Electronic Supplementary Material (ESI) for Environmental Science: Processes & Impacts.
This journal is © The Royal Society of Chemistry 2022

Supplementary information

RNA-seq reveals potential gene biomarkers in fathead minnows (*Pimephales promelas*) for exposure to treated wastewater effluent

*Peter G. Schumann[a], Emma B. Meade[a], Hui Zhi[b], Gregory H. LeFevre[b], Dana W. Kolpin[c], Shannon M. Meppelink[c], Luke R. Iwanowicz[d], Rachael F. Lane[e], Angela Schmoldt[f], Olaf Mueller[f], Rebecca D. Klaper[a, f, *]*

[a]University of Wisconsin-Milwaukee, Milwaukee, WI USA; [b]University of Iowa, Iowa City, IA USA; [c]U.S. Geological Survey, Iowa City, IA USA; [d]U.S. Geological Survey, Kearneysville, WV USA; [e]U.S. Geological Survey, Lawrence, KS USA; [f]Great Lakes Genomics Center, Milwaukee, WI, USA

*Corresponding Author:
rklaper@uwm.edu; Phone: 414-382-1713; School of Freshwater Sciences, University of Milwaukee, Milwaukee WI, United States

Any use of trade, firm, or product, names is for descriptive purposes only and does not imply endorsement by the authors or the U.S. Government.

CONTAINS:

Field site details, exposure details, data processing and analysis details, 3 Supplementary Tables, 9 Supplementary Figures, 12 pages in total

S1. Materials and Methods

Chemical data for Muddy Creek pharmaceutical characterization is available online at https://nwis.waterdata.usgs.gov.

S1.1. Bioluminescent Yeast Estrogenicity Screen (BLYES)

20 µL of sample extract was added in triplicate to the wells of white, solid-bottom 96-well microtiter plates and evaporated at room temperature in a Class II biological safety cabinet. Following evaporation, 200 µL of a 48-hour culture of strain BLYES adjusted to 0.4 (OD_{600}) in fresh yeast minimal media (YMM eu^-, ura^-) was added to each well. A 12-point standard curve of

17β-estradiol (E2; Sigma-Aldrich Co.) was included on each plate. A media control was included on all plates to establish background luminescence. Plates were covered and incubated in the dark at 30 °C for 4 hours. Luminescence was quantified using a SpectraMax M4 microplate reader (Molecular Devices) in luminescence mode (1,000 millisecond integration time), and estrogen equivalents (E_2Eq) of each sample were determined via interpolation to a 4-parameter curve within SoftMax Pro 6.2.2 (Molecular Devices). Relative net agonistic activity per liter of sample was then calculated based on sample concentration. The detection limit for this assay was 0.18 ng/L $E_2Eq_{(BLYES)}$.

S1.2 Caged-fish Exposure

Fish were transported to the field site in sex-segregated, aerated coolers on the morning of deployment. Water temperature during the 4-hour transit period was maintained at 20°C. Upon arrival, fish were acclimated to temperatures within 2°C of corresponding site water (20-25°C) and deployed in cages consisting of modified vinyl-coated stainless-steel mesh minnow traps with an interior volume of approximately 20 L. Minnow cages were lined inside with a fine nylon mesh which was secured with zip ties and fishing line across all wire mesh surfaces to prevent fish from escaping. Each cage also contained a 3 by 6-inch arched shelter cut from polyvinylchloride pipe which was secured to the inside of the wire cage with zip ties (**Fig. S.2**). Three cages each containing six male and six female fathead minnows were deployed at each site. Each cage was secured to a submerged cinderblock arranged in a triangular formation as illustrated in **Fig. S.3**. Cages were positioned approximately 1-3 inches above the stream bed and were monitored daily to ensure they remained submerged and prevent sediment from building up around the cages. After the 96 h exposure, the fish were transported in buckets for sample processing (transit time < 1 hour). Additionally, a control group that consisted of 18 males and 18 females was also transported to the Muddy Creek field site but was not deployed into the stream. Tissue samples were collected from these control fish on Day 1 of the caged fish exposure (July 14).

S1.3 Data Processing using DaMiRseq

Using the DaMiRseq package available through R, the estimated count data from Kallisto was first filtered by removing genes with less than 10 counts across 50% of samples. This was done to help ensure that genes with an extremely low expression were excluded from the analysis as often these genes can be indistinguishable from sampling noise. We also decided to remove hypervariant genes, which were defined as genes with all sample "class" coefficients of variance greater than 3. Sample "classes" are determined by the combination of the cage number and the site (i.e., "Upstream_1, Upstream_2, Upstream_3, Downstream_4, Downstream_5, Downstream_6). After filtering, the estimated count data was normalized using the variance stabilizing transformation (VST) method, which removes the dependence of count data variation on the mean.

S1.4 RNA-seq validation using RT-qPCR

Genes were considered good candidates to validate RNA-seq results by RT-qPCR if they (1) were expressed in both female and male fish, (2) were differentially expressed in the upstream vs. downstream sites with a Benjamini-Hochberg (B-H) adjusted p-value < 0.05, and (3) had Log_2(Fold Change) values of > 1 or < -1. A total of only 7 genes met each of these criteria. Of these 7 genes, 3 genes with the greatest Log_2(Fold Change) values and lowest p_{adj}-values were selected for validation. Hypoxanthine phosphoribosyltransferase 1 (hprt1) was selected as a housekeeper gene as it was found to be highly expressed in both male and female fish (baseMean values >1000) and was not found to be differentially expressed between sites.

Candidate gene expression was assayed using RT-qPCR only in the upstream and downstream samples that were used in RNA-sequencing. Briefly, cDNA was synthesized from extracted mRNA (Direct-zol RNA Miniprep, Zymo Research) using a High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) with random primers according to the manufacturer's instructions. The synthesized cDNA was then used to perform RT-qPCR with

Applied Biosystems StepOnePlus™ using Absolute qPCR SYBR Green (Thermo Scientific). Target gene expression was normalized to hprt1 for each sample, and the ΔΔCt method was used to determine the relative fold change of target genes (1) with respect to upstream samples.

All statistical analyses on qPCR results were performed in R. The normality for each gene was tested separately in each site for both male and female data using a Shapiro-Wilk test. Outliers were defined as relative fold change values 1.5x outside the interquartile range. Normally distributed data that contained no outliers was then analyzed using a Welch two-sample t-test ($\alpha = 0.05$). Non-normally distributed data or data that contained outliers was analyzed using a Wilcoxon rank sum exact test ($\alpha = 0.05$). Effect size was also analyzed by calculating a Cohen's d value for each comparison. A summary of these statistics can be found in **Table S.3**.

S1.5 Weighted gene co-expression network analysis (WGCNA)

A weighted gene co-expression network analysis was conducted using the WGCNA package in R for both male and female upstream vs. downstream DEGs. The gene co-expression network is specified by an adjacency matrix of co-expression similarity between any pair of two genes among all the samples. Given that we did not directly characterize the phenotypic states of the FHMs after exposure, we chose to construct unsigned networks. Additionally, we chose unsigned networks as it is possible that certain biological pathways can incorporate both negative and positive controls of gene expression. This allowed modules to be formed that included both negatively and positively correlated genes within the adjacency matrix. Thus, through this approach we were able to broaden the scope of the downstream functional enrichment analyses.

Regarding the cluster dendrograms generated with the male and female datasets (**Fig. 2**), the male FHM dendrogram (**A**) shows more clear association and grouping of genes into modules in contrast to the female FHM dendrogram (**B**). One factor that likely contributed to the differences

in clarity of gene associations between the male and female dendrograms is the overall size of the networks. After filtering and normalization of the count data, 27,065 genes remained in the male dataset and 32,876 genes remained in the female dataset. However, the soft-thresholding power selected for the female dataset (power = 8) represents the ideal threshold as any greater value results in modules that are too large for more specific biological characterizations via pathway analysis.

S1.6 Protein-protein interaction (PPI) and co-expression networks to identify hub genes

Genes found in the center of a network usually exhibit more important functions. Therefore, it is important to not only consider the correlations of individual genes with module eigengenes, but to also analyze the degree of the interconnectivity of those genes within a module. To assess this, STRING was utilized to construct PPI networks using the genes contained within significant WGCNA modules. These PPI networks were then visualized using Cytoscape software (**Fig. 5a, 5b** and **Fig. S.7a, S.7b**) and MCODE was used to identify the most densely interconnected gene (i.e., molecular) complexes. The molecular complexes identified by MCODE were compared to the genes that met the WGCNA hub gene criteria.

Table S.1 Primer pairs used for RT-qPCR analysis.

Gene	Primer Sequence		Amplicon Size (bp)	Tm (°C)	GenBank ID
hprt1	Fwd	5'-GATGAAGAGCAAGGTTATGAC-3'	214	51.0	XM_039665711.1
	Rev	5'-ACACAGAGCAACGATATGG-3'		52.2	
cops2	Fwd	5'-CAAACAGGCAGTGTCAAAGAAG-3'	99	54.6	XM_039695148.1
	Rev	5'-GCATCCAGTGTGGTCTCATAAA-3'		54.9	
dnaja	Fwd	5'-CACGAAGACAACCTCACCAT-3'	101	54.9	XM_039648216.1
	Rev	5'-ACGAGCAGCGATCTGTTATC-3'		54.7	
serpinh1b	Fwd	5'-CCACCACAAGATGGTCGATAA-3'	104	54.6	XM_039662686.1
	Rev	5'-TGTCCTCATAGAAGCCGTAGA-3'		54.7	
	Rev	5'-TGTCCTCATAGAAGCCGTAGA-3'		54.7	

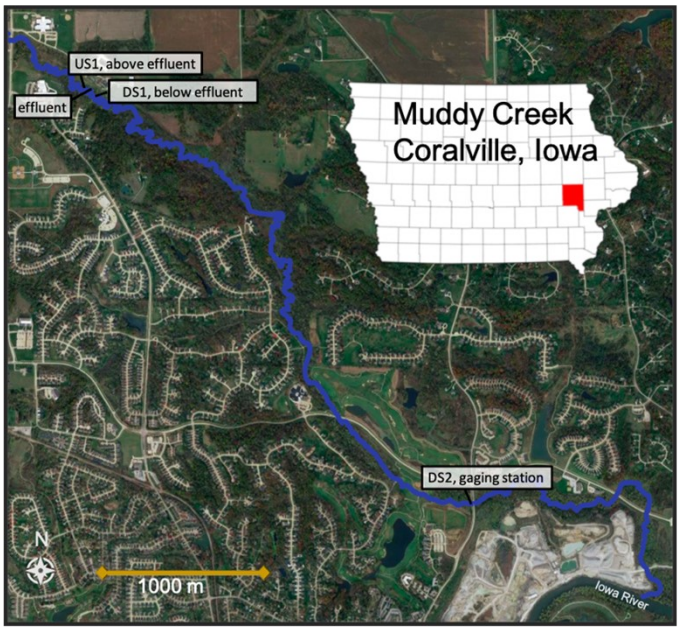

Fig. S.1 Stream sites used for caged fathead minnow (*Pimephales promelas*) exposures in Muddy Creek, Coralville, Iowa (Latitude 41°42'00", Longitude 91°33'46"): US1 (100m above effluent), DS1 (0.1 km below effluent), and DS2 (5.1 km below effluent). Background image from Google Maps.

Fig. S.2 Cages consisted of a cylindrical vinyl-coated steel minnow trap (18" length, 9" diameter) lined with fine nylon mesh. Each cage contained two polyvinylchloride shelters (one strapped to each half). Cages were securely closed with zip ties and transported to the field site in buckets.

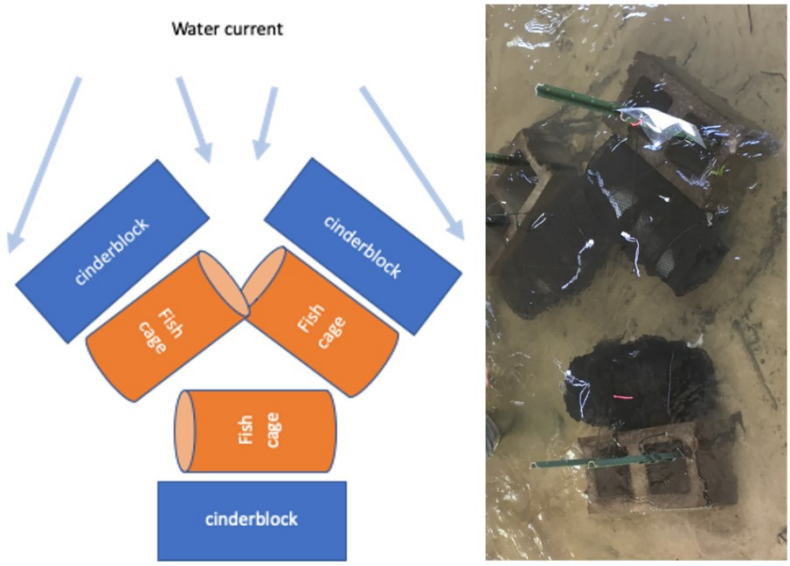

Fig. S.3 Fish cages were strapped to cinderblocks using zip ties and nylon cord such that cages remained completely submerged and had at least 1-3 inches clearance above stream bed. Two 5' fence posts (3" width) were used to stake each cinderblock into the sandy streambed. The open triangular formation of the cinder blocks was used to lessen the force of the current on the fish cages and prevent sediment from building up between and around the cages.

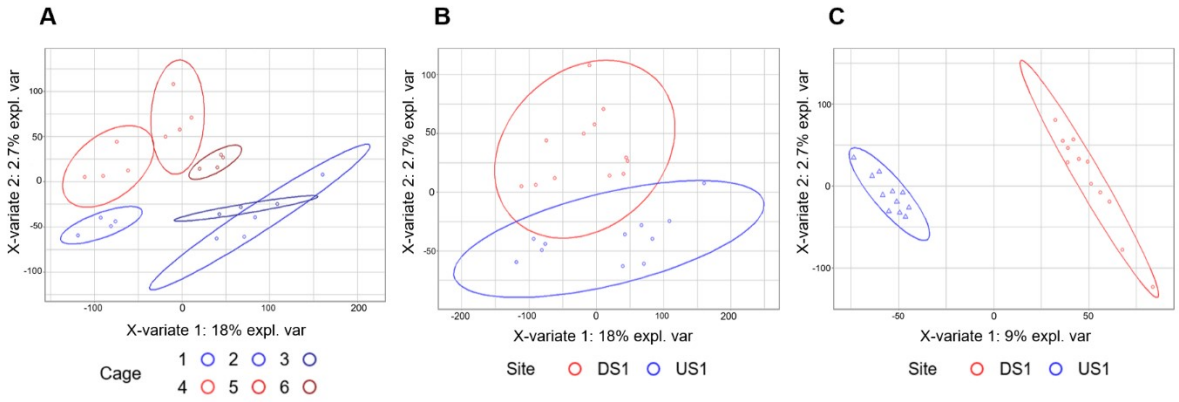

Fig. S.4a Partial least squares-discrimination analysis (PLS-DA) to assess potential batch effects resulting from cage differences (i.e., cage effects) in the female fathead minnow dataset with US1 data represented in blue and DS1 data represented in red. Normalized RNA-seq count data was first (**A**) clustered by cage with US1 cages (1-3) and DS1 cages (4-6) to visualize cage-dependent separation. The data was then (**B**) clustered by site to compare against (**C**) the batch corrected data. Batch corrected data resulted in separate clusters with a 9% explanation of variance for X-variate 1.

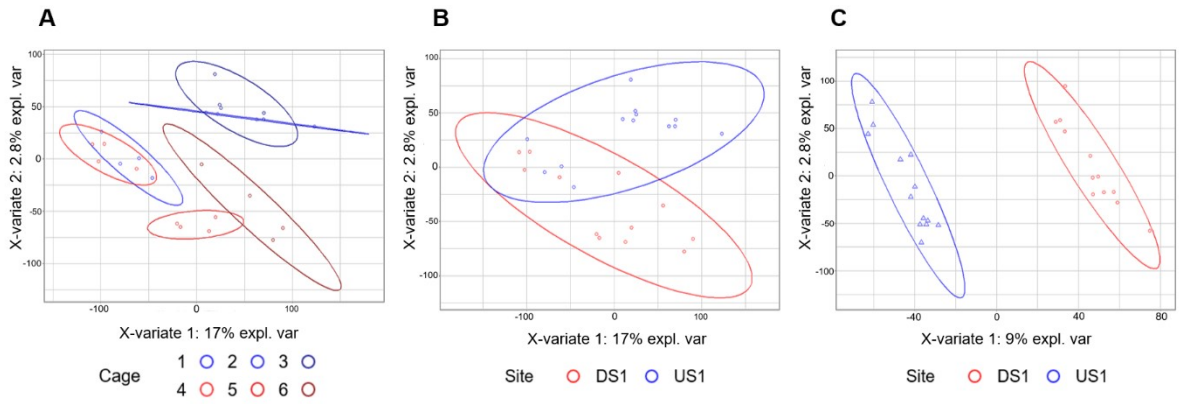

Fig. S.4b Partial least squares-discrimination analysis (PLS-DA) to assess potential batch effects resulting from cage differences (i.e., cage effects) in the male fathead minnow dataset with US1 data represented in blue and DS1 data represented in red. Normalized RNA-seq count data was first (**A**) clustered by cage with US1 cages (1-3) and DS1 cages (4-6) to visualize cage-dependent separation. The data was then (**B**) clustered by site to compare against (**C**) the batch corrected data. Batch corrected data resulted in separate clusters with a 9% explanation of variance for X-variate 1.

S2. Supporting Results

S2.1 RNA-seq validation using RT-qPCR

The gene expression patterns found in the RT-qPCR assays of the same RNA samples used for RNA-seq were comparable in terms of directionality and significance (except for the female *dnaja* results), giving support to the sequencing data (**Fig. S.5**). The magnitudes of expression were not always reflected as strongly in the qPCR results. Notably, the Log_2(Fold Change) values for *cops2* was about half of that found in the RNA-seq results and the female, downstream *dnaja* qPCR result was not found to be significantly different from the upstream samples. However, these differences in expression patterns found using RT-qPCR and RNA-seq are likely due to sensitivity differences between the two methods, which could be due to a combination of differences in reaction chemistry, detection limits, and data analysis pipelines (2,3).

Table S.2 Bioluminescent yeast estrogenicity in Muddy Creek at each of the four sampling sites during the 96-h FHM exposure. The measured estrogen equivalents (E_2Eq) ranged from 0.49 ng/L in the US1 sample to 1.04 in the DS1 sample.

LABID	Project Code	Site ID	Site Location	Collection Date	Collection Time	Rec Date	MeOH Final Volume (µL)	E2Eq (BLYES) ng/L
MM-57374A	MCD	5454050	Muddy Cr above Treatment Effluent at N Liberty, IA	7/16/2019	825	7/17/2019	100	0.49
MM-57372A	MCD	5454051	Muddy Cr at Treatment Effluent at N Liberty, IA	7/16/2019	835	7/17/2019	100	0.88
MM-57375A	MCD	5454052	Muddy Cr blw Treatment Effluent at N Liberty, IA	7/16/2019	845	7/17/2019	100	1.04
MM-57373A	MCD	5454090	Muddy Cr at Coralville, IA	7/16/2019	930	7/17/2019	100	0.7

Table S.3 Summary of statistics for RT-qPCR results. Upstream male: n = 13, Downstream male: n = 12, Upstream female: n = 11, Downstream female: n = 12. Genes with significant ($p<0.05$) fold change differences between the downstream and upstream samples are bolded.

Sex	Gene	Test	p-value	Cohen's d
M	**cops2**	wilcox.test	**0.04571**	0.8349365
	dnaja	t.test	**0.0003225**	1.760961
	serpinh1b	t.test	**0.0006813**	1.69829
F	**cops2**	t.test	**0.02353**	1.065585
	dnaja	t.test	0.78	0.1195489
	serpinh1b	t.test	**0.004415**	1.413549

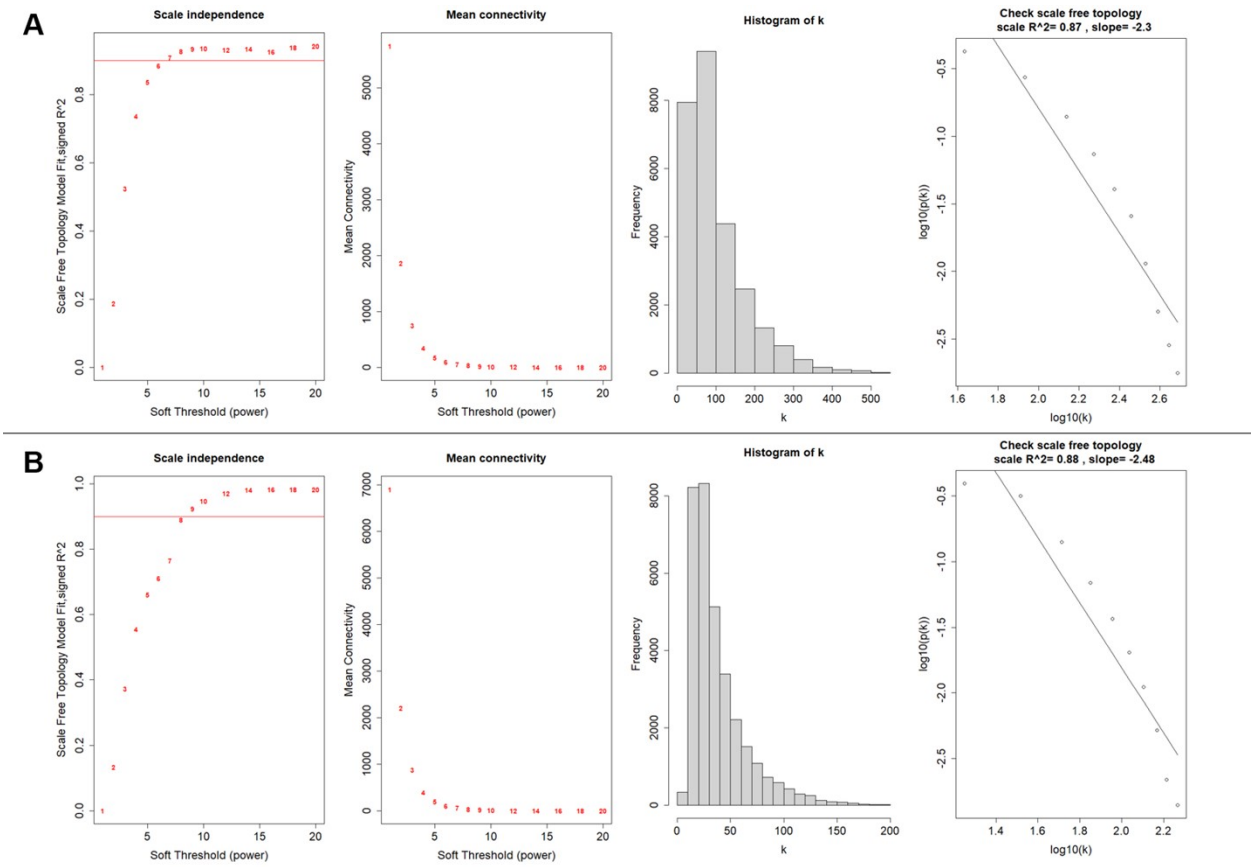

Fig. S.5 Analysis of weighted gene co-expression network analysis (WGCNA) network topology for (**A**) male and (**B**) female fathead minnow datasets. The scale-free fit index (y-axis) was plotted as a function of the soft-thresholding power (x-axis) where the red horizontal line indicates the R^2 cutoff for approximating scale-free topology. The mean connectivity (degree, y-axis) was plotted as a function of the soft-thresholding power (x-axis). Histograms of connection frequencies. Log-log plots of the histogram connectivity measured against a linear model with an R^2 value of 0.87 for the male dataset and 0.88 for the female dataset.

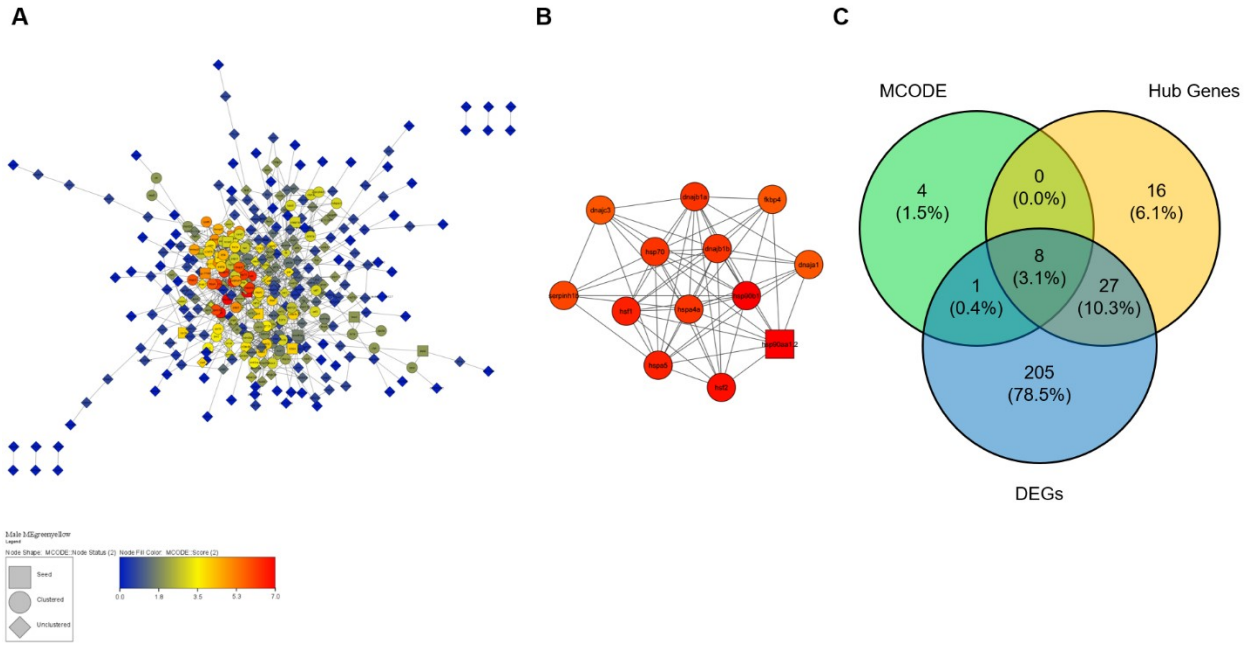

Fig. S.6a (**A**) Protein-protein interaction (PPI) network of genes in the male fathead minnow MEgreenyellow module. Node colors are scaled by molecular complex detection (MCODE) scores where green nodes are less interconnected and red nodes are highly interconnected. (**B**) Highest scoring (i.e., most densely interconnected) molecular complex as identified by MCODE. (**C**) Venn diagram of genes identified by MCODE clustering, genes that met the WGCNA hub gene criteria (|MM|>0.8 and |GS|>0.5) and genes that were found to be significantly differentially expressed (DEGs). 8 genes met all these criteria: *hspa4a*, *dnajb1a*, *hspa5*, *dnajb1b*, *fkbp4*, *dnaja1*, *hsp90aa1.2*, and *serpinh1b*.

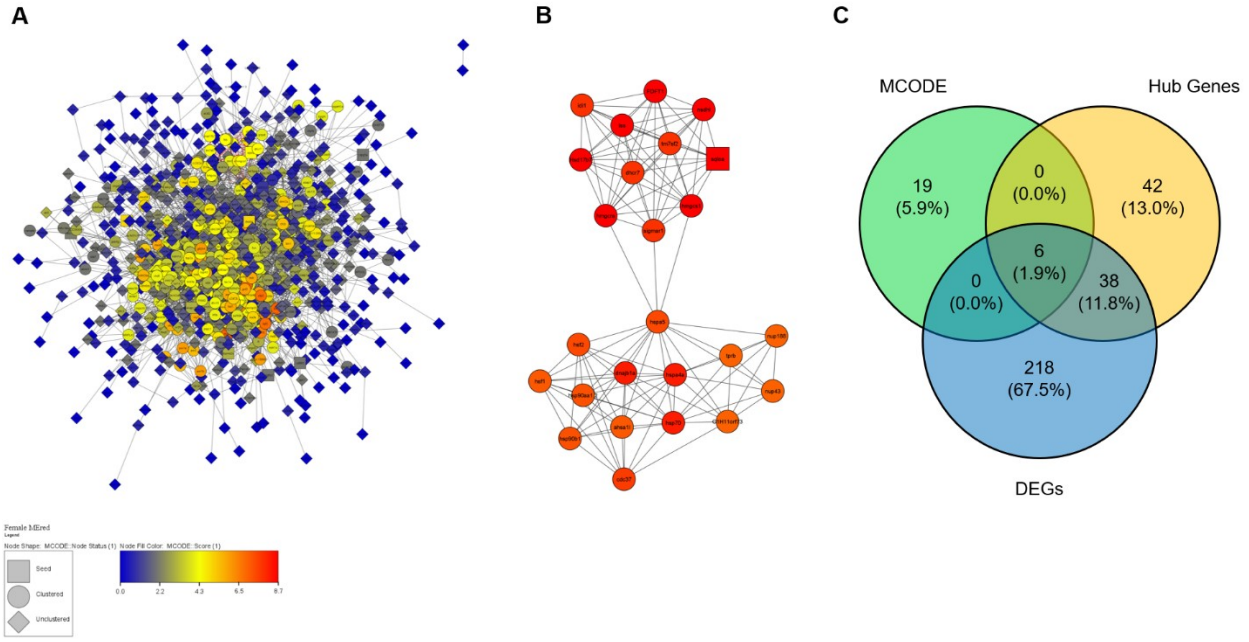

Fig. S.6b (**A**) Protein-protein interaction (PPI) network of genes in the female fathead minnow MEred module. Node colors are scaled by molecular complex detection (MCODE) scores where green nodes are

less interconnected and red nodes are highly interconnected. (**B**) Highest scoring (i.e., most densely interconnected) molecular complex as identified by MCODE. (**C**) Venn diagram of genes identified by MCODE clustering, genes that met the WGCNA hub gene criteria (|MM|>0.8 and |GS|>0.5) and genes that were found to be significantly differentially expressed (DEGs). 6 genes met all these criteria: *hspa4a*, *dnajb1a*, *hspa5*, *hsf1*, *hsp70.3*, and *hsp90aa1.2*.

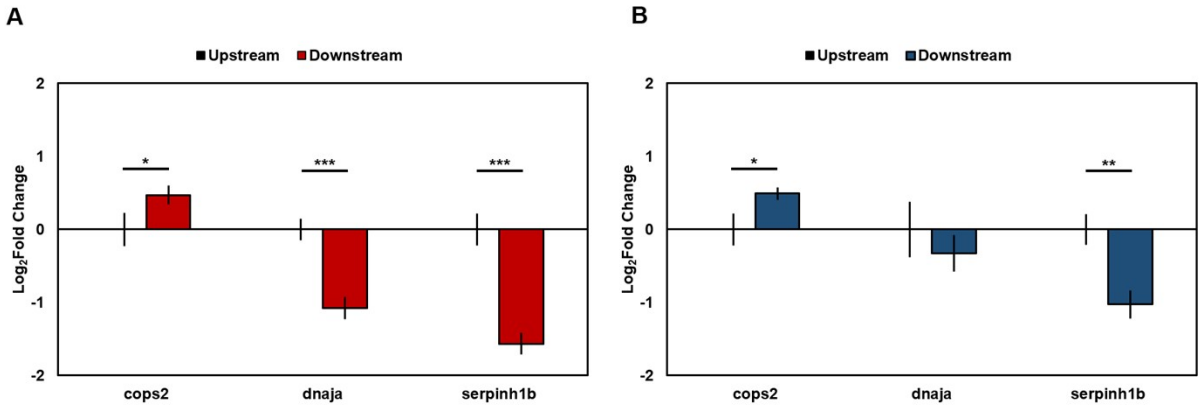

Fig. S.7 Fold change of selected validation genes from RNA-seq data shows significance in all validation genes in (**A**) male fathead minnows and only 2 of the 3 genes in (**B**) female fathead minnows. Relative Log_2(Fold Change) generated from RT-qPCR assays on the same samples used for RNA-seq show comparable trends in the validation genes. Error bars represent standard error of the mean (SEM). Significance code is as follows: * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Supporting References

1. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. Methods. 2001 Dec 1;25(4):402–8.

2. Huang IJ, Dheilly NM, Sirotkin HI, McElroy AE. Comparative transcriptomics implicate mitochondrial and neurodevelopmental impairments in larval zebrafish (Danio rerio) exposed to two selective serotonin reuptake inhibitors (SSRIs). Ecotoxicol Environ Saf [Internet]. 2020 Oct 15;203:110934. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0147651320307739

3. Benjamin Alexander-Dann, Lorena Pruteanu L, Erin Oerton, Nitin Sharma, Ioana Berindan-Neagoe, Dezső Módos, et al. Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data. Mol Omi [Internet]. 2018 Aug 6 [cited 2021 Jul 22];14(4):218–36. Available from: https://pubs.rsc.org/en/content/articlehtml/2018/mo/c8mo00042e