcript Analysis. a) Trinity had the most proteins with greater than 70 % alignment coverage. b) Trinity had the most proteins with greater than 90% alignment coverage. Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE (<http://journals.plos.org/plosone/>).

### *2.5.7 Open Reading Frames Analysis*

The contigs in this project were sequenced from mRNA; therefore, the best assembly strategy should produce a large number of open reading frames (ORFs). Overall, the assemblies performed in the same rank order for the presence of ORFs in sequence lengths ranging from >799 bps, >999 bps, and >1,199 bps. Assemblies with the highest to lowest amount of ORFs for each sequence length were: Bridger > Trinity > Oases SMK > Oases LMK > SOAP SMK > SOAP LMK. The Bridger assembly had the highest amount of ORFs for all three lengths. This assembly had 3,189 ORFs > 799 bps, 1,377 ORFs > 999 bps, and 593 ORFs > 1,999 bps (Fig. 16). The SOAP LMK assembly had the lowest amount of ORFs for all three lengths. This assembly had 699, 187, and 62 ORFs for sequence lengths ranging from > 799 bps, > 999 bps, and > 1,999 bps, respectively (Fig. 16.)



**Fig. 16.** Open Reading Frames Analysis. The Bridger assembly produced the most amount of open reading frames for sequences with >799 bps (red), >999 bps (blue), and >1,199 bps (green). Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE (<http://journals.plos.org/plosone/>).

### 2.5.8 Detonate RSEM-EVAL Score

A novel metric to evaluate the quality of each assembly is Detonate's RSEM-EVAL score. This score is based on a probabilistic model that only requires the clean reads and the overall assembly for inputs. Assemblies with higher RSEM-EVAL scores are considered better. The assemblies with highest to lowest RSEM-EVAL scores are as follows: Trinity > Bridger > SOAP LMK > Oases LMK > SOAP SMK > Oases SMK (Table 9). The scores for all six assemblies ranged from  $-5,426.0 \times 10^6$  to  $-6,125.0 \times 10^6$  indicating that all assemblies were very similar (Table 9).

**Table 9.** The Detonate's RSEM-EVAL scores suggests that the Trinity assembly performed the best. The higher the number value, the better the assembly.

Assembly	SOAP LMK	SOAP SMK	Oases LMK	Oases SMK	Bridger	Trinity
Score	-5,488.0 $\times 10^6$	-5,715.0 $\times$ $10^6$	-5,602.0 $\times 10^6$	-6,125.0 $\times 10^6$	-5,448.0 $\times 10^6$	-5,426.0 $\times$ $10^6$

Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE (<http://journals.plos.org/plosone/>).

### *2.5.9 Overview of Assembly Strategies*

Based on the eleven evaluation metric categories used to assess the six different assembly strategies, it was determined that no one particular assembly strategy performed the best in all categories tested (Table 10). The assembly strategies that performed within the top three of most metrics was the Bridger assembly (10 of 11), followed by the Oases SMK assembly (9 of 11), and then the Oases LMK assembly (6 of 11). The assembly strategies that occurred within the top three the least were the Trinity assembly (5 of 11), followed by the SOAP LMK assembly (4 of 11) and then the SOAP SMK assembly (3 of 11). Therefore, the Bridger assembler was chosen to use for the selected transcriptomes.



**Table 10.** Summary of the top three performers for each evaluation metric category.

	First	Second	Third
Contig Number	Bridger	SOAP LMK	SOAP
N50 Length	Oases SMK	Oases LMK	SMK
Contigs >1kb	Bridger	Oases SMK	Bridger
RMBT	Bridger	Oases LMK	SOAP
southern platyfish			SMK
DB	Oases LMK	Oases SMK	Oases
Amazon molly			SMK
DB	Oases LMK	Oases SMK	
CEGMA	Oases LMK and Oases SMK	Bridger, SOAP LMK, and SOAP SMK	Bridger
BUSCO	Trinity	Bridger	Oases
Full Length			SMK
Transcripts	Trinity	Oases SMK	Oases
ORFs	Bridger	Trinity	LMK
Detonate	Trinity	Bridger	Oases
			SMK
			SOAP
			LMK

Modified from Rana et al., 2016. Reprinted with permission from PLOS ONE (<http://journals.plos.org/plosone/>).

## *2.6. Next Generation Sequencing of male and female gonadal tissue*

Although this study's primary focus was on male killifish, it was important to identify female-sexually-oriented genes in order to detect endocrine disruption associated with feminization. Therefore, each of the following groups had a male and female representative transcriptome: sexually active exposed (EXP), sexually active control (CON), and non-sexually active (NSA). It should be noted that in some cases two killifish representatives were required to meet the minimum amount of RNA required for sequencing the EXP male and CON male transcriptomes.

To globally profile the gonad transcriptomes of all the killifish groups, we employed Illumina NextSeq 500 technology to sequence the libraries. All the raw sequencing reads were deposited into the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI), and can be accessed under the following accession numbers: EXP male (SRR4384820), EXP female (SRR4384823), CON female 1 (SRR4384824), CON female 2 (SRR4384822), CON male (SRR4384819), NSA male (SRR4384821), and NSA female (SRR4384825).

To improve the accuracy of the assembly, the raw sequence reads were cleaned to remove Illumina adaptor sequences, low quality reads with a Phred score value less than 20, and contaminating sequences. The filtering of the raw sequence reads resulted in 2,494,610 clean reads for EXP male, 1,656,757 clean reads for EXP female, 2,576,992 clean reads for CON male, 1,820,504 clean reads for CON female 1, 1,352,102 clean reads for CON female 2, 2,988,495 clean reads for NSA male, and 1,466,660 clean reads

for NSA female (Table 11). These cleans reads were then assembled using the *de novo* assembler, Bridger.

**Table 11.** Statistics of Reference Derived Transcriptome Assemblies

	<b>EXP Male D7 GE 1&amp;6</b>	<b>EXP Female D7 GE11</b>	<b>CON Male D7 GC 3&amp;11</b>	<b>CON Female 1 D7 GC6</b>	<b>CON Female 2 D7 GC7</b>	<b>NSA Male G5</b>	<b>NSA Female G8</b>
Contig Number	2,494,610	1,656,757	2,576,992	1,820,504	1,352,102	2,988,495	1,466,660
N50 Length	1,096	1,869	1,503	1,971	1,642	1,280	2,119
Minimum Contig Length	201	201	201	201	201	201	201
Largest Contig Length	23,252	16,452	21,154	21,385	15,253	23,074	18,415
Average contig length	670.37	1,004.47	861.43	1,030.44	954.82	710.84	1147.74
Contigs Over 1k	464,551	544,273	715,640	597,152	425,377	580,034	543,034
RMBT to Reference	97.04%	98.47%	96.49%	97.89%	98.4%	96.28%	98.31%

### *2.6.1 De novo transcriptome assembly of the sexually active exposed and non-sexually active male killifish groups*

Based on the results of the *de novo* transcriptome assembly comparison study, the Bridger assembler was utilized to perform the *de novo* assembly for all the killifish groups. As seen in Table 11, the EXP male paired-end sequence reads were assembled into 2,494,610 contigs with an N50 length of 1,096 bp and average contig length of 670.37 bp. The EXP female paired-end sequence reads were assembled into 1,656,757 contigs with an N50 length of 1,869 bp and average contig length of 1004.47 bp. The CON male paired-end sequence reads were assembled into 2,576,992 contigs with an N50 length of 1,503 bp and average contig length of 861.43 bp. The CON female 1 paired-end sequence reads were assembled into 1,820,504 contigs with an N50 length of 1,971 bp and average contig length of 1030.44 bp. The CON female 2 paired-end sequence reads were assembled into 1,352,102 contigs with an N50 length of 1,642 bp and average contig length of 954.82 bp. The NSA male paired-end sequence reads were assembled into 2,988,495 contigs with an N50 length of 1,280 bp and average contig length of 710.84 bp. The NSA female paired-end sequence reads were assembled into 1,466,660 contigs with an N50 length of 2,119 bp and average contig length of 1,147.7 bp. Overall, one interesting pattern observed was that all three male transcriptomes had more contigs compared to the three female transcriptomes.

### *2.6.3 Assembly Assessment: RMBT Analysis*

To verify the quality of the reference killifish transcriptome, the percentage of reads that can be mapped back to transcripts (RMBT) constructed by the Bridger assembler was determined. The RMBT percentages for the seven transcriptomes used to make the reference transcriptome are as follows: EXP Male (97.04%), EXP Female (98.47%), CON Male (96.49%), CON Female 1 (97.89%), CON Female 2 (98.4%), NSA Male (96.28%), and NSA Female (98.31%). These percentages were better than what was observed in the *de novo* transcriptome assembly comparison study where the highest RMBT percentage was performed by Bridger at 87.71% (Table. 6). Overall, the RMBT percentages for the transcriptomes used to make the reference killifish transcriptome ranged from 96.49% to 98.31% signifying that all the transcriptomes were reliably assembled (Table 11).

#### *2.6.4 Assembly and Annotation Assessment: BUSCO Alignment Analysis*

BUSCO alignment analysis was also used to verify the quality of the reference killifish transcriptome. The transcriptome was not able to incorporate all of the 3,023 vertebrate BUSCOs genes as seen in Fig. 17. However, the BUSCO analysis showed that only 563 (18.6%) genes were missing. This was better than what was observed during the *de novo* transcriptome assembly comparison study where Trinity performed the best in terms of having the least amount of missing genes 675 (22.3%) as seen in Fig. 16.

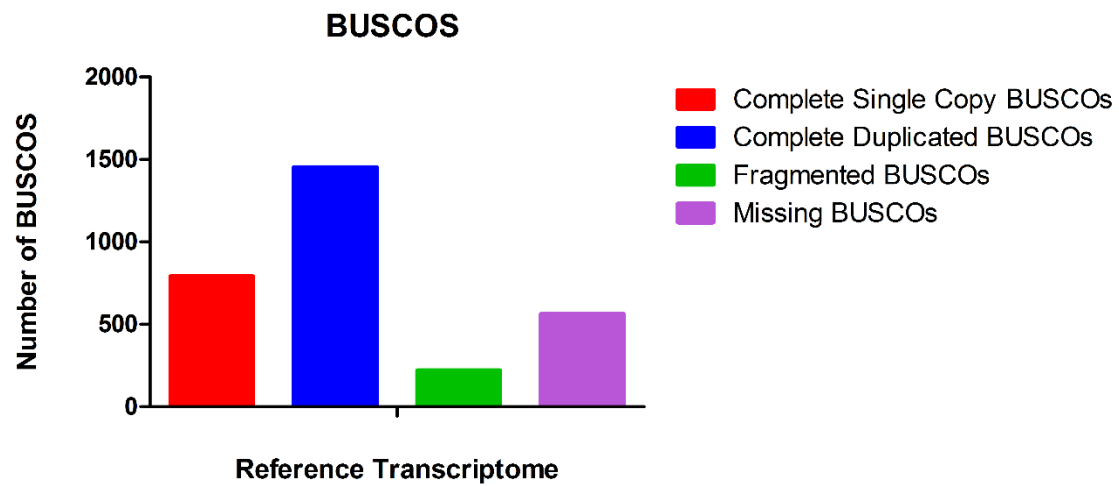


Fig. 17 BUSCO Analysis of the reference killifish transcriptome. The transcriptome had a lower percentage of missing BUSCOs to what was observed in the *de novo* transcriptome assembly comparison study.

## 2.7 Visualization and Analysis of Transcriptome Data

### 2.7.1 Overview of data processing and visualization

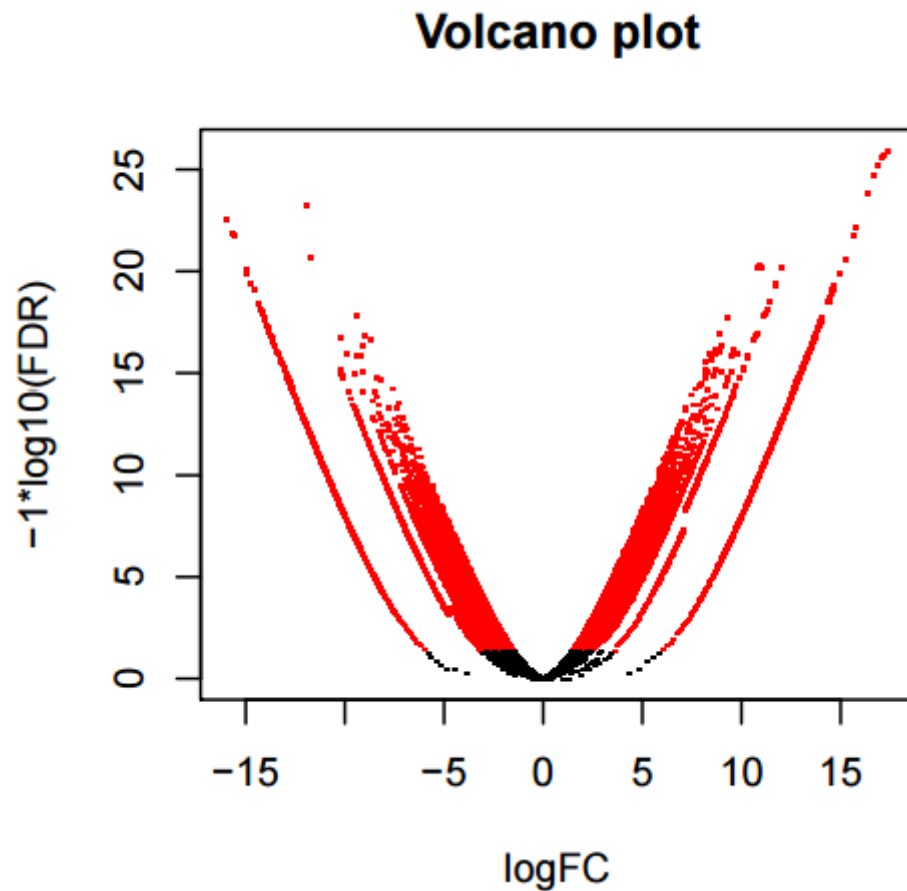
The transcriptomes were analyzed using the pipeline displayed in Fig 1. Briefly, the Trianotate pipeline was utilized to annotate each *de novo* assembled transcriptome and to determine the differentially expressed genes between the EXP vs CON groups and the NSA vs CON groups. The annotated sequences from each group were subjected to two filtering steps. The first filtering step was to screen out select annotated GO terms of interest from each transcriptome. Secondly, these genes were further filtered into a group that met the selection criteria of having a log fold change (LFC) of  $\pm 2$ . This concentrated group of genes was then used as the input for Cytoscape analysis. The large Cytoscape network was filtered by using the search term “Androgen” to concentrate the network down. This resulted in the visualization of molecular interaction networks and biological pathways impacting the androgen receptor pathway. Heatmaps were used to visualize the differentially expressed genes found within the following interacting networks with the androgen receptor pathway: Steroid Hormone Receptor Pathway, Apoptosis, and Response to Heat. Select genes from these heatmaps were highlighted in a concentrated volcano plot.

### 2.7.2 Sexually Active Exposed Testis vs Sexually Active Control Testis Volcano Plot

The comparison between the sexually active exposed (EXP) testis vs the sexually active control (CON) testis yielded a total of 223,024 genes as being differentially expressed between the two groups (Fig 18). The EXP testis had 76,931 genes down-

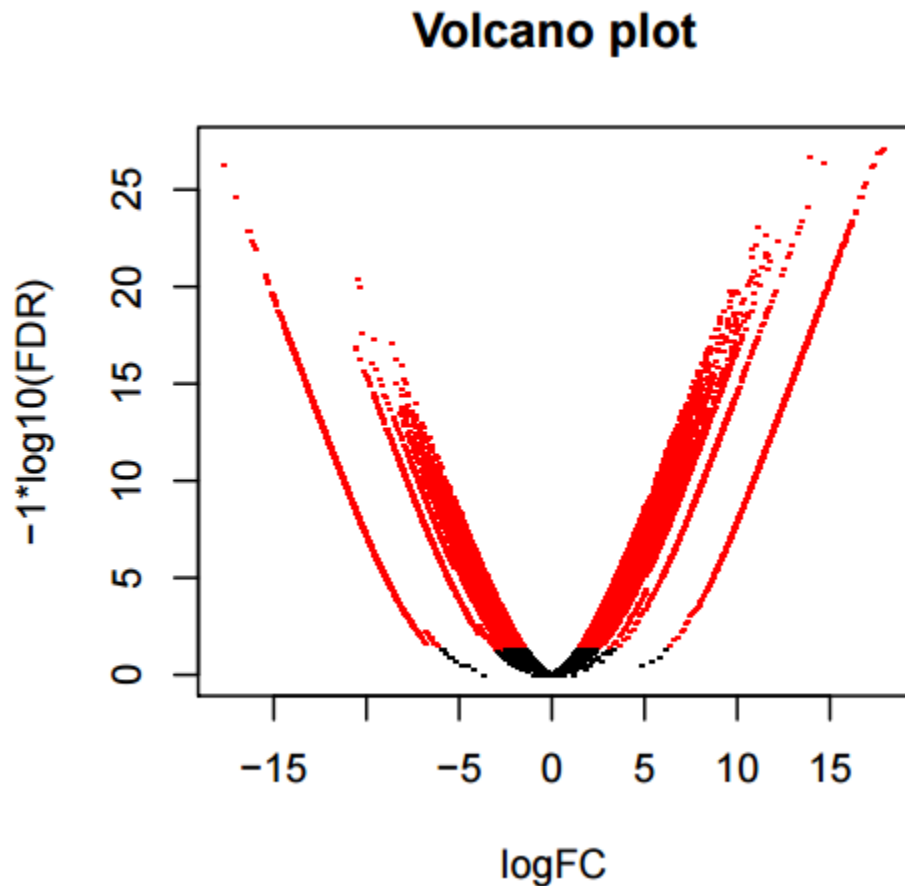


regulated and 146,093 genes up-regulated. The converse was true for the CON testis where 76,931 genes were up-regulated and 146,093 genes down-regulated. Results indicated that the EXP group had various genes and pathways turned on in response to the crude oil.



**Fig. 18.** Volcano plot showing log fold-change and FDR for the comparisons of sexually active exposed (EXP) testis vs the sexually active control (CON) testis. The genes with a negative log fold-change are down-regulated in EXP (up-regulated in CON) and the genes with a positive log fold-change are up-regulated in EXP (down-regulated in CON). Genes that are significantly differentially expressed at a false discovery rate (FDR) of 0.05 are shown in red, and the genes that are not significantly differentially expressed at a FDR of 0.05 are shown in black.

The comparison between the non-sexually active (NSA) testis vs the sexually active control (CON) testis yielded a total of 246,180 genes as being differentially expressed between the two groups (Fig 19). The NSA testis had 139,931 genes down-regulated and 106,249 genes up-regulated. The converse was true for the CON testis, which means that more genes associated with sexual activity were up-regulated than down-regulated. Results globally indicated that various genes and pathways are becoming up-regulated during sexual activation.

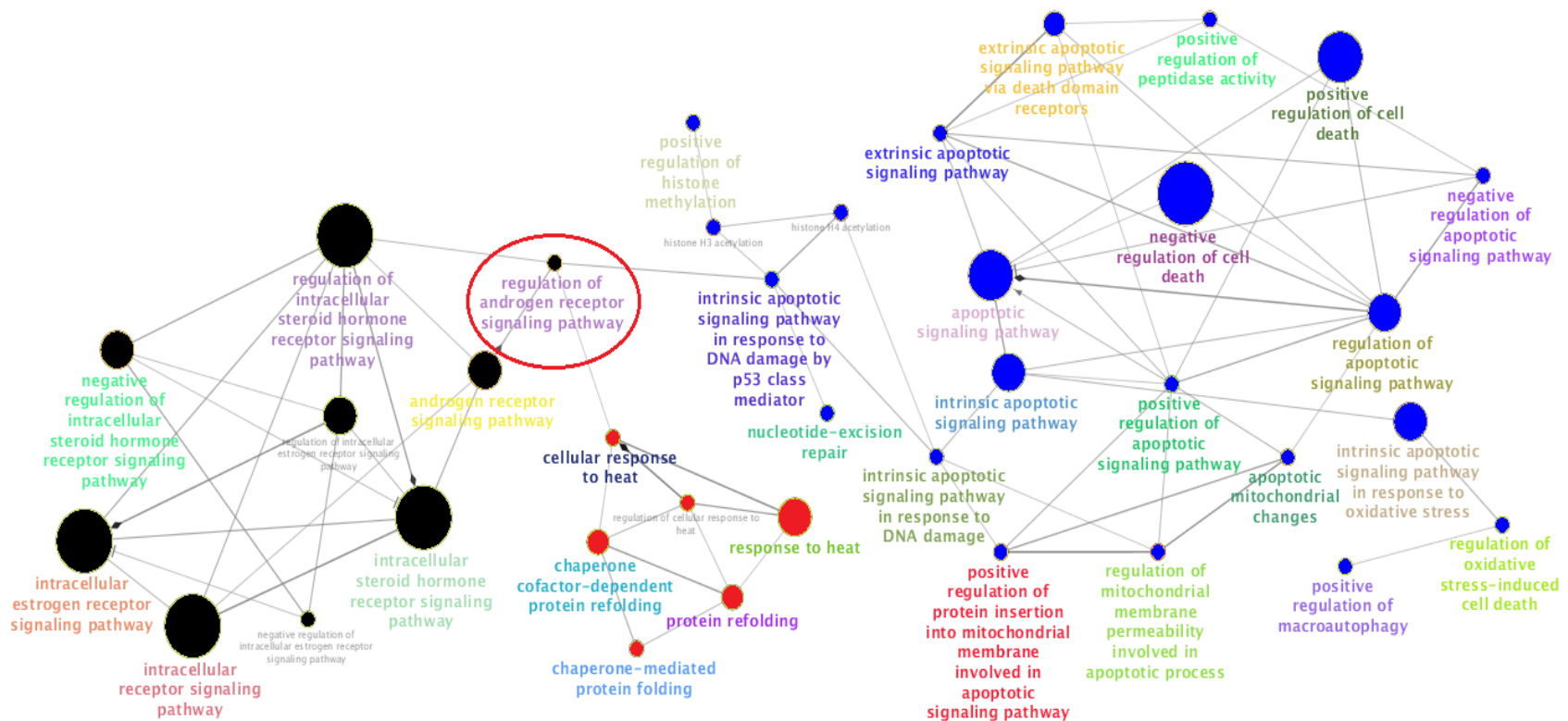


**Fig. 19.** Volcano plot showing log fold-change and FDR for the comparisons of non-sexually active (NSA) testis vs sexually active control (CON) testis. The genes with a negative log fold-change are down-regulated in NSA (up-regulated in CON) and the genes with a positive log fold-change are up-regulated in NSA (down-regulated in tCON). Genes that are significantly differentially expressed at a false discovery rate (FDR) of 0.05 are shown in red and the genes that are not significantly differentially expressed at a FDR of 0.05 are shown in black.

### 2.7.3 Cytoscape Network: Regulation of Androgen Receptor Signaling Pathway

The Cytoscape App ClueGo was used to investigate the differentially expressed genes with an LFC of at least  $\pm 2$  within the EXP vs CON group to determine GO categories interacting with the search term “Androgen”. Results showed an interesting GO category node, “Regulation of Androgen Receptor Signaling Pathway” (circled in

red), within a subcluster of nodes associated with “Steroid Hormone Receptor Signaling Pathways” (shown in black). The “Regulation of Androgen Receptor Signaling Pathway” node was interacting with various groups of other GO category nodes including “Response to Heat” (shown in red) and “Apoptosis” (shown in blue) as seen in the Cytoscape network shown in Fig. 20. The connection of differentially expressed genes in the “Regulation of Androgen Receptor Signaling Pathway” node with those in the “Response to Heat” and “Apoptosis” nodes signifies that the crude oil gavaged fish were responding to a stressful condition at the cellular and molecular level.



**Fig. 20. Cytoscape Network: Steroid Hormone Receptor Signaling Pathway.** Cytoscape network utilizing the ClueGO App to show input genes with a logFC  $\pm 2$  grouped into GO category nodes interacting with other GO categories via edges (lines). The “Regulation of Androgen Receptor Signaling Pathway” node (circled in red) within the “Steroid Hormone Receptor Signaling Pathway” in black is impacted by the GO categories associated with “Response to Heat” in red and “Apoptosis” in blue.

#### 2.7.4 Heatmap:Apoptosis Network

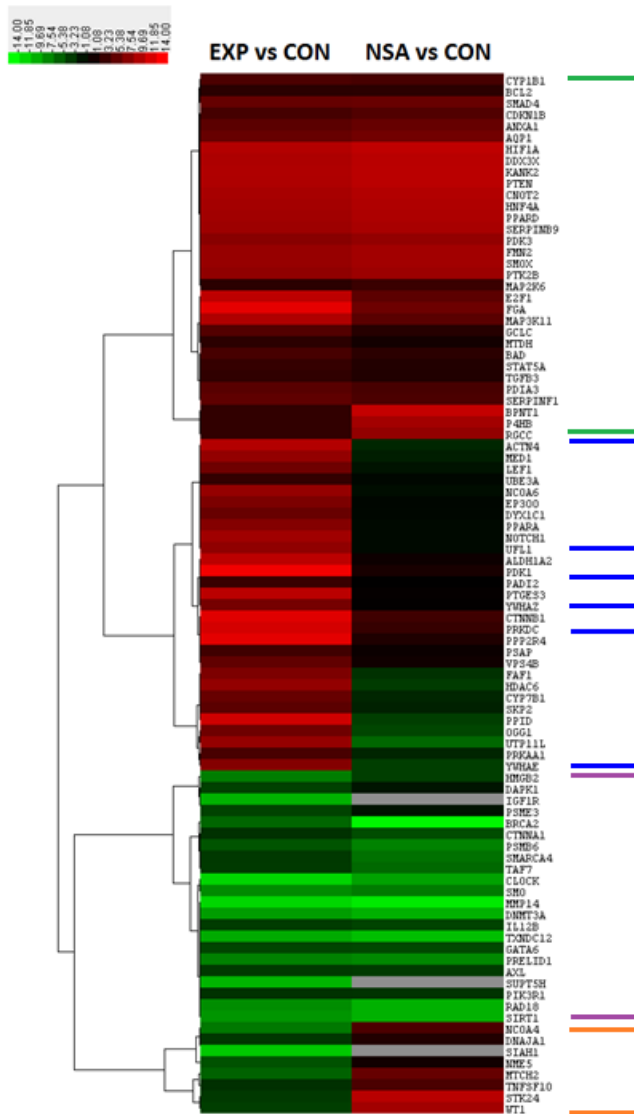
Heatmaps were generated by comparing NSA vs CON and EXP vs CON. Red color indicates up-regulation, green color down-regulation and grey color no expression detected in one of the two groups compared. In general, if the NSA vs CON column is the same color as the EXP vs CON column, then the crude oil exposed fish, which were sexually active, were responding in the same manner as the non-sexually active fish. This indicates that expression of genes associated with sexual activity were modified, either up (red) or down (green) regulated. A green color in the NSA vs CON column indicates that the gene is up-regulated with sexual activity. A red color in the NSA vs CON column indicates that the gene is down-regulated with sexual activity. A green color in the EXP vs CON column indicates that the gene was up-regulated during sexual activation but negatively impacted by crude oil exposure. A red color in the EXP vs CON column indicates that the gene was up-regulated in response to the crude oil. It should be noted that a single gene can be present in more than one node and hence in more than one heatmap.

The expression of the genes incorporated in the numerous nodes affiliated with the “Apoptosis” network (shown in black) in Fig. 20 can be found in the heatmaps shown in Fig. 21. Both heatmaps show significant, differentially expressed genes for EXP vs CON and for NSA vs CON. Overall, 61 genes were up-regulated and 30 genes were down-regulated in the EXP group. The blue brackets contain genes that were up-regulated by crude oil exposure and down-regulated in the NSA group. This was 20.8 % (19 of 91) of the up-regulated genes presented and included PPID and PDK1. The orange

brackets show genes that were down-regulated in EXP vs CON but up-regulated or similar in NSA vs CON. This indicates that crude oil can down-regulate genes that were not associated with sexual activity. These types of genes accounted for 8.7 % (8 of 91) of the down-regulated genes represented and included DNAJA1.

When genes in both columns share the same color, it indicates that the response of the EXP group was similar to that of the NSA group. This is seen for genes highlighted by the green and purple brackets. The green bracket shows genes up-regulated by crude oil in sexually active fish that were normally down-regulated by sexual activity. These genes accounted for 46.1 % (42 of 91) of those presented and included PDK3 and PDK1. The purple bracket shows crude oil down-regulated genes in sexually active fish that were normally up-regulated by sexual activity. These genes accounted for 24.1 % (22 of 91) of those presented and included MMP14 and SIRT1. Overall, similar responses of EXP and NSA groups indicated that crude oil exposed fish were responding like non-sexually active fish and these accounted for 70.3 % (64 of 91) of the genes found in the apoptosis network.





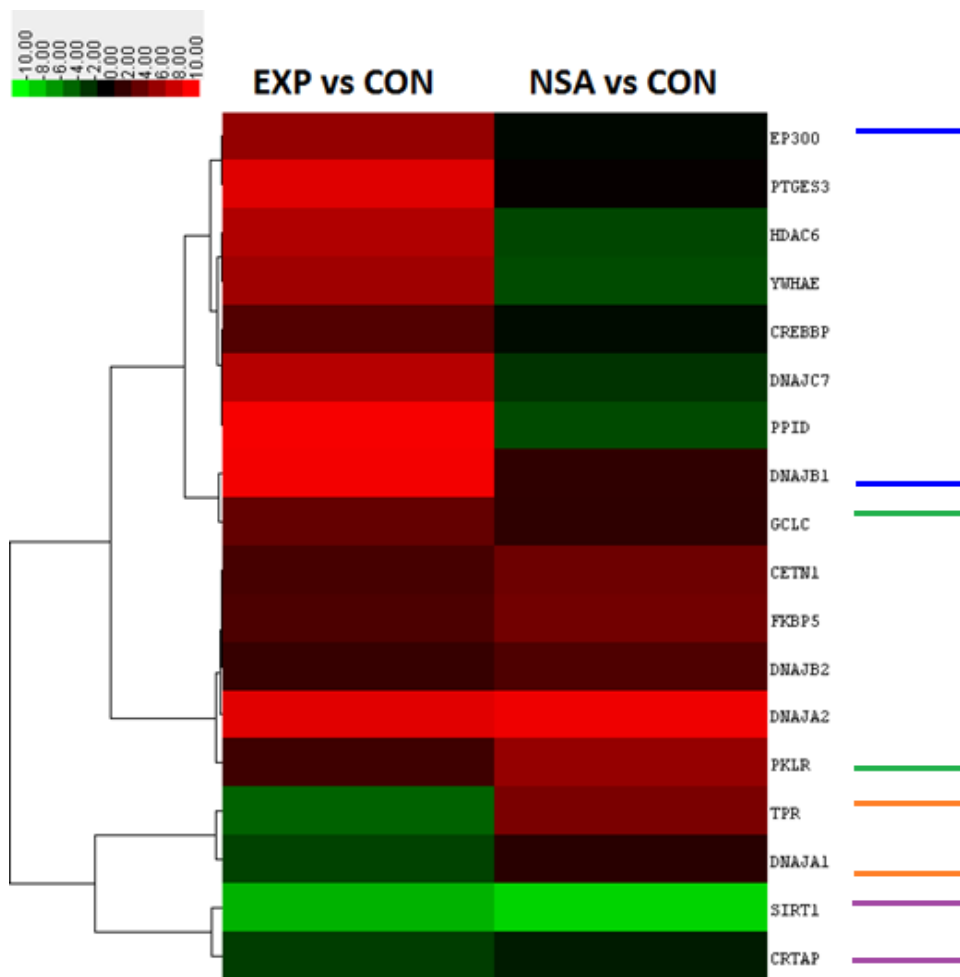
**Fig. 21 Apoptosis heatmap.** Both heatmaps show significant, differentially expressed genes for the EXP vs CON testis (column) and the NSA vs CON testis (column). The log fold-change of each gene's (row) expression value is contingent upon the expression value of the CON testis. The color bar indicates the degree of the log fold-change where green is down-regulated and red is up-regulated compared to CON testis. Grey indicates that the testis group being compared to the CON testis had no detectable expression.

### *2.7.5 Heatmap:Response to Heat Network*

Expression levels of genes found in the “Response to Heat” Cytoscape network (Fig. 20) are shown in Fig. 22. Overall, 14 genes were up-regulated and 4 genes were down-regulated in the EXP group. Genes that were up-regulated in the EXP group compared to both the NSA and CON groups are highlighted by the blue bracket. This was 38.8% (7 of 18) of the up-regulated genes presented and included HDAC6 and YWHAE. The response signifies that genes associated with sexual-activation were further up-regulated by crude oil exposure. The down-regulated EXP genes compared to the CON group are highlighted by the orange bracket. This response indicated that crude oil can down-regulate genes not associated with sexual activity. These types of genes accounted for 11.1 % (2 of 18) of the down-regulated genes represented and included DNAJA1.

As observed in the Apoptosis heatmap, numerous genes within the EXP group had the same expression as the NSA group. This is seen within the group of genes highlighted by the green and purple brackets. The green bracket represents genes that were up-regulated in both the EXP and NSA groups and down-regulated in the CON group. These genes accounted for 38.8 % (7 of 18) of those presented and included DNAJA2 and FKBP5. The purple bracket represents a group of genes that were down-regulated in the EXP and NSA groups and up-regulated in the CON group. Sexual activation may account for the differential expression observed between the NSA and CON groups. However, the crude oil exposure appeared to account for the down-regulated expression in EXP and NSA groups. This gene expression pattern accounted for 11.1 % (2 of 18) of the genes and included SIRT1. Overall, it is quite evident that the

crude oil exposure caused differentially expressed genes amongst the EXP, CON, and NSA groups within the “Response to Heat” network.



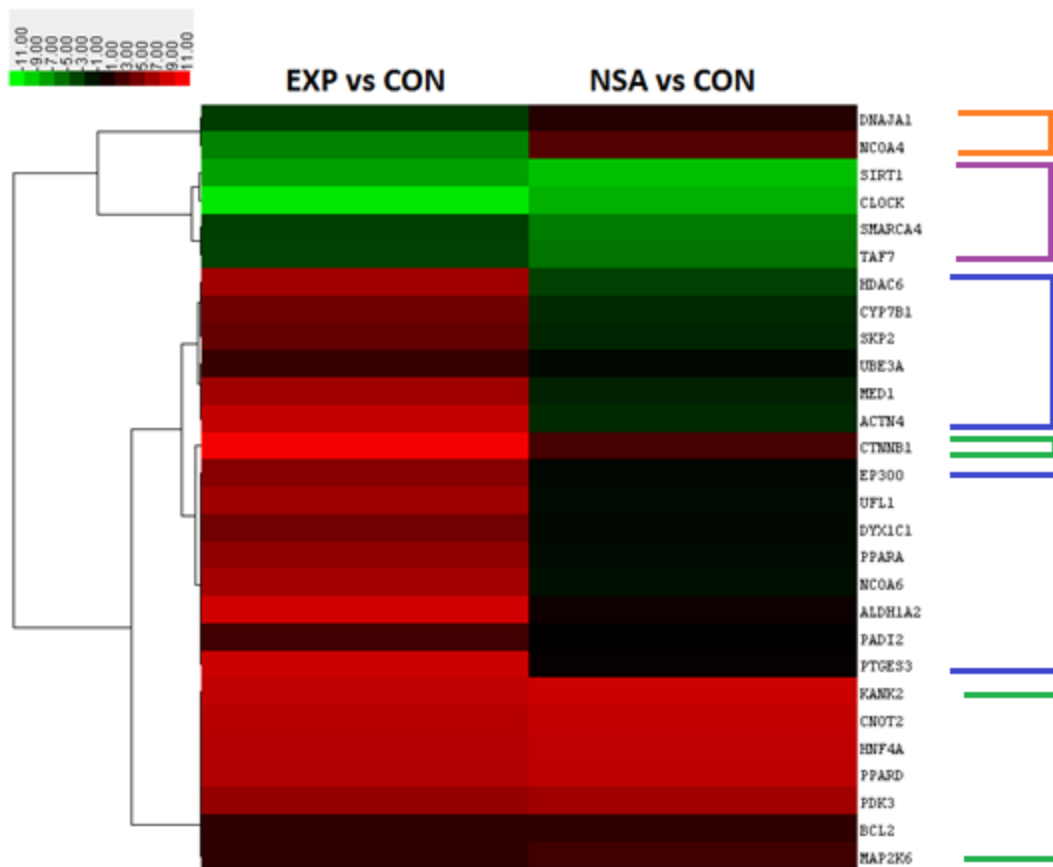
**Fig. 22 Response to Heat heatmap.** Both heatmaps show significant, differentially expressed genes for the EXP vs CON testis (column) and the NSA vs CON testis (column). The log fold-change of each gene's (row) expression value is contingent upon the expression value of the CON testis. The color bar indicates the degree of the log fold-change where green is down-regulated and red is up-regulated compared to CON testis. Grey indicates that the testis group being compared to the CON testis had no detectable expression.

### *2.7.6 Heatmap: Steroid Hormone Receptor Signaling Pathway Network*

Expression levels of genes found in the “Steroid Hormone Receptor Signaling Pathway” cytoscape network (Fig. 20) are shown in Fig. 23. Both heatmaps show significant, differentially expressed genes for EXP vs CON and NSA vs CON. When comparing EXP vs CON, 18 genes were up-regulated and 6 were down-regulated. This indicated that crude oil exposure was modulating 85.7% (24 of 28) of genes associated with the steroid hormone receptor signaling pathway. Comparison of EXP vs CON with NSA vs CON showed that some genes up-regulated in response to crude oil were down-regulated or similar in both the NSA and CON groups (highlighted in the blue brackets). These accounted for 42.9 % (12 of 28) of the genes presented and included HDAC6. These genes are associated with the steroid hormone receptor signaling pathway but not necessarily sexual activation. The orange brackets show genes that were down-regulated in EXP vs CON but up-regulated in NSA vs CON. This also indicated that non-sexually active genes can be impacted by crude oil. This group of genes consisted of 7.1% (2 of 28) of the down-regulated EXP genes. Overall, a majority of the differentially expressed genes in the steroid hormone receptor signaling pathway were modulated by crude oil exposure. However, comparisons between EXP vs CON and NSA vs CON showed that some genes responding to crude oil were not necessarily related to sexual-activation even though they were within GO categories associated with the steroid hormone receptor signaling pathway.

As observed in the Response to Heat and Apoptosis heatmaps, some genes in the EXP and NSA groups had similar levels of expression. The green bracket represents

genes that were both up-regulated in the EXP and NSA groups and down-regulated in the CON group. This group of genes accounted for 35.7% (10 of 28) of the genes within the EXP group and included PPARD and KANK2. The purple bracket shows crude oil down-regulated genes in sexually active fish that were normally up-regulated by sexual activity. These genes accounted for 14.3 % (4 of 28) of those presented and included CLOCK and SIRT1. Overall, similar responses of EXP and NSA groups indicated that crude oil exposed fish were responding like non-sexually active fish and these accounted for 50.0 % (14 of 28) of the genes found in the Steroid Hormone Signaling Pathway network. These genes represented an endocrine disrupting effect of crude oil exposure.



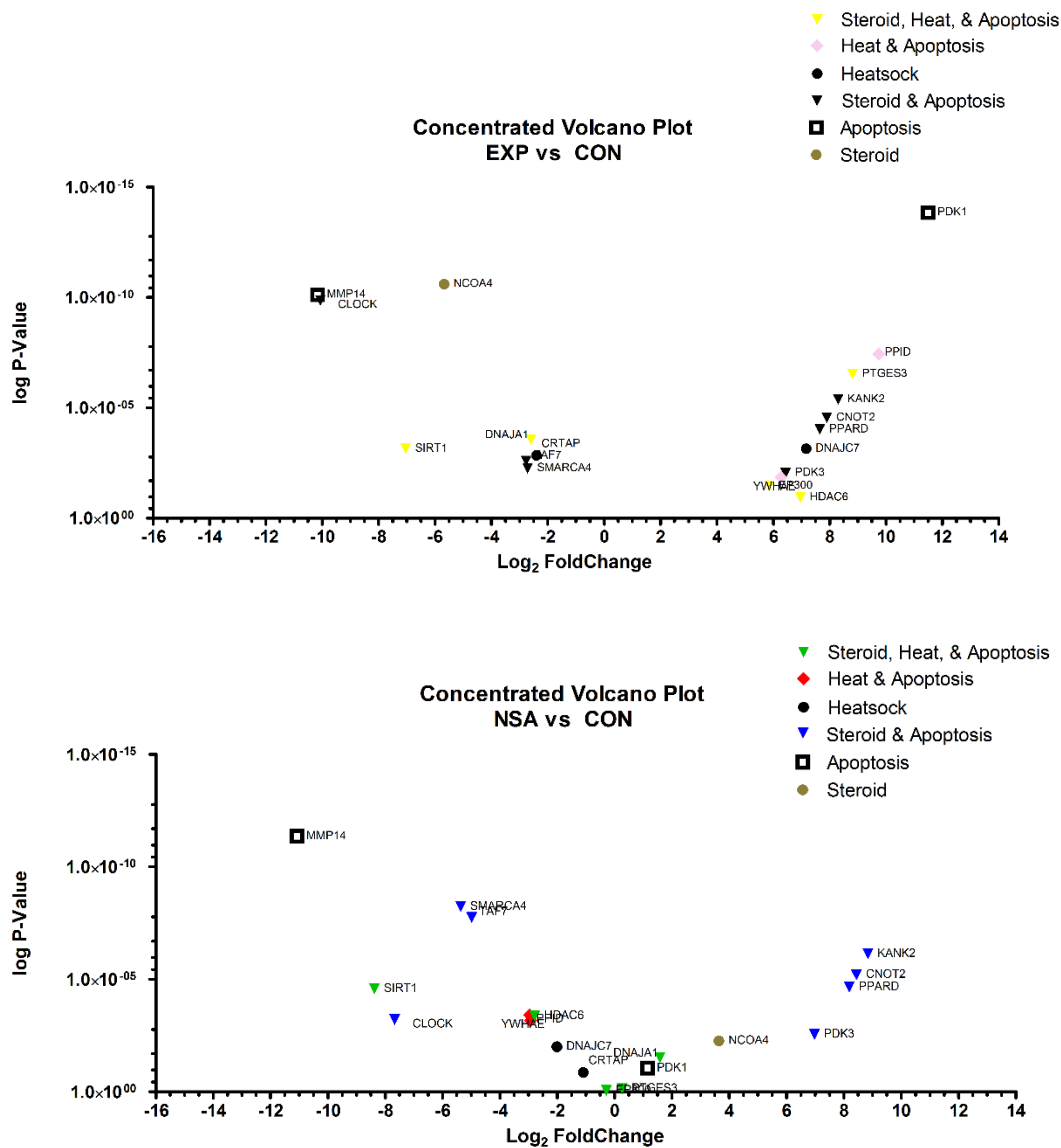
**Fig. 23 Steroid Hormone Receptor Signaling Pathway heatmap.** Both heatmaps show significantly differentially expressed genes for the EXP vs CON testis (column) and the NSA vs CON testis (column). The log fold-change of each gene's (row) expression value is contingent upon the expression value of the CON testis. The color bar indicates the degree of the log fold-change where green is down-regulated and red is up-regulated compared to CON testis. Grey indicates that the testis group being compared to the CON testis had no detectable expression.

#### 2.7.8 Concentrated Volcano plot.

EXP vs CON and NSA vs CON concentrated volcano plots (Fig. 24) were created based on select genes of interest from the Apoptosis, Response to Heat, and Steroid Hormone Receptor Signaling Pathway heatmaps (Figs. 21 to 23). The selection focused on genes appearing in more than one heatmap, the EXP group acting like the NSA group

with the same expression pattern, and the appearance of a heatmap category-specific response to the crude oil. The following genes appeared in all three heatmaps signifying that they played an integral part within the androgen receptor signaling pathway in response to the crude oil: DNAJA1, EP300, SIRT1, HDAC6, and PTGES3. The Steroid Hormone Receptor Signaling Pathway and Apoptosis groups contained the following genes with the same expression patterns observed in both the EXP and NSA: PPARD, SMARCA4, CLOCK, TAF7, CNOT2, KANK2, and PDK3. The genes PPDI and YWHAЕ were also found in the Response to Heat group. Both of these genes were significantly up-regulated compared to the CON and NSA groups signifying that they were involved in the response to crude oil. Within the Response to Heat group, DNAJC7 was significantly up-regulated compared to both CON and NSA, and CRTAP's EXP expression pattern was similar to what was observed in the NSA group. PDK1 and MMP14 are representatives of the Apoptosis group and were used to validate the transcriptomes with qPCR. PDK1 was involved in the response to the crude oil as it was significantly upregulated compared to the CON and NSA groups. MMP14 shared the same expression patterns in both the EXP and NSA groups which was the opposite of what was observed in the CON group. A striking representative within the Steroid Hormone Receptor Signaling Pathway heatmap was NCOA4, which appeared to be negatively impacted by the crude oil exposure. Overall, these highlighted genes appear to be candidate biomarkers for gonadal-derived endocrine disruption due to crude oil exposure.





**Fig. 24.** Concentrated Volcano Plots. Relative expression of genes in EXP vs CON (top) and NSA vs CON (bottom) heatmaps. Genes were selected from the Steroid Hormone Receptor Signaling Pathway in the Cytoscape network (Fig. 20). Each point represents a gene found within more than one of the following heat map groups: Response to Heat, Apoptosis and Steroid Hormone Receptor Signaling Pathway.

## 2.8. qPCR validation

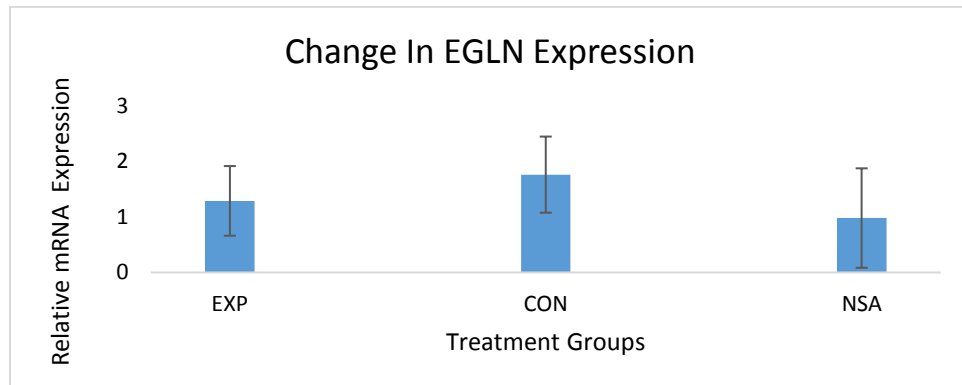
Ten primer sets were developed for qPCR using sequence information obtained from NGS. PCR bands were cloned, sequenced and aligned using BLASTX to nr databases in GenBank. Results for all ten primer sets showed high identity to fish sequences that ranged from 83 to 100% identity (Table 12). Eight of the ten sequences best matched Atlantic killifish, *Fundulus heteroclitus*. These findings supported the quality of the transcriptome assembly.

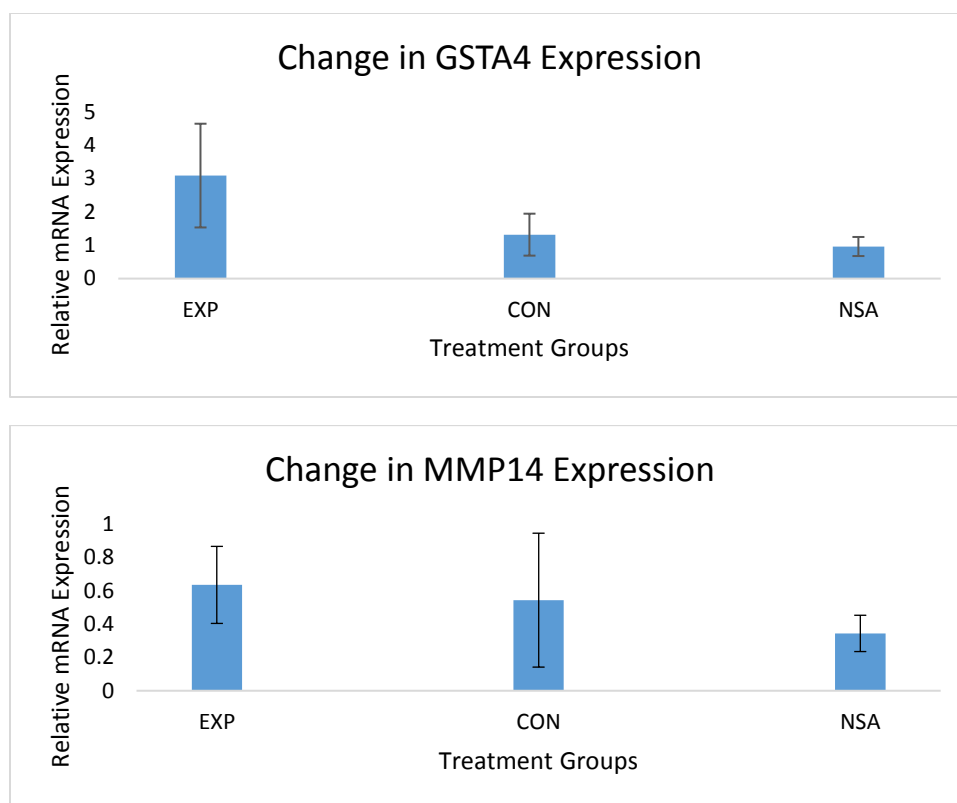
qPCR analysis was performed using six of the primer sets. Three biological replicates were analyzed using the  $\Delta\Delta C_t$  method to validate the bioinformatics expression analysis for EGLN, GSTA4, MMP14, AK7, TDRD7b, and PDK. All data were normalized against  $\beta$ -actin, as a housekeeping gene. The mRNA expression levels were presented as mean  $\pm$  SD (n= 3). The patterns of differential expression were consistent with the results from the expression analysis for EGLN, GSTA4, AK7, and PDK1 for the EXP vs CON, and NSA vs CON groups as shown in Figs. 25-26. The expression pattern for MMP14 and TDRD7b in both EXP vs CON groups are not in line with the bioinformatics analysis but the NSA vs CON relationships are. Overall, 10 of the 12 expression relationships corresponded with the qPCR as seen in Table 13. Variability was observed for biological triplicates of each treatment group. The correlations for the average of all fish used in the qPCR as well as for just the fish used in the transcriptome can be seen in Figs. 27. The transcriptome fish had a better match with the Trinotate data than the average of the three fish used in qPCR. Both positive (GSTA4 and PDK1) and negative correlations (ENLG, MMP14, and TDRD7b) were observed. The positive

correlations were not statistically significant with  $p = 0.342$  and  $0.301$  for transcriptome only and all fish, respectively. However, a strong relationship for Trinotate and qPCR was found for those genes showing a negative correlation with  $p = 0.015$  and  $0.037$  for transcriptome only and all fish, respectively. Overall, the correlations do show a good relationship between the qPCR data and Trinotate data, especially for the negative correlations.

**Table 12.** Cloned sequences aligned-percent-identities with sequences in GenBank

Sequence Name	Primer	Blast Hit	Hit Identity %
EGLN2	Forward	<i>Fundulus heteroclitus</i>	96%
EGLN2	Reverse	<i>Fundulus heteroclitus</i>	99%
PDK1	Forward	<i>Nothobranchius furzeri</i>	93%
PDK1	Reverse	<i>Poecilia reticulata</i>	96%
DNAJB	Forward	<i>Fundulus heteroclitus</i>	96%
DNAJB	Reverse	<i>Fundulus heteroclitus</i>	97%
ERCC2	Forward	<i>Fundulus heteroclitus</i>	85%
ERCC2	Reverse	<i>Fundulus heteroclitus</i>	96%
AK7	Forward	<i>Fundulus heteroclitus</i>	90%
AK7	Reverse	<i>Fundulus heteroclitus</i>	95%
TDRD7	Forward	<i>Poecilia reticulata</i>	83%
TDRD7	Reverse	<i>Salmo salar</i>	91%
PPARD	Forward	<i>Fundulus heteroclitus</i>	93%
PPARD	Reverse	<i>Fundulus heteroclitus</i>	95%
MMP14	Forward	<i>Fundulus heteroclitus</i>	94%
MMP14	Reverse	<i>Fundulus heteroclitus</i>	94%
PTEN	Forward	<i>Fundulus heteroclitus</i>	100%
PTEN	Reverse	<i>Fundulus heteroclitus</i>	98%
GSTA4	Forward	<i>Fundulus heteroclitus</i>	97%
GSTA4	Reverse	<i>Fundulus heteroclitus</i>	95%





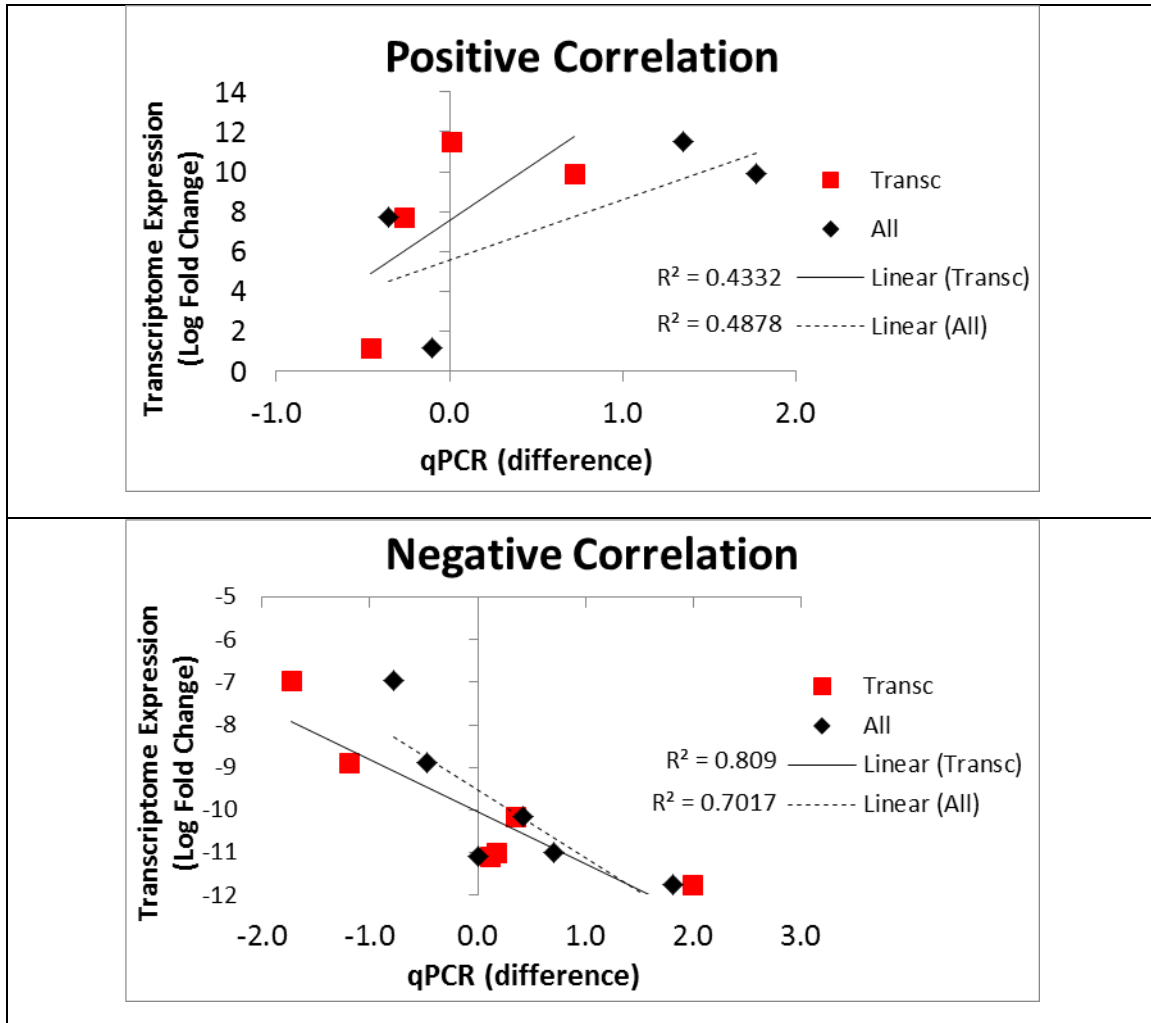
**Fig. 25** Changes in EGLN, GSTA4, and MMP14 gene expression measured by qPCR. The mRNA expression levels are presented as mean  $\pm$  SD (n= 3)



**Fig. 26** Changes in AK7, TDRD7b, and PDK1 gene expression measured by qPCR. The mRNA expression levels are presented as mean  $\pm$  SD (n= 3)

**Table 13.** Trinotate expression relationships with qPCR analysis

Gene	Trinotate EXP vs CON	qPCR EXP vs CON	Trinotate NSA vs CON	qPCR NSA vs CON
EGLN2	↓	↓	↓	↓
GSTA4	↑	↑	↑	↑
MMP14	↓	↑	↓	↓
AK7	↓	↓	↓	↓
TDRD7b	↓	↑	↓	↓
PDK1	↑	↑	↔	↔



**Fig. 27** Correlation between transcriptome expression and qPCR. Genes were separated into positive (GSTA4, PDK1) or negative (ENGLN2, MMP14, TDRD7b) correlations. AK7 results were not included. Transcriptome expression data are for EXP vs CON and NSA vs CON. qPCR data are the difference of EXP minus CON and NSA minus CON averages. Averages for all three fish used in the qPCR are shown (All) or just those fish selected to make transcriptomes (Transc). Trend lines are shown with  $R^2$  values calculated using Pearson Bivariate Correlation (SSPS). Negative correlations are statistically significant with  $p = -0.838$  and  $-0.899$  for All and Transc, respectively



## Discussion

### 3.1. Selection of fish for NGS

Currently, no research has been presented on testis-derived reproductive disruption due to crude oil exposure in killifish (*F. heteroclitus*). One of the goals of this work was to establish exposure times at which the reproductive system of *F. heteroclitus* responds to crude oil at the molecular level and to generate gonadal RNAs for transcriptome analysis. Another objective was to create an annotated characterization of the testis transcriptomes in *F. heteroclitus* to identify key genes and pathways associated with disrupted physiological function of testis due to crude oil exposure. This was accomplished by applying Illumina NextSeq 500 technology to the following three experimental treatments in *F. heteroclitus* to determine the genes and molecular pathways that get turned on during sexual activation and how they are impacted due to crude oil exposure: 1) an exposed spawning male gavaged with crude oil collected from the DeepWater Horizon oil rig prior to the accident; 2) a control spawning male gavaged with fish oil; and 3) a control non-spawning male. In doing so, modern biomarkers were identified to assess natural resource damage caused by oil spills. These biomarkers can be used in future studies to link molecular responses to population level effects. Hence, impaired reproductive responses will be associated with detectable molecular responses.

In order to generate transcriptomes in which endocrine disruption is occurring, it is necessary to distinguish between reproductively active and inactive females and males and to determine at what time point crude oil is causing endocrine disruption. Male killifish in spawning condition have bright colors on their anal fin and sides.

Reproductively active females undergo vitellogenesis. During this process, vitellogenin, the egg yolk protein precursor is synthesized in the liver, secreted into the plasma, and transported to the oocytes for uptake (Levi et al., 2009, Nicolas et al., 1998, and Bermanian et al., 2004). The most dominant trigger of vitellogenin expression is the ovarian steroid hormone 17 $\beta$ -estradiol (E2) that is synthesized by CYP19a in the gonad (Cheshenko et al., 2008). Therefore, expression of the VTG gene in liver as well as CYP19a in ovary is used to determine whether or not female killifish are reproductively active.

The approach of the experiment was to select endocrine disrupted fish for gonadal transcriptome analyses that had known levels of crude oil exposure. Select biomarkers (CYP19a, VTG, and CYP1A) known to respond to EDCs such as PAHs were used to select affected fish. Several laboratory studies have documented oil-related declines in reproductive parameters in marine teleosts such as alterations in levels of reproductive hormones, inhibited gonadal development, and reduced egg and larval viability (Idler et al., 1995, Thomas et al., 1995, Truscott et al., 1992). In field studies, Bugel et al (2010) showed that the reproductive health of male *F. heteroclitus* can be impacted by endocrine disrupting compounds (EDCs). He found that male killifish from Newark Bay, a chronically contaminated estuary, had decreased gonad weight, altered testis development, and decreased CYP19a mRNA expression that corresponded with increased PAHs in gall bladders. Additionally, there is *in vivo* evidence that PAH exposure resulted in reproductive and developmental deficits in female killifish collected from PAH-impacted sites (Patel et al., 2006). This study showed that the field collected

female killifish exposed to the PAH, benzo(a)pyrene (BaP), in the laboratory for 15 days experienced inhibited ovarian aromatase (CYP19a) activity. In other studies, PAHs have had deleterious effects on the vitellogenesis in female fish such as the reduction in circulating hormones and plasma vitellogenin, estrogenic and antiestrogenic effects, retardation of oocyte maturation, and reduction of reproductive success (Nicolas et al., 1998). In males, the VTG gene is present but normally silent (Matozzo et al., 2008). However, it may be activated by xeno-estrogens and this activation leads to the feminization of males, signifying endocrine disruption. Aryl hydrocarbon receptor (AhR) ligands, such as PAHs, have also been shown to be capable of inducing CYP1A expression while disrupting 17 $\beta$ -estradiol-induced expression of VTG and reducing ER $\alpha$  levels (Bemanian et al., 2004 and Patel et al., 2006). Therefore, in the present study, evidence of endocrine disruption was expected to be associated with 1) up-regulation of CYP1A in both genders, 2) up-regulation of VTG and down-regulation of CYP19a in male killifish and 3) down-regulation of VTG and CYP19a in female killifish.

The selection of sexually active female fish for transcriptomes was based on the following criteria. Of the three possible Day 7 female controls, GC6 and GC7 were selected as both showed higher levels of CYP19a than found in the non-sexually active fish (Fig 19). However, expression of VTG in these control female fish decreased between Day 0 and Day 7 for unknown reasons. Loss of VTG expression was not consistent with the high level of CYP19a. Both LC6 and LC7 were expressing CYP1A signifying a potential previous exposure; although, levels were less than half of that in exposed fish. Previous exposure is possible as these were wild fish. In addition, most

species of fish have relatively low levels of CYP1A constitutively expressed (Patel et al., 2006 and Bemanian et al., 2004). Of the five Day 7 female exposed fish, only GE11 was selected for transcriptome analysis. GE11 showed down regulation of CYP19a, which indicated endocrine disruption, as well as high levels of PAH-like compounds in its EEMS scan. Research has shown down regulation of CYP19A mRNA in females exposed to PAHs (Patel et al, 2006). In terms of the other females, two had gonads that were too small for RNA isolation and two showed unexpectedly low levels of CYP1A in liver despite exposure. In summary, GC6 and GC7 were chosen to represent female, sexually active, unexposed fish, and GE 11 was chosen to represent female, sexually active, exposed fish.

Of the six possible Day 7 male control fish, GC3 and GC11 were chosen for transcriptome analysis. The other four candidates were eliminated because gonadal RNA was not available for GC1 and GC2, and both GC5 and GC8 were expressing high levels of CYP1A similar to those of the exposed fish (Fig. 9). It should be noted that GC11's liver was lost during RNA preparation so both VTG and CYP1A expression analysis could not be analyzed. GC3 and GC11 were selected because they showed low levels of CYP19a as expected for male fish and had levels of PAH-like compounds about 10 fold lower than the exposed group in EEMS scans. Of the eight Day 7 male exposed fish, both GE1 and GE6 were selected. GE1 and GE6 had low levels of CYP19a: it was suppressed below those in control (Fig. 9). This result did not indicate feminization with potentially increased estrogen. It did correspond with Bugel et. al.'s finding of smaller gonads and reduced CYP19a in male killifish from Newark Bay (2010). Results for VTG were mixed

with expression levels in LE1 about 100x higher than in LE6. It was decided not to use this biomarker for transcriptome selection as expression levels declined from Day 1 to Day 7. Both fish also showed increased levels of CYP1A and PAH-like compounds supporting exposure to AH ligands. In summary, GC3 and GC11 were chosen to represent male, sexually active, unexposed fish, and GE 1 and GE6 were chosen to represent male, sexually active, exposed fish.

The results for the non-sexually active (NSA) expressions for CYP19a, CYP1A, and VTG can be seen in Fig. 8. The NSA Male, G5, was chosen for transcriptome analysis because it expectedly was not expressing CYP19a or VTG and had very little CYP1A. It was expected that VTG would not be expressed in NSA males because VTG is the egg yolk protein precursor. The NSA female, G7, was selected for transcriptome analysis because it was normally producing CYP19a and not producing CYP1A. However, it was unexpectedly expressing VTG. Non-sexually active females should not have been producing high levels of VTG because they were out of spawning season (Nicolas et al, 1999). The unexpected expression of VTG may be related to their laboratory food and estrogen build up within the aquatic atmosphere of the tank. The NSA Female, G8, was also selected for transcriptome analysis because it was normally producing CYP19a.

The influence of fish size on gene expression was evaluated as shown in Fig. 10. Statistical analyses showed that the EXP and NSA groups were significantly larger than the CON group for both weight and length (Table 4). Pearson Correlation tests were performed to determine if the differences in size had an impact on the expression of the

three biomarkers used for NGS selection. All fish weights (g) for Day 7 EXP, Day 7 CON, and NSA were correlated to the semi-quantitative PCR expression for CYP19a, VTG, and CYP1a and it was shown that weight did not influence the expression of genes used to select the fish for NGS. Therefore, the only factor contributing to the change in gene expressions was state of sexual activity and crude oil exposure.

The CYP19a, VTG, and CYP1A biomarker analysis for all the fish did not always correspond with the expected results seen in other studies (Bugel et. al., 2010, Nicolas et. al., 1998, Bemanian et. al., 2004 and Patel et al., 2006). It should be noted that in these other studies the biomarker response was to PAH exposure and not crude oil exposure. Crude oil is a complex mixture of various chemicals including PAHs. On the other hand, the exposure for non-responsive fish might have not been long enough or concentrated enough to trigger an expression change in some biomarkers. These fish might have become tolerant to PAHs due to previous exposures given that they were wild fish. The review by Nicolas states that elevated levels of PAHs affect the vitellogenic cycle differently between species, populations, and even between individuals (Nicolas, 1999). Therefore, it was very important in the present study to validate that the fish were successfully exposed to the crude oil.

EEMs was utilized to detect the presence of PAHs in the gall bladder as another parameter for selecting fish for sequencing. Crude oil is a complex mixture that consists of four major chemical groups with alkanes and cycloalkanes being the major constituents along with aromatic hydrocarbons and asphaltics (Robbins and Hsu., 1996). Of these four major chemical groups, PAHs have been routinely used for biomonitoring

crude oil (Jung et al., 2011) because of the ease of detected by fluorescence (Kim et al., 2010). The aromatic ring structures of PAHs allow them to be detected and distinguished by unique excitation (light absorption) and emission (fluorescence) wavelengths. The EEMS spectra for gall bladders showed the presence of PAH-like compounds. The 3D scans for control fish primarily showed a protein signature with no apparent PAH-like compounds, whereas the 3D scans for exposed fish showed PAH profiles consistent with the presence of PAHs (Fig. 11). The spectra of exposed, Day 3 fish showed maximum fluorescence at Ex300/Em370-380, which was consistent with the minor peak of the standard for 9-phenanthrol, and at Ex340/ Em385, which was consistent with the minor peak of the standard for 1-hydroxypyrene. By Day 7, the spectra had changed with the major peak at Ex260/EM370-380 nm, which was consistent with the major peak of the standard for 9-phenanthrol. The change in spectra over time indicated not only the presence of PAHs but also the preferential metabolism of hydroxypyrene-like PAHs over phenanthrol-like PAHs. The retention of phenanthrene-like compounds compared to other types of PAHs is supported by the literature (Meador et al., 2008). Indeed, these authors recommend using phenanthrene to represent PAH exposure in field studies.

The general pattern of PAH-like compounds in gall bladders was observed by converting 3D to 2D scans. In these scans, emission was held at 385 nm and excitation was scanned from 260 to 400 nm. Data was normalized by dividing the fluorescence intensity by fish weight. The resulting figure showed that exposed fish had approximately 10x as much PAHs as control fish (Fig. 12). Presence of PAHs in controls was probably due to the fact that the fish were wild and exposed to PAHs in their environment prior to

collection. They may also have received PAHs from their laboratory diet - commercial silversides. The overall outcome from these data was evidence that the gonads containing the molecular biomarkers for endocrine disruption were exposed to the crude oil as seen by the EEMS data.

Based on biomarker analysis, RNA availability, and EEMS analysis, the following fish were chosen for transcriptome analysis: Day 7 male GE1 and GE6, Day 7 female GE11, Day 7 male GC3 and GC11, Day 7 female GC6 and GC7, NSA female 7 and 8, and NSA male 5. Even though this study just concentrated on the endocrine disruption in males exposed to crude oil, it was important to generate transcriptomes of both males and females to help detect female-sexually orientated genes that could be associated with feminization in males. As seen in the Trianotate pipeline, this strategy was employed for the annotating and quantifying of all six transcriptome categories.

### *3.2 De Novo Assembly Comparison Study*

To generate the highest quality transcriptomes, a *de novo* transcriptome assembly comparison study was executed to evaluate the performance of four popular *de novo* assemblers (Bridger, Trinity, Velvet/Oases, and SOAPdenovo-Trans) along with various *k-mer* strategies. Currently, there is not a gold standard for a *de novo* transcriptome assembler. It is known that different *de novo* assemblers using the same transcripts with similar user defined parameters have produced assemblies that vary amongst each other (Moreton et al., 2014; Chopra et al., 2014; He et al., 2015; Schliesky et al., 2012). One of the goals of this work was to compare four recently published *de novo* assemblers and



one user defined parameter, the *k-mer* value, to determine if one strategy was better than another for *de novo* assembly of a non-model organism transcriptome. The assemblers included Trinity, Oases, and SOAPdenovo-Trans (commonly used *de Bruijn* graph-based *de novo* assemblers) as well as Bridger (uses the minimum path cover model to construct splice graphs that are used to build compatibility graphs). Additionally, eleven metrics of assembly quality were compared in order to better form a consensus as to which ones should be used to assess *de novo* transcriptome assemblies. Overall, eleven metrics were used to evaluate six assembly strategies.

Commonly used evaluation statistics, such as the number of contigs, the N50 value, and contig length, were developed for genome assemblies but have also been used to evaluate transcriptome assemblies (Baker, 2012). Better assemblies should theoretically have more reads assembled into longer contigs and thereby higher N50 values. However, importance of the N50 metric has been questioned (Li et al., 2014; Salzberg et al., 2012). Research indicated that N50 values could be artificially increased based on *k-mer* strategy and or the user defined minimum contig length. Short contigs occurred when high *k-mer* values did not assemble short reads of low abundance transcripts or low *k-mer* values assembled short fragmented transcripts due to lack of overlap (Surget-Groba et al., 2010; Chopra et al., 2014). In both cases, if the resulting transcripts were shorter than the minimum contig length parameter established by the user, they were eliminated by the program ultimately generating artificially high N50 values. In the study presented, the influence of high (LMK) and low (SMK) *k-mer* values on evaluation statistics was determined for SOAP and Oases assemblies. Both showed

somewhat higher N50 values for low *k-mer* assemblies, but greater differences were found for the assembler. For example, the N50 lengths for SOAP LMK and SMK were 917 and 1,042, respectively, while those for Oases LMK and SMK were 1,676 and 1,743, respectively (Table 2). Bridger also used a low *k-mer* strategy, and results showed that contig number as well as contigs over 1 kb were twice those of the other assemblies. Taken together, the assembler program had more influence on these traditional, genomic metrics than the *k-mer* values, and the Bridger assembler performed the best in two out of the three metrics.

The quality of an assembly can be assessed by its ability to construct transcripts that align to genes in publically available databases. In the study presented, the six assemblies were evaluated using two well-annotated phylogenetic tree relatives, the CEGMA database, and the BUSCO database. For the phylogenetic tree analyses, killifish transcriptome assemblies were aligned to the reference databases of southern platyfish (*X. maculatus*) and Amazon molly (*P. formosa*) (Fig. 13). Oases assemblies had the best percentages of alignments ranging from 47.00 to 51.82%. Results for the other assemblies were similar but lower with alignments ranging from 32.68 to 39.35%. The low percentage of killifish matches in general could be attributed to the evolutionary distance between killifish and the other two species as other *de novo* transcriptome assembly studies using non-model fish had similar findings (Ji et al., 2012; Huth and Place, 2013; Salisbury et al., 2015). Additionally, the BLASTX results were based on aligning the testis transcriptome of killifish with the entire genomes of the southern platyfish and Amazon molly. This would have inherently caused a lower percentage of

matches. Furthermore, some of the unannotated contigs may be short and consist of sequences lacking well-characterized protein domains, comprised of 3' or 5' untranslated regions, or be non-coding RNAs (Gao et al., 2014). Results also showed that the general performance of each assembly's BLASTX alignment was more dependent on the assembler software than *k-mer* value. For example, assemblies using Oases SMK and LMK had percent alignments of 47.00 and 50.14% in southern platyfish while SOAP had percent alignments of 35.22% and 34.60% for the same species (Table 8).

CEGMA and BUSCO evaluated assembly quality by aligning contigs to core proteins. For CEGMA, most of the 248 core proteins were found with each assembler; therefore, this metric did not distinguish between them (Table 9). Interestingly, some assemblers left out CEGMA proteins found by others. This indicated that particular assemblers had different proficiencies in reassembling clean reads into transcripts as previously reported by Naksugi et al. (Schliesky et al., 2012). For example, KOG0292 (Vesicle coat complex COPI, alpha subunit) was only present in the SOAP SMK and LMK strategies, and KOG2311 (NAD/FAD-utilizing protein possibly involved in translation) was only present in Trinity. In addition, both SOAP and Oases were unable to assemble properly the same transcripts regardless of the *k-mer* strategy. For example, both SOAP SMK and LMK did not assemble at full length KOG0261, KOG0209, KOG0462 or KOG2311. Both Oases SMK and LMK did not assemble at full length KOG0292, KOG2311 or KOG4392. This indicated that the assembler's algorithms had a more influential role than the *k-mer* selection when assembling transcripts. BUSCO showed greater differences between assemblers than CEGMA most likely due to the

larger database of 3,023 vertebrate genes (Fig. 14). Missing BUSCO genes for the six different assemblies ranged from 675 (22.3%) in Trinity to 1,030 (34.1%) in SOAP LMK. As with CEGMA and BLASTX alignments, this metric showed that the assembler rather than the *k-mer* strategy had more influential over the outcome. Both the SMK (28.7%) and LMK (31%) strategies of Oases outperformed the SMK (32.4%) and LMK (34.1%) strategies of SOAP. Overall, results indicated that for either genomes or transcriptomes, if a user's goal was to annotate the most genes possible for a non-model organism, Oases was the assembler of choice. If the goal were to assemble the most accurate transcriptome, Trinity or Bridger would be the best choice. As of 2015, CEGMA is no longer actively supported and has been essentially replaced by BUSCO.

This study incorporated several evaluation metrics specifically for transcriptomes including number of full-length transcripts, ORFs, Detonate's RSEM-EVAL and RMBT. Shown here as well as other studies, the Trinity assembler proved able to reconstruct the most full-length transcripts (Chopra et al., 2014; Grabherr et al., 2011; Duan et al., 2012). Alignment coverages were greater than 70% (8,168) and 90% (5,664) even though it utilized a single *k-mer* value (Figs. 14a and 14b). The Bridger assembler was able to reconstruct the most ORFs (Fig. 16). Results showed that transcripts of >799 bps, >999 bps, and > 1199 bps had 3,189, 1,377 and 593 ORFs, respectively. This number of ORFs was 23.9% (>799 bs), 19.3% (>999 bps), and 16.3% (>1199 bps) higher than those for Trinity, the next closest assembler. SMK and LMK assemblies of Oases and SOAP were similar for both full-length transcripts and ORFs; therefore, the overall quality of each assembly's performance was not reliant on the *k-mer* strategy. Detonate's RSEM-EVAL

provided a reference-free evaluation score where a high value indicated an accurate transcriptome assembly. Results showed that the Trinity assembly appeared slightly better than the Bridger and SOAP LMK (Table 10). However, all scores were similar (ranging from  $5,426.0 \times 10^6$  to  $6,125.0 \times 10^6$ ); and therefore, this metric did not distinguish well between the assemblers. The RMBT statistic determined assembly accuracy based on the philosophy that the higher the amount of processed reads that can be mapped back to an assembly the fewer the errors (i.e. mis-assembly) introduced by the assembler program. Results for the six assembly strategies evaluated showed RMBT percentages ranging from 87.71% for the Bridger assembly to 80.77% for the SOAP SMK assembly (Table 7). Other studies comparing assembly strategies also reported differences in RMBT percentages (Haznedaroglu et al., 2012; Moreton et al., 2014; Chopra et al., 2014; Zhao et al., 2011; Schliesky et al., 2012). In the study presented, the assembler software and not the *k-mer* strategy appeared to have the greater impact. RMBT results showed that the SMK and LMK assemblies for both SOAP (80.77 and 83.78%, respectively) and Oases (84.18 and 85.28%, respectively) performed similarly. In summation, both Trinity and Bridger performed well according to transcriptome-specific metrics and *k-mer* sizes were not a factor.

Overall, the study presented compared assembly programs, *k-mer* strategies, and various metrics for determining *de novo* transcriptome assembly quality. Assembly performance was evaluated using the testis transcriptome of *Fundulus heteroclitus*, an estuary fish that is a sentinel teleost species commonly used in environmental toxicology studies (Burnett et al., 2007). To date, this is the first study to compare the performances

of the commonly used *de Bruijn* graph-based *de novo* assemblers, Trinity, Oases, and SOAPdenovo-Trans, with a new *de novo* method employed by Bridger. This is also the first study to evaluate the effects of using a small and large multiple *k-mer* strategy for the Oases and SOAPdenovo-Trans assemblers within the same study. Based on the eleven evaluation metrics presented above, it was found that the product of those assemblies was more influenced by the assembler itself than the *k-mer* strategy. Bridger performed more often within the top three of each evaluation metric than the *de Bruijn* graph-based programs for the *de novo* transcriptome assembly of the killifish RNA-Seq reads. Therefore, Bridger was chosen as the assembler for all the other transcriptomes based on its performance in this study and the fact that its run time was substantially quicker compared to Trinity.

### 3.3 Validation of Reference Transcriptomes

Having chosen Bridger to assembly transcriptomes, it was then necessary to decide on the approach that would allow comparison of gene expression among the different treatment groups. Incorporating female killifish transcriptomes into the analyses was imperative as it allowed male feminization due to endocrine disruption to be assessed. Therefore, each of the following groups had a male and female representative transcriptome: sexually active exposed (EXP), sexually active control (CON), and non-sexually active (NSA). Two possible approaches for transcriptome assembly were considered. The first was to assemble reads for each sample separately and then compare them. However, this approach has proven extremely complicated by the necessity to

match the same transcripts derived from each of the assemblies (Haas et al., 2013). In addition, the killifish were wild outbred animals expected to have genetic variability between individuals further complicating this approach. The more straightforward alternative was to first combine all reads across all samples into a single data set and assemble them to generate a single reference assembly. The transcripts from each sample could then be quantified by aligning each sample's (not normalized) reads to the reference transcriptome assembly and counting the number of reads that align to each transcript (Haas et al., 2013; Farrell et al., 2014). The quality and accuracy of each transcriptome assembled was determined by statistical analyses.

Transcriptome quality and accuracy was assessed using both *in silico* and *in vitro* approaches. RMBT analysis coupled with BUSCO alignment analysis constituted the two *in silico* approaches employed to validate the accuracy and quality of the reference killifish transcriptome. The percentage of reads that could be mapped back to the transcripts (RMBT) ranged from 96.49% to 98.31% signifying that all the transcriptomes were accurately assembled (Table 11). The results of the BUSCO alignments further strengthen the quality of the reference killifish transcriptome by having less missed genes (563/3,023, 18.6%) compared to the top performer in the *de novo* transcriptome assembly study, Trinity (675/3,023, 22.3%). It should be noted that some of the contigs may be missing because they are short and therefore 1) do not consist of well characterized protein domains, 2) predominately be 3' or 5' untranslated regions, or 3) be non-coding RNAs (Gao et al, 2014). Overall, these two *in silico* assessment tools validated the quality and accuracy of the *de novo* assembly of the reference transcriptome.

Trinotate was used to generate gene expression levels of each male transcriptome. An *in vitro* approach to validating these expression levels included designing primers based off the reference transcriptome and performing qPCR analyses. Ten gene specific primers were designed based on sequences assembled by Bridger and annotated by BLASTx. The primers were used to amplify partial sequences from male killifish cDNA by PCR. The resulting bands were the expected size (Table 3). DNA from the PCRs were cloned, sequenced and aligned to sequences in GenBank to find the percent identity to *F. heteroclitus* as shown in Table 12. Of the ten cloned genes, only PDK1 and TDRD7b did not have a *F. heteroclitus* match. However, these genes still had a high percent identity match to other fish species. Overall, the results showed that all of the amplified DNA fragments aligned with their respective gene with identities of  $\geq 83\%$  , which validated the accuracy of the assembly.

Of the ten genes cloned and sequenced, six were used for qPCR analysis (EGLN, GSTA4, MMP14, AK7, TDRD7b, and PDK.). There were a total of 12 Trinotate expression relationships for the six genes (EXP vs CON and NSA vs CON), and ten of these corresponded with the qPCR analysis (Table 13). The main feasible reason causing the discrepancy between the qPCR data and Trinotate expression data may be associated with individual variance between the fish. Additionally, according to the FPKM calculation, the length of transcripts and incomplete paired-end data could affect the FPKM value (Hsu et al., 2015).

$$FPKM = \frac{\text{mapped specific exon fragments}}{\text{total mapped exon fragments (millions)} \times \text{specific exon length (kB)}}$$



Correlating qPCR results with Trinotate results further validated the transcriptome data. Correlations for the average of all fish used in the qPCR as well as for just those used in the transcriptome can be seen in Fig. 27. Trinotate data was presented as EXP vs CON and NSA vs CON; therefore, a similar relationship was needed for qPCR data. It was decided to represent the data as EXP minus CON and NSA minus CON. Interestingly, some genes showed a positive correlation (GSTA4 and PDK1), while others showed a negative correlation (ENLG, MMP14, and TDRD7b). The positive correlations were not statistically significant with  $p = 0.342$  and  $0.301$  for transcriptome only and all fish, respectively. However, a strong relationship for Trinotate and qPCR was found for those genes showing a negative correlation with  $p = 0.015$  and  $0.037$  for transcriptome only and all fish, respectively. Overall, the correlations do show a good relationship between the qPCR data and Trinotate expression data, especially for the negative correlations.

### *3.4 Bioinformatic Visualization of Potential Biomarkers*

The transcriptome data was visualized using various bioinformatics tools. The influence of sexual activation was visualized by comparing NSA vs CON and the influence of crude oil exposure was visualized by comparing EXP vs CON. The volcano plots in Figs. 18-19 both showed that the majority of the differentially expressed genes were up-regulated during sexual activation and that the crude oil exposure elicited a response at the genetic level. The comparison between the NSA testis vs the CON testis showed that the NSA testis had 139,931 genes down-regulated and 106,249 genes up-regulated (which is to say sexual activation up-regulated 139,931 and down-regulated

106,249 genes). The comparison between the EXP testis vs the CON showed that the EXP testis had 76,931 genes down-regulated and 146,093 genes up-regulated. A study performed by Garcia *et al.*, showed a volcano plot representing *de novo* assembled liver transcriptomes of (*Fundulus grandis*) exposed to the crude oil from the Deep Horizon oil spill in 2010 (2012). Similar complex genetic responses was observed with 1,070 down-regulated and 1,251 up-regulated genes. It should be noted that the Garcia *et al.* study only focused on differentially expressed genes with a significance level of  $p < 0.01$  while the present study did not. Also, the Garcia *et al.* study used different bioinformatics tools for their differential expression analysis (Bowtie for the aligner and DESeq for differential expression where this study employed STAR for the aligner and edgeR for the differential expression analysis). Nevertheless, the present study and that of Garcia *et al.* showed crude oil up-regulated more genes than it down-regulated. In order to discover novel genes involved in endocrine disruption and to investigate their interactions with other genes and biochemical pathways, the data visualization software, Cytoscape, was employed.

The Cytoscape App ClueGo was used to investigate the interaction of the differentially expressed genes with in selected GO Categories (Estrogen, Ovary, Ovulation, Reproduction, Sex, Spermatogenesis, Testis, Steroidogenesis, Xenobio, Hypoxia, and Heat Shock) that had at least a LFC of  $\pm 2$  within the EXP vs CON group. The search term “Androgen” was used to further concentrate down the network. It was found that the “Regulation of Androgen Receptor Signaling Pathway” GO category node (circled in red) found with in the “Steroid Hormone Receptor Signaling Pathway” (black

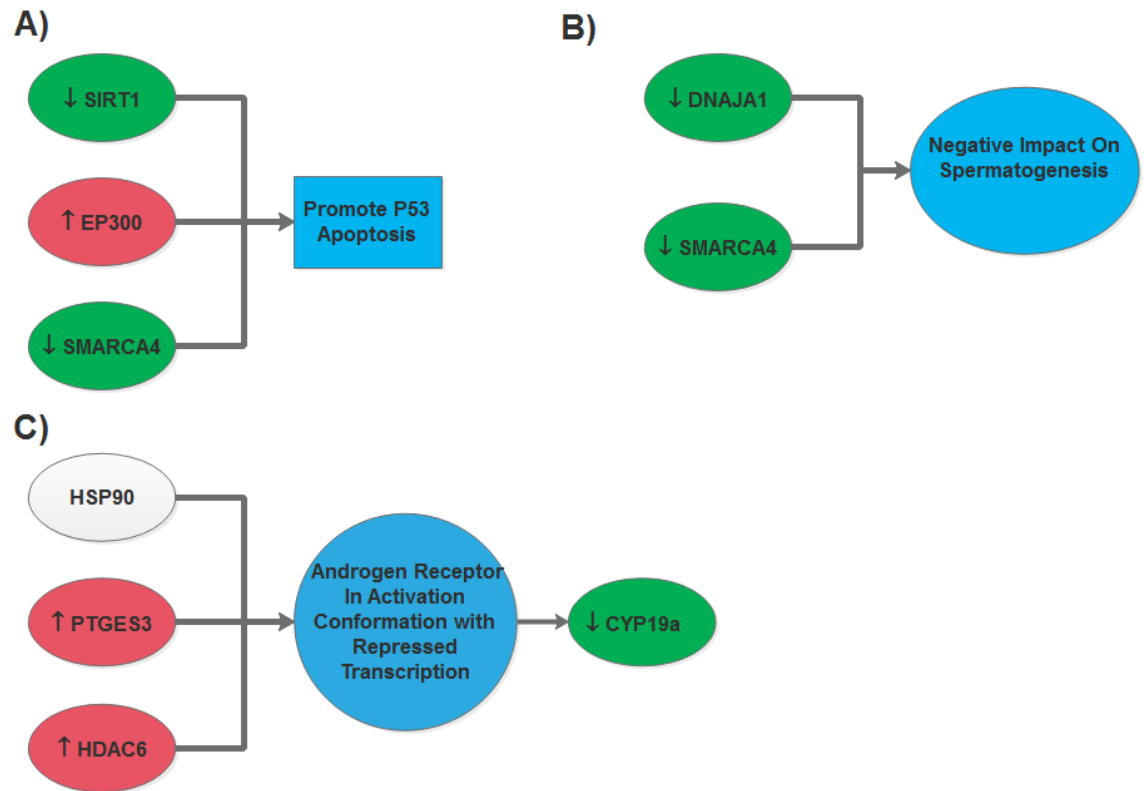
nodes) was interacting with various groups of GO category nodes associated with “Response to Heat” (red nodes) and “Apoptosis” (blue nodes) as shown in the Cytoscape network Fig. 20. The interaction with the “Response to Heat” and “Apoptosis” GO category nodes supported the premise that the crude oil gavaged fish were responding to a stressful condition at both the cellular and molecular level

To further this point, the chaperone nodes found within the “Response to Heat” network contained genes associated with chaperone functions such as stabilizing new proteins for proper folding and refolding proteins that were damaged by cellular stress (De Maio et al, 1999). The presence of “Apoptosis” related GO category nodes also supported stressful conditions in crude oil gavaged fish. It has been shown that PAHs in crude oil can cause DNA damage through the formation of reactive oxygen species (ROS) (Taban et al., 2004). This is evident as various nodes within the “Apoptosis” network were related to responses associated with DNA repair and oxidative stress. Overall, this network shows that the androgen receptor pathway and associated steroid receptor pathways were responding to crude oil exposure both at the DNA and protein levels.

The “Regulation of Androgen Receptor Signaling Pathway” node found within the “Steroid Hormone Receptor Signaling Pathway” network (Fig. 20) consisted of the following four differentially expressed genes within the EXP group: EP300, HDAC6, SIRT1, and SMARCA4. Interestingly, all four of these genes were also found in at least one of the response categories, “Response to Heat”, Fig. 22 or “Apoptosis”, Fig. 21, suggesting that these genes played a key role in the response to crude oil involving the

androgen receptor-signaling pathway. EP300, HDAC6, and SIRT1 were within both the “Response to Heat” and “Apoptosis” groups while SMARCA4 was only within the “Apoptosis” group. EP300 and HDAC6 were up-regulated in the EXP group compared to both the CON and NSA groups signifying that these genes were directly responding to the stress caused by the crude oil exposure. SIRT1 and SMARCA4 had the same down-regulated expression patterns in both of the EXP and NSA groups showing that the sexually active crude oil exposed fish was behaving in the same fashion as the non-sexually active fish.

The down-regulated SIRT1 (Sirtuin 1) and up-regulated EP300 (histone acetyltransferase p300) have interesting relationships with detrimental pathways that might have been affected by the crude oil as seen in Fig. 28A. SIRT1 (Sirtuin 1) has been shown to deacetylate and thereby deactivate p53’s oxidative stress-induced apoptotic activity (Hori et al., 2013). The functional role of the EP300 is to regulate p53-dependent apoptosis after DNA damage (Lyer et al., 2004). Therefore, up-regulation of EP300 due to DNA damaged and the lack of inhibition of p53 activity due to down-regulation of SIRT1 would have increased levels of apoptosis in response to crude oil.



**Fig. 28** Candidate biomarkers. A) Candidates associated with apoptosis promotion (EP300, SIRT1 and SMARCA4). B) Candidates associated with impaired spermatogenesis (SMARCA4 and DNAJA1). C) Candidates responsible for suppressed androgen receptor transcriptional activation (HDAC6 and PTGES3) causing less androgen to be converted into estrogen by CYP19a.

The up-regulated HDAC6 (Histone deacetylase 6) and down-regulated SMARCA4 (Transcription activator BRG1) have interesting relationships with reproductive pathways that might have been affected by the crude oil as seen in Fig. 28. HDAC6 enhances histone deacetylase activity and thereby represses transcription (Aldana-Masangkay et al., 2011). SMARCA4 also represses transcription and research shows that SMARCA4 represses p53-dependent transcription (Lee et al., 2002). Additionally, it has been demonstrated that SMARCA4's activity is crucial to the development of sperm (Kim et al., 2012). Therefore, the down-regulation caused by crude oil exposure would make males less fertile and be in keeping with the low expression levels of SMARCA4 found in the NSA group. Repression of transcription by up-regulated HDAC6 and down-regulated SMARCA4 and their possible influence on sperm development may be a significant finding of this work.

In addition, up-regulated HDAC6 is associated with activation of HSP90 and Prostaglandin E Synthase 3 (PTGES3) as seen in Fig. 28C. (Ai et al, 2009, Chan et al, 2015). PTGES3 is a co-chaperone to heat shock protein 90 that localizes to genomic response elements in a hormone-dependent manner and disrupts receptor-mediated transcriptional activation by promoting disassembly of transcriptional regulatory complexes (Forsythe et al., 2001). HSP90 and PTGES3 play a key role in androgen-induced and androgen-independent nuclear localization as well as androgen receptor (AR) stabilization (Ai et al., 2009, Reebye et al., 2012). Stabilization of the AR is associated with maintaining the receptor in the active conformation while repressing its transcriptional activities, allowing it to bind AR ligands (Ai et al., 2009, Chan et al.,

2015). Therefore, up-regulation of HDAC6, HSP90 and PTGES3 in the present study should result in suppression of CYP19a transcription, which supports the down-regulation of CYP19a observed in PCR analysis of Day 7 exposed males (Figs. 3 and 9). Overall, these four key genes within the “Regulation of the Androgen Receptor Signaling Pathway” were responding to the crude oil making them great candidate crude oil biomarkers that can link molecular responses to reproductive effects.

The genes highlighted within the concentrated volcano plots (Fig. 24) are other candidate biomarkers to link genetic responses to the crude oil exposure. These genes were selected based on meeting one of the following three criteria. First, they appeared in more than one heatmap. The involvement of a gene in more than one category gives the impression that they are one of the main genes either positively or negatively impacted by the crude oil. Second, gene expression patterns had a major influence in the selection of candidate biomarkers. Genes in the EXP group with same expression patterns as the NSA group were selected because the results implied that the sexually active fish in the exposed group were endocrine disrupted to the point of responding like non-sexually active fish. Third, in keeping with gene expression patterns, genes were selected in the EXP group that had the opposite response in the NSA group.

The following genes appeared in all three heatmaps constituting the Steroid Receptor Signaling Pathway in Cytoscape (Fig. 20): PTGES3, EP300, SIRT1, HDAC6, and DNAJA1. The responses and consequences of PTGES3, EP300, SIRT1 and HDAC6 are discussed above. DNAJA1, heat shock protein family (Hsp40) member A1, acts as a heat shock protein 70 co-chaperone by facilitating protein folding, trafficking, prevention

of aggregation, and proteolytic degradation (Terada et al., 2005). Terada et al. has shown that DNAJA1 mutant mice led to severe defects in spermatogenesis that involved aberrant androgen signaling (Terada et al., 2005). In this study, this gene was down-regulated in EXP compared to CON indicating possible detrimental impacts on spermatogenesis by crude oil. This response was supported by the observed down-regulation of SMARCA4, which is known to have a similar effect. Overall, the concentrated volcano plots showed filtered candidate biomarkers that can be used in future studies.

## **Conclusion**

To date, this research is the first report of an annotated overview for the testes transcriptome in *F. heteroclitus*, resulting in the most comprehensive genetic reproductive resource available for the species. In this study, the global expression patterns in response to crude oil exposure in killifish were profiled to shed light on the complex genetic responses to crude oil exposure. This research provided the groundwork for modern, relevant tools for studying the effects of crude oil on killifish populations by linking molecular biomarkers and impaired reproduction. As a result, candidate gene modules and biomarkers responsible for gonadal-derived reproductive disruption at the functional genomic level were identified: PTGES3, EP300, SIRT1, HDAC6, and DNAJA1. This work can provide a repository for future gene expression analysis, functional studies, and reproductive investigations in *F. heteroclitus*. This will enhance the capabilities of population monitoring and can be used as a benchmark in comparative studies in other fish models. This data will be of interest to researchers using the killifish



in estuaries as bio-monitors to crude oil exposure. Overall, this research will open new opportunities and bring new insights for researchers studying *F. heteroclitus*.

## References

- Aas, E., Beyer, J., Goksøyr, A. 2000b. Fixed wavelength fluorescence (FF) of bile as a monitoring tool for polyaromatic hydrocarbon exposure in fish: an evaluation of compound specificity, inner filter effect and signal interpretation. *Biomarkers* **5**, 9e23.
- Ai, J., Wang, Y., Dar, J. A., Liu, J., Liu, L., Nelson, J. B., & Wang, Z. (2009). HDAC6 Regulates Androgen Receptor Hypersensitivity and Nuclear Localization via Modulating Hsp90 Acetylation in Castration-Resistant Prostate Cancer. *Molecular Endocrinology*, **23** (12), 1963–1972. doi.org/10.1210/me.2009 0188
- Aldana-Masangkay, G., Sakamoto, K. (2011) The Role of HDAC6 in Cancer. *Journal of Biomedicine and Biotechnology*:875824. doi:10.1155/2011/875824.
- Anders, S., McCarthy, D., Chen, Y., Okoniewski, M., Smyth, G., Huber, W., et al., (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols* **8**, 1765–1786 (2013) doi:10.1038/nprot.2013.099
- Anderson, M., Miller, M., Hinton, D. (1996) In vitro modulation of 17- $\beta$ -estradiol-induced vitellogenin synthesis: Effects of cytochrome P4501A1 inducing compounds on rainbow trout (*Oncorhynchus mykiss*) liver cells. *Aquatic Toxicology* **34**, 327-350
- Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Ariese, F., Kok, S.J., Verkaik, M., Gooijer, C., Velthorst, N.H., Hofstra, J.W. (1993) Synchronous fluorescence spectrometry of fish bile: a rapid screening method for the biomonitoring of PAH exposure. *Aquat. Toxicol.* **26**, 273e286.
- Ashrafi, H., Hill, T., Stoffel, K., Kozik, A., Yao, J., Chin-Wo, S.R, et al. (2012) *De novo* assembly of the pepper transcriptome (*Capsicum annuum*): a benchmark for *in silico* discovery of SNPs, SSRs and candidate genes. *BMC Genomics*. **13**:571 doi:10.1186/1471-2164-13-571
- Baker M. (2012) *De novo* genome assembly: what every biologist should know. *Nat Meth.* **9**(4):333–7. doi: 10.1038/nmeth.1935.
- Bemanian, V., Male, R., Goksøyr, A. (2004) The aryl hydrocarbon receptor-mediated disruption of vitellogenin synthesis in the fish liver: Cross-talk between AHR- and ER $\alpha$ -signalling pathways. *Comparative Hepatology* **3**:2

- Bentivegna, C.S., DeFelice, C.R., Murphy, W.R.** (2016) Excitation–emission matrix scan analysis of raw fish oil from coastal New Jersey menhaden collected before and after Hurricane Sandy, Marine Pollution Bulletin. doi.org/10.1016/j.marpolbul.2016.01.023
- Bindea, G., Mlecnik, B., Hacki, H., Charoentong, P., Tosolini, M, Kirilovsky, A., Fridman, W., Pages, F., Trajanoski, Z., Galon, J,** (2009) Bioinformatics. Apr 15;25(8):1091-3. doi: 10.1093/bioinformatics/btp101.
- Bodkin, J.L., Ballachey, B.E., Dean, T.A., Fukuyama, A.K., Jewett, S.C., McDonald, L., Monson, D.H., O'Clair, C.E., VanBlaricom, G.R.** (2002) Sea otter population status and the process of recovery from the 1989 'Exxon Valdez' oil spill. Marine Ecology Progress Series **241**, 237-253.
- Bolger, A.M., Lohse, M., & Usadel, B.** (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics. **30** (15): 2114–2120. doi: 10.1093/bioinformatics/btu170
- Burnett, K.G., Bain, L.J., Baldwin, W.S., Callard, G.V., Cohen. S., et al.** *Fundulus* as the premier teleost model in environmental biology: opportunities for new insights using genomics. Comparative Biochemistry and Physiology Part D, Genomics & Proteomics. **2**: 257–286.
- Breyne, P., Dreesen, R., Cannoot, B., Rombaut, D., Vandepoele, K., Rombauts, S., Vanderhaeghen, R., Inzé, D., Zabeau, M.** (2003) Quantitative cDNA-AFLP analysis for genome-wide expression studies. Molecular Genetics and Genomics **269**, 173-179.
- Brulle, F., Jeffroy, F., Madec, S., Nicolas, J.L., Paillard, C.** (2012) Transcriptomic analysis of *Ruditapes philippinarum* hemocytes reveals cytoskeleton disruption after in vitro *Vibrio tapetis* challenge. Developmental & Comparative Immunology. Volume **38**, Issue 2, Pages 368–376
- Bugel, S.M., White, L.A., Cooper, K.R.** (2010) Impaired reproductive health of killifish (*Fundulus heteroclitus*) inhabiting Newark Bay, NJ, a chronically contaminated estuary. Aquatic Toxicology. **96** 182-193
- Burnett, K.G., Bain, L.G., Baldwin, W.S., Callard, G.V., Cohen, C., Di Giulio, R.T., et al.** (2007) *Fundulus* as the premier teleost model in environmental biology: opportunitites for new insights using genomics. Comp. Biochem. Physiol. Part D: Genomic Proteomic **2** (4), 257-286
- Carls, M.G., Rice, S.D., Hose, J.E.** (1999) Sensitivity of fish embryos to weathered crude oil: Part I. Low-level exposure during incubation causes malformations,

- genetic damage, and mortality in larval pacific herring (*Clupea pallasii*). Environmental Toxicology and Chemistry **18**, 481-493.
- Chan, S.C., Dehm, S.M.** (2014) Constitutive activity of the androgen receptor. Advances in pharmacology;**70**:327–366
- Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., et al.** (2015) Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biology. **16**:30 doi: 10.1186/s13059-015-0596-2
- Cheshenko, K., Pakdel, F., Segner, H., Kah, O., Eggen, R.** (2008) Interference of endocrine disrupting chemicals with aromatase CYP19 expression or activity, and consequences for reproduction of teleost fish. General and Comparative Endocrinology **155** 31–62
- Chopra, R., Burow, G., Farmer, A., Mudge, J., Simpson, C.E., et al.** (2014) Comparisons of De Novo Transcriptome Assemblers in Diploid and Polyploid Species Using Peanut (*Arachis spp.*) RNA-Seq Data. PLoS ONE. **9** (12): e115055. doi:10.1371/journal.pone.0115055
- Cloonan, N., Grimmond, S.M.** (2008) Transcriptome content and dynamics at single-nucleotide resolution. Genome biology **9**, 234.
- Davail B, Pakdel F, Bujo H, Perazzolo L, Wacławek M, Schneider W, Le Menn F.** (1998) Evolution of oogenesis: The receptor for vitellogenin from the rainbow trout. J Lipid Res **39**:1929-1937.
- Davidovici, B., Orion, E., Wolf, R.,** (2008) Cutaneous manifestations of pituitary gland diseases Clinics in Dermatology. Volume **26**, Issue 3, Pages 288–295
- Debouck, C.,** (1995) Differential display or differential dismay? Current Opinion in Biotechnology **6**, 597-599.
- De Maio, A.** (1999) Heat shock proteins: facts, thoughts, and dreams Shock. **11** (1): 1–12. doi:10.1097/00024382-199901000-00001. PMID 9921710.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S. et al.** (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics. **29**:15–21
- Duan, J., Xia, C., Zhao, G., Jia, J., Kong, X.** (2012) Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genomics. **13**: 392. doi: 10.1186/1471-2164-13-392
- Dubansky, B., Whitehead, A., Miller, J., Rice, C., Galvez, F.** (2013) Multi-tissue molecular, genomic, and developmental effects of the Deepwater Horizon oil spill on resident Gulf killifish (*Fundulus grandis*). Environ. Sci. Technol.DOI: 10.1021

- Elcoro, A.S., de Juan, A., García, J.A., Durana, N., Alonso, L.** (2014) Comparison of second order multivariate methods for screening and determination of PAHs by total fluorescence spectroscopy. *Chemom. Intell. Lab. Syst.* **132**, 63–74.
- Esler, D., Schmutz, J.A., Jarvis, R.L., Mulcahy, D.M.** (2000) Winter Survival of Adult Female Harlequin Ducks in Relation to History of Contamination by the "Exxon Valdez" Oil Spill. *The Journal of wildlife management*, 839-847.
- Esler, D., Bowman, T.D., Trust, K.A., Ballachey, B.E., Dean, T.A., Jewett, S.C. et al.** (2002) Harlequin duck population recovery following the 'Exxon Valdez' oil spill: Progress, process and constraints. *Marine Ecology Progress Series* **241**, 271-286.
- Farrell, J.D., Byrne, S., Paina, C., Asp, T.,** (2014) *De Novo* Assembly of the Perennial Ryegrass Transcriptome Using an RNA-Seq Strategy. *PLoS ONE* **9**(8): e103567. doi:10.1371/journal.pone.0103567
- Fan, H., Xiao, Y., Yang, Y., Xia, W., Mason, A.S., et al.** (2013) RNA-Seq Analysis of *Cocos nucifera*: Transcriptome Sequencing and Subsequent Functional Genomics Approaches. *PLoS ONE*; **8**(3): e59997. doi:10.1371/journal.pone.0059997
- Feldmeyer, B., Wheat, C., Krezdorn, N., Rotter, B., Pfenninger, M.** (2011) Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, *Basommatophora*, *Pulmonata*), and a comparison of assembler performance. *BMC genomics* **12**, 317.
- Ferretto, N., Tedetti, M., Guigue, C., Mounier, S., Redon, R., Goutx, M.** (2014) Identification and quantification of known polycyclic aromatic hydrocarbons and pesticides in complex mixtures using fluorescence excitation–emission matrices and parallel factor analysis. *Chemosphere* **107**, 344–353.
- Forsythe, H.L., Jarvis, J.L., Turner, J.W., Elmore, L.W., Holt, S.E.** (2001) Stable Association of hsp90 and p23, but Not hsp70, with Active Human Telomerase. *The Journal of Biological Chemistry* **276**, 15571-15574. doi: 10.1074/jbc.C100055200
- Fuentes-Rios, D., Orrego, R., Rudolph, A., Mendoza, G., Gavilan, J.F., Barra, R.** (2005) EROD activity and biliary fluorescence in *Schroederichthys chilensis* (Guichenot 1848): biomarkers of PAH exposure in coastal environments of the South Pacific Ocean. *Chemosphere* **61**, 192e199.
- Gao, J., Wang, X., Zou, Z., Jia, X., Wang, Y., Zhang, Z.** (2014) Transcriptome

- analysis of the differences in gene expression between testis and ovary in green mud crab (*Scylla paramamosain*). BMC Genomics. **15**:585 doi:10.1186/1471-2164-15-585
- Garg, R., Patel, R.K., Tyagi, A.K., Jain, M.** (2011) *De Novo* Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification. DNA Research. **18** (1): 53-63. doi: 10.1093/dnares/dsq028
- Garcia, O., Saveanu, C., Cline, M., Fromont-Racine, M., Jacquier, A., Schwikowski, B., Aittokallio, T.** (2007) Golorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring Bioinformatics. **23** (3): 394-396. doi: 10.1093/bioinformatics/btl605
- Garcia, T., Shen, Y., Crawford, D., Oleksiak, M., Whitehead, A., Walter, R.** (2012) RNA-Seq reveals complex genetic response to deepwater horizon oil release in *Fundulus grandis*. BMC Genomics. **13**:474.
- Golet, G.H., Seiser, P.E., McGuire, A.D., Roby, D.D., Fischer, J.B., Kuletz, K.J., Irons, D.B., Dean, T.A., Jewett, S.C., Newman, S.H.** (2002) Long-term direct and indirect effects of the 'Exxon Valdez' oil spill on pigeon guillemots in Prince William Sound, Alaska. Marine Ecology Progress Series **241**, 287-304.
- Gordo, S.M., Pinheiro, D.G., Moreira, E.C., Rodrigues, S.M., Poltronieri, M.C., de Lemos, O.F., et al.** (2012) High-throughput sequencing of black pepper root transcriptome. BMC Plant Biology. **12**:168 doi:10.1186/1471-2229-12-168
- Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., et al.** (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. **29**: 644–652. doi: 10.1038/nbt.1883. pmid:21572440
- Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., et al.** (2010) The developmental transcriptome of *Drosophila melanogaster*. Nature **471**, 473-479.
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., et al.** (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols **8**: 1494–1512.
- Haznedaroglu, B.Z., Reeves, D., Rismani-Yazdi, H., Peccia, J.** (2012) Optimization of *de novo* transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. BMC Bioinformatics **13**:170 doi:10.1186/1471-2105-13-170

- He, B., Zhao, S., Chen, Y., Cao, Q., Wei, C., Cheng, X., et al.** (2015) Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms. *BMC Genomics*. **16**:65 doi:10.1186/s12864-014-1192
- Hori, Y.S., Kuno, A., Hosoda, R., Horio, Y.** (2013) Regulation of FOXOs and p53 by SIRT1 Modulators under Oxidative Stress. *PLoS ONE* **8**(9): e73875. doi:10.1371/journal.pone.0073875
- Hose, J. E., McGurk, M. D., Marty, G. D., Hinton, D. E., Brown, E. D., Baker, T.** (1996) Sublethal effects of the Exxon Valdez oil spill on herring embryos and larvae: morphological, cytogenetic, and histopathological assessments, 1989-1991 : Effects of the Exxon Valdez oil spill on young Pacific herring in Prince William Sound, Alaska. *Canadian journal of fisheries and aquatic sciences* (Print) **53**, 2355-2365.
- Hsu, H.Y., Chen, S.H., Cha, Y.R., Tsukamoto, K., Lin, C.Y., Han, Y.S.,** (2015) *De Novo* Assembly of the Whole Transcriptome of the Wild Embryo, *Preleptocephalus*, *Leptocephalus*, and Glass Eel of *Anguilla japonica* and Deciphering the Digestive and Absorptive Capacities during Early Development. *PLoS ONE* **10**(9): e0139105. doi:10.1371/journal.pone.0139105
- Huth, T.J., Place, S.P.** (2013) *De novo* assembly and characterization of tissue specific transcriptomes in the emerald notothen, *Trematomus bernacchii*. *BMC Genomics*. **14**:805 doi:10.1186/1471-2164-14-805
- Idler, D., So, Y., Fletcher, G., Payne, J.** (1995) Depression of blood levels of reproductive steroids and glucuronides in male winter flounder (*Pleuronectes americanus*) exposed to small quantities of Hibernia crude, used crankcase oil, oily drilling mud and harbour sediments in the 4 months prior to spawning in late May-June, 1995.
- Incardona, J.P., Collier, T.K., Scholz, N.L.** (2004) Defects in cardiac function precede morphological abnormalities in fish embryos exposed to polycyclic aromatic hydrocarbons. *Toxicology and Applied Pharmacology* **196**, 191-205.
- Inoue, A., Yoshida, N., Omoto, Y., et al.** (2002) Development of cDNA microarray for expression profiling of estrogen-responsive genes. *Journal of Molecular Endocrinology*. **29**(2):175–192.
- Irvine, G.V., Mann, D.H., Short, J.W.** (2006) Persistence of 10-year old Exxon Valdez oil on Gulf of Alaska beaches: The importance of boulder-armoring. *Marine pollution bulletin* **52**, 1011-1022.

- Ji, P., Liu, G., Xu, J., Wang, X., Li, J., et al.** (2012) Characterization of Common Carp Transcriptome: Sequencing, *De Novo* Assembly, Annotation and Comparative Genomics. *PLoS ONE*. **7**(4): e35152. doi:10.1371/journal.pone.0035152
- Jung, J., Kim, M., Yim, U., Ha, S., An, J., Won, J., et al.** (2011) Biomarker responses in pelagic and benthic fish over 1 year following the Hebei Spirit oil spill (Taean, Korea). *Mar. Pollut. Bull.* **62**, 1859–1866.
- Kim, Y., Fedoriw, A. M., Magnuson, T.** (2012) An essential role for a mammalian SWI/SNF chromatin-remodeling complex during male meiosis. *Development* (Cambridge, England), **139**(6), 1133–1140.
- Kreitsberg, R., Zemit, I., Freiberg, R., Tambets, M., Tuvikene, A.** (2010) Responses of metabolic pathways to polycyclic aromatic compounds in flounder following oil spill in the Baltic Sea near the Estonian Coast. *Aquatic Toxicology* **99**:473-478 doi: 10.1016/j.aquatox.2010.06.005.
- Laffon, B., Rábade, T., Pásaro, E., Méndez, J.** (2006) Monitoring of the impact of Prestige oil spill on *Mytilus galloprovincialis* from Galician coast. *Environment International* **32**, 342-348.
- Langmead, B., Salzberg, S.L.** (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*. **9**(4):357–9. doi: 10.1038/nmeth.1923.
- Letunic, I., Bork, P.** (2011) Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucl Acids Res.* **39**: 475–478
- Lee, D., Kim, J.W., Seo, T., Hwang, S.G., Choi, E.J., Choe, J.** (2002) SWI/SNF complex interacts with tumor suppressor p53 and is necessary for the activation of p53-mediated transcription. *J Biol Chem* **277**, 22330-22337
- Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J.A., Stewart, R., Dewey, C.** (2014) Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. *Genome Biology*. **15**:553 doi: 10.1186/s13059-014-0553-5
- Li, W., Godzik, A.** (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**: 1658–1659.
- Liang, P., Pardee, A.B.,** (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967-971.
- Lin, E.L.C., Cormier, S.M., Racine, R.N.** (1994) Synchronous fluorometric measurement of metabolites of polycyclic aromatic hydrocarbons in the bile of brown bullhead. *Environ. Toxicol. Chem.* **13**, 707e715.
- Lyer, N.G., Chin, S.F., Ozdag, H., Yataro Daigo, Hu, D.E., Cariati M., et al.** (2004)



- p300 regulates p53-dependent apoptosis after DNA damage in colorectal cancer cells by modulation of PUMA/p21 levels 7386–7391 PNAS vol. **101** no. 19  
doi/10.1073/pnas.0401002101
- Nandety, R.S., Kamita, S.G., Hammock, B.D., Falk, B.W.** (2013) Sequencing and De Novo Assembly of the Transcriptome of the Glassy-Winged Sharpshooter (*Homalodisca vitripennis*). PLoS ONE. 2013; **8**(12): e81681.  
doi:10.1371/journal.pone.0081681
- Norcross, B. L., Hose, J. E., Frandsen, M., Brown E. D.** (1996) Distribution, abundance, morphological condition, and cytogenetic abnormalities of larval herring in Prince William Sound, Alaska, following the Exxon Valdez oil spill: Effects of the Exxon Valdez oil spill on young Pacific herring in Prince William Sound, Alaska. Canadian journal of fisheries and aquatic sciences (Print) **53**, 2376-2387.
- Mateos, J., Mananos, E., Carrillo, M., Zanuy, S.** (2002) Regulation of follicle-stimulating hormone (FSH) and luteinizing hormone (LH) gene expression by gonadotropin-releasing hormone (GnRH) and sexual steroids in the Mediterranean Sea bass . Comparative Biochemistry and Physiology Part B **132** 75–86
- Matozzo, V., Gagné, F., Marin, M., Ricciardi, F., Blaise, C.** (2008) Vitellogenin as a biomarker of exposure to estrogenic compounds in aquatic invertebrates. Environment International **34** 531–545
- Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y.** (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research **18**, 1509-1517.
- McClelland, M., Mathieu-Daude, F., Welsh, J.** (1995) RNA fingerprinting and differential display using arbitrarily primed PCR. Trends in Genetics **11**, 242-246.
- McGurk, M. D., Brown, E. D.** (1996) Egg-larval mortality of Pacific herring in Prince William Sound, Alaska, after the Exxon Valdez oil spill: Effects of the Exxon Valdez oil spill on young Pacific herring in Prince William Sound, Alaska. Canadian journal of fisheries and aquatic sciences (Print) **53**, 2343-2354.
- Meador, J., Buzitis, J., Bravo, C.** (2008) Using fluorescent aromatic compounds in bile from juvenile salmonids to predict exposure to polycyclic aromatic hydrocarbons. Environmental Toxicology and Chemistry, Vol. **27**, No. 4, pp. 845–853
- Metzker, M.** (2010) Sequencing technologies — the next generation Nature Reviews Genetics **11**, 31-46

- Meyer, J., Nacci, D., Di Giulio, R.** (2002) Cytochrome P4501A (CYP1A) in Killifish (*Fundulus heteroclitus*): Heritability of Altered Expression and Relationship to Survival in Contaminated Sediments. *Toxicological Sciences* **68**, 69–81
- Mills, L., Chichester, C.** (2005) Review of evidence: Are endocrine-disrupting chemicals in the aquatic environment impacting fish populations? *Science of the Total Environment* **343** 1– 34
- Morales-Caselles, C., Jiménez-Tenorio, N., de Canales, M., Sarasquete, C., DelValls, T.** (2006) Ecotoxicity of Sediments Contaminated by the Oil Spill Associated with the Tanker “Prestige” Using Juveniles of the fish *Sparus aurata*. *Archives of Environmental Contamination and Toxicology* **51**, 652-660.
- Morin, R.D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T.J., et al.** (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**, 81-94.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B.** (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628.
- Moreton, J., Dunham, S., Emes, R.** (2014) A consensus approach to vertebrate *de novo* transcriptome assembly from RNA-seq data: assembly of the duck (*Anas platyrhynchos*) transcriptome. *Front Genet.* **5**:190 doi: 10.3389/fgene.2014.00190
- Morozova, O., Marra, M.A.** (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255-264.
- Monson, D.H., Doak, D.F., Ballachey, B.E., Johnson, A., Bodkin, J.L.** (2000) Long-term impacts of the Exxon Valdez oil spill on sea otters, assessed through age-dependent mortality patterns. *Proceedings of the National Academy of Sciences* **97**, 6562-6567.
- Murawski, S.A., Hogarth, W.T., Peebles, E.B., Barbeiri, L.** (2014) Prevalence of external skin lesions and polycyclic aromatic hydrocarbon concentrations in Gulf of Mexico fishes, post-deepwater horizon. *Trans. Am. Fish. Soc.* **143**, 1084e1097.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M.,** (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349.
- Nakasugi, K., Crowhurst, R., Bally, J., Waterhouse, P.** (2014) Combining Transcriptome Assemblies from Multiple *De Novo* Assemblers in the Allo-Tetraploid Plant *Nicotiana benthamiana*. *PLoS ONE.* **9**(3): e91776. doi:10.1371/journal.pone.0091776

- Navas, J., Segner, H.** (2000) Antiestrogenicity of [beta]-naphthoflavone and PAHs in cultured rainbow trout hepatocytes: Evidence for a role of the aryl hydrocarbon receptor. *Aquatic Toxicology* **51**:79-92.
- Nicholas, J.** (1999) Vitellogenesis in fish and the effects of polycyclic aromatic hydrocarbon contaminants. *Aquatic Toxicology* **45** 77–90
- O'Neil, S.T., Emrich, S.J.** (2013) Assessing *De Novo* transcriptome assembly metrics for consistency and utility. *BMC Genomics*. **14**: 465 doi:10.1186/1471-2164-14-465.
- Ordás, M., Albaigés, J., Bayona, J., Ordás, A., Figueras, A.** (2007) Assessment of In Vivo Effects of the Prestige Fuel Oil Spill on the Mediterranean Mussel Immune System. *Archives of Environmental Contamination and Toxicology* **52**, 200-206.
- Ozsolak, F., Milos, P.M.** (2010) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* **12**, 87-98.
- Pait, A., Nelson, J.** (2003) Vitellogenesis in male *Fundulus heteroclitus* (killifish) induced by selected estrogenic compounds. *Aquatic Toxicology* **64** 331/342
- Patel, M., Scheffler, B., Wang, L., Willett, K.** (2006) Effects of benzo(a)pyrene exposure on killifish (*Fundulus heteroclitus*) aromatase activities and mRNA. *Aquatic Toxicology* **77**(3): 267-278
- Parra, G., Bradnam, K., Korf, I.** (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. **23**: 1061–1067.
- Peterson, C.H.** (2001) The “Exxon Valdez” oil spill in Alaska: Acute, indirect and chronic effects on the ecosystem, *Advances in Marine Biology*. Academic Press, 1-103.
- Peterson, C.H., Rice, S.D., Short, J.W., Esler, D., Bodkin, J.L., Ballachey, B.E., Irons, D.B.** (2003) Long-Term Ecosystem Response to the Exxon Valdez Oil Spill. *Science* **302**, 2082-2086.
- Popa, S., Clifton, D., Steiner, R.** (2008) The Role of Kisspeptins and GPR54 in the Neuroendocrine Regulation of Reproduction. *Annu. Rev. Physiol.* **70**:213–38
- Reebye, V., Cano, L., Lavery D. et al.** (2012) Role of the HSP90-associated cochaperone p23 in enhancing activity of the androgen receptor and significance for prostate cancer. *Molecular Endocrinology*, vol. **26**, no. 10, pp. 1694–1706
- Rey-Salgueiro, L., Martínez-Carballo, E., García-Falcón, M.S., González-Barreiro, C., Simal-Gándara, J.** (2009) Occurrence of polycyclic aromatic hydrocarbons and their Hydroxylated metabolites in infant foods. *Food Chem.* **115**, 814–819.

- Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D. et al.** (2010) *De novo* assembly and analysis of RNA-seq data. *Nat Methods* **7**: 909–912. doi: 10.1038/nmeth.1517.
- Robbins, W.K., Hsu, C.S.** (1996) *Kirk-Othmer Encyclopedia of chemical Technology*, vol. **18**, 4<sup>th</sup> ed., J. Wiley, pp. 352-370.
- Salisbury, J.P., Sirbulescu, R.F., Moran, B.M., Auclair, J.R., Zupanc, G.K.H., Agar J.** (2015) The central nervous system transcriptome of the weakly electric brown ghost knifefish (*Apteronotus leptorhynchus*): *de novo* assembly, annotation, and proteomics validation. *BMC Genomics*. **16**:166 doi:10.1186/s12864-015-1354-2
- Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., et al.** (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Research*. **22**: 557–567. doi: 10.1101/gr.131383.111
- Schliesky, S., Gowik, U., Weber, A.P.M., Braeutigam, A.** (2012) A: RNA-Seq assembly—Are we there yet?. *Front Plant Sci*. **3**:220.
- Schmieder, R., Edwards, R.** (2011) Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS ONE*. **6**(3): e17288. doi:10.1371/journal.pone.0017288
- Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E.** (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*. **28**: 1086–1092. doi: 10.1093/bioinformatics/bts094.
- Seiser, P.E., Duffy, L.K., McGuire, D.A., Roby, D.D., Golet, G.H., Litzow, M.A.** (2000) Comparison of Pigeon Guillemot, *Cephus columba*, Blood Parameters from Oiled and Unoiled Areas of Alaska Eight Years After the Exxon Valdez Oil Spill. *Marine pollution bulletin* **40**, 152-164.
- Sharma, N., Jung, C.H., Bhalla, P.L., Singh, M.B.** (2014) RNA Sequencing Analysis of the Gametophyte Transcriptome from the Liverwort, *Marchantia polymorpha*. *PLoS ONE*. **9**(5): e97497. doi:10.1371/journal.pone.0097497
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al.** (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*. Nov; **13**(11):2498-504
- Shen, B., Zhang, Z., Wang, Y., Wang, G., Chen, Y., Lin, P., et al.** (2009) Differential expression of ubiquitin-conjugating enzyme E2r in the developing ovary and

- testis of penaeid shrimp *Marsupenaeus japonicus*. Molecular biology reports **36**, 1149-1157.
- Shen, Y., Garcia, T., Walter, R.** (2011) Gene Expression Analysis Using RNA-Seq from Organisms Lacking Substantial Genomic Resources, Systems and Computational Biology - Molecular and Cellular Experimental Systems, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-280-7, InTech, DOI: 10.5772/19790
- Shendure, J., Ji, H.,** (2008) Next-generation DNA sequencing. Nature biotechnology **26**, 1135-1145.
- Short, J.W., Lindeberg, M.R., Harris, P.M., Maselko, J.M., Pella, J.J., Rice, S.D.** (2003) Estimate of Oil Persisting on the Beaches of Prince William Sound 12 Years after the Exxon Valdez Oil Spill. Environmental Science & Technology **38**, 19-25.
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.V.** (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. doi: 10.1093/bioinformatics/btv351
- Schulz, R.W., Vischer, H.F., Cavaco, J.E., Santos, E.M., Tyler, C.R., Goos, H.J., et al.** (2001) Gonadotropins, their receptors, and the regulation of testicular functions in fish. Comp Biochem Physiol B Biochem Mol Biol **129**:407– 417
- Sonnenschein, C., Soto, A.** (1998) An Updated Review of Environmental Estrogen and Androgen Mimics and Antagonists. J. Steroid Biochem. Molec. Biol., Vol. **65**, No. 1-6, pp. 143-150
- Surget-Groba, Y., Montoya-Burgos, JI.** (2010) Optimization of *de novo* transcriptome assembly from next-generation sequencing data. Genome Res. **20**(10):1432-40. doi: 10.1101/gr.103846.109
- Taban, I.C., Bechmann, R.K., Torgrimsen, S., Baussant, T., Sanni, S.** (2004) Detection of DNA damage in mussels and sea urchins exposed to crude oil using comet assay. Marine Environmental Research Volume **58**, Issues 2–5, Pages 701–705
- Tena-Sempere, M.** (2006) KiSS-1 and reproduction: focus on its role in the metabolic regulation of fertility. Neuroendocrinology **83**, 275–281.
- Terada, K., Yomogida, K., Imai, T., Kiyonari, H., Takeda, N., Kadomatsu, T Mori, M.** (2005) A type I DnaJ homolog, DjA1, regulates androgen receptor signaling and spermatogenesis. The EMBO Journal, **24**(3), 611–622. <http://doi.org/10.1038/sj.emboj.7600549>
- Teo, S.L.H., Able, K.W.** (2003) Habitat use and movement of the mummichog (*Fundulus heteroclitus*) in a restored salt marsh. Estuaries **26**:720-730.

- Thomas, P., Budiantara, L.** (1995) Reproductive life history stages sensitive to oil and naphthalene in Atlantic croaker. *Marine Environmental Research* **39**, 147-150.
- Tilseth, S., Solberg, T.S., Westrheim, K.** (1984) Sublethal effects of the water-soluble fraction of ekofisk crude oil on the early larval stages of cod (*Gadus morhua*). *Marine Environmental Research* **11**, 1-16.
- Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J. et al.** (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* **28**:511-5.
- Truscott, B., Idler, D.R., Fletcher, G.L.** (1992) Alteration of Reproductive Steroids of Male Winter Flounder (*Pleuronectes americanus*) Chronically Exposed to Low Levels of Crude Oil in Sediments. *Canadian Journal of Fisheries and Aquatic Sciences* **49**, 2190-2195.
- Trust, K., Esler, D., R. Woodin, B., J. Stegeman, J.** (2000) Cytochrome P450 1A Induction in Sea Ducks Inhabiting Nearshore Areas of Prince William Sound, Alaska. *Marine pollution bulletin* **40**, 397-403.
- Tyler, C., Jobling, S., Sumpter, J.** (1998) Endocrine disruption in wildlife: a critical review of the evidence. *Crit Rev Toxicol* **28**:319– 61.
- Van Der Kraack, G., Hewitt, M., Lister, A., McMaster, M., Munkittrick, K.** (2001) Endocrine toxicants and reproductive success in fish. *Hum Ecol Risk Assess* **7**:1017– 25.
- Varanasi, U., Reichert, W.L., Eberhart, B.T.L., Stein, J.E.** (1989) Formation and persistence of benzo(a)pyrene-diolepoxide-DNA adducts in liver of English sole (*Parophrys vetulus*). *Chem. Biol. Interact.* **69**, 203e216.
- Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W.** (1995) Serial analysis of gene expression. *Science* **270**, 484-487.
- Wang, X.W., Luan, J.B., Li, J.M., Bao, Y.Y., Zhang, C.X., Liu, S.S.** (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC genomics* **11**, 400.
- Wang, Z., Gerstein, M., Snyder, M.** (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63.
- Whitehead, A., Dubansky, B., Bodinier, C., Garcia, T.I., Miles, S., Pilley, C., et al.** (2011) Genomic and physiological footprint of the Deepwater Horizon oil spill on resident marsh fishes. *Proc. Natl Acad. Sci. USA* **109**, 20298–20302

- Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., et al.** (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243.
- Xiang, L., He, D., Dong, W., Zhang, Y., Shao, J.,** (2010) Deep sequencing-based transcriptome profiling analysis of bacteria-challenged *Lateolabrax japonicus* reveals insight into the immune-relevant genes in marine fish. *BMC genomics* **11**, 472.
- Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J. et al.** (2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*. **30**: 1660–1666. doi: 10.1093/bioinformatics/btu077. pmid:24532719
- Yednock, B.K., Sullivan, T.J, Neigel, J.E.** (2015) *De novo* assembly of a transcriptome from juvenile blue crabs (*Callinectes sapidus*) following exposure to surrogate Macondo crude oil. *BMC Genomics*. **16**:521.
- Zadlock, F.J., Rana, S.B., Alvi, Z.A., Zhang, Z., Murphy, W., et al.** (2017) *De Novo* Assembly and Analysis of the Testes Transcriptome from the Menhaden, *Bervoortia tyrannus*. *Fish Aqua J* **8**: 186. doi:10.4172/2150-3508.1000186
- Zerbino, D.R., Birney, E.** (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. **18**(5): 821–829. doi: 10.1101/gr.074492.107
- Zhao, Q.Y., Wang, Y., Kong, Y.M, Luo, D., Li, X. et al.** (2011) Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*. **12**(Suppl 14): S2.
- Zohar, Y., Munoz-Cueto, J., Elizur, A., Kah, O.** (2010) Neuroendocrinology of reproduction in teleost fish. *General and Comparative Endocrinology* **166** 438-455.

## **Appendix A**

PLOS ONE publishes all of the content in the articles under an open access license called “CC-BY.” This license allows you to download, reuse, reprint, modify, distribute, and/or copy articles or images in PLOS journals, so long as the original creators are credited (e.g., including the article’s citation and/or the image credit). Additional permissions are not required: <http://journals.plos.org/plosone/s/licenses-and-copyright>.