STUDENT SUMMER INTERNSHIP TECHNICAL REPORT

Investigation of Heavy Metal Biomarkers for the Assessment of Remediated Surface Waters

DOE-FIU SCIENCE & TECHNOLOGY
WORKFORCE DEVELOPMENT PROGRAM

Date submitted:

December 4, 2020

Principal Investigators:

Juan C. Morales (DOE Fellow)
Florida International University

Katrina Waters (Division Director of Biological Sciences)
Lisa Bramer (Data Scientist)
Pacific Northwest National Laboratory

Ravi Gudavalli Ph.D. (Program Manager)
Florida International University

Leonel Lagos Ph.D., PMP® (Program Director)
Florida International University

Submitted to:

U.S. Department of Energy
Office of Environmental Management
Under Cooperative Agreement # DE-EM0000598

FIU | Applied Research Center
FLORIDA INTERNATIONAL UNIVERSITY

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Heavy metals are natural constituents of the earth´s geological formations and their prolonged exposure is known to cause deleterious health effects in humans [1][2][3]. In theory, metal compounds are classified as toxic in nature, regardless of their density or atomic mass [4]. Moreover, much of the heavy metal contamination seen today in soils and surface waters across many rivers in the United States began in the late 18^{th} century. With the rapid increase in technology, there has also been an increase in many activities such as fossil fuel burning, mining, agriculture, and landfill contamination, which are shown to affect the water quality [22][23][24].

The use of fish, particularly zebrafish, has become important in areas of toxicology and drug discovery [5]. This vertebrate model can be used to reveal effects in embryonic development, chemical toxicity or molecular mechanisms using microarray and RNA-Seq technologies [6][7]. Hence, during the early 1980s, the field of toxicology first proposed to study the frequency of mutations in response to environmental carcinogens [8]. Ever since, zebrafish have been shown to effectively and rapidly uncover toxicological mechanisms for many contaminants and improve the understanding on the impact on vertebrates [9][10].

Microarray technology is a novel tool in molecular biology which quantifies hundreds to thousands of gene transcripts from a given tissue or cell sample simultaneously [11]. A microarray has thousands of oligonucleotides or DNA fragments of a known sequence in a chip. After hybridization, the gene expression profiling can be used as an important source to discover molecular mechanisms and toxicity patterns post exposure in many species [12]. Meanwhile, zebrafish toxicological studies suggest that acute exposure to environmental heavy metals can suppress transcription factors on the DNA and block access into the DNA methylation machinery [13]. Meanwhile, toxic effects in site-specific patterns of methylation activation or repression result in gene-specific synthesis and influence a response to adaptation or defense mechanisms in response to stress [13]. Patterns of change in genes from acute exposure to Cd, Co and Cu suggests an effect on zebrafish in areas of motor and neuromast development and blockage to cellular transport pathways, activating oxidative stress responses [14][15]. Since the central nervous system structures and organ functions are highly conserved regions, a wide range of toxic alterations are mostly universal between zebrafish and human species [16].

Random Forest

The prediction of environmental biomarkers using large amounts of data (microarray and RNA-Seq studies) is of great need. To help circumvent this problem, an important technique used as a standard in data analysis is the Random Forest (RF) method. RF is a classification and regression algorithm that is based on the aggregation of trees after training the gene expression data, and it validates sets of variables using predictors for future observations [17]. In addition, RF has proven to be excellent in analyzing large numbers of variables and can function by predicting

measures using variables of importance. This method uses supervised machine-learning algorithms that process and analyze large numbers of predictor variables in high throughput data and is representative of an ensemble of learning methods [18].

Several important characteristics have determined that RF is effective in determining genes and the role of each variable from response in prediction [19]. Also, RF predictive power shows that it is excellent at predicting variables which have noise and also different types of classes [19]. The model uses supervised random sampling strategies and addresses the predictive variability scores using variable of importance measures (VIMs). Also, the algorithm automatically computes and ranks variables according to their class and predictive ability [17].

Motivation

Currently, regulatory toxicology does not have effective methods that can provide testing capacity to measure change and biological alterations from heavy metal stress. Both state and federal regulatory guidelines use traditional dose-response thresholds which measure changes through trophic levels, and results are compared with analytical sediment benchmarks to create hazard quotients [20]. Moreover, as outlined by the Environmental Protection Agency (EPA), the compounds of highest concern for human health are As, Cd, Co, Cr, Cu, Hg, Ni, Pb and U [21]. Understanding the many mechanisms by which genes are modulated due to stress responses may have important implications for treatment and surveillance.

In this report, a novel model is proposed for the identification of biomarkers used to evaluate risk factors involved in toxicity mechanisms associated with the exposure of heavy metals using bioinformatics and machine-learning. We hypothesize that gene biomarkers can be used to discriminate important pathways associated with (As, Cd, Hg) toxicity and their corresponding activity differences among exposures. The identified biomarkers may be used as unique fingerprints post-exposure to assess the long-term consequences, for example in a safety genomic evaluation.

The models will be tested using:

1. Candidate gene selection and classification.
2. Principal Component Analysis (PCA).
3. Random Forest (RF) Analysis.
4. Ranking of genes using Variables of Importance Measures (VIMs).

2. EXECUTIVE SUMMARY

This research work has been supported by the DOE-FIU Science & Technology Workforce Development Initiative, an innovative program developed by the U.S. Department of Energy's Office of Environmental Management (DOE-EM) and Florida International University's Applied Research Center (FIU-ARC). During the summer of 2020, a DOE Fellow intern Juan Carlos Morales, spent 10 weeks participating in a virtual summer internship with Pacific Northwest National Laboratory (PNNL) under the supervision and guidance of Dr. Katrina Waters, Director of the Biological Sciences division.

The intern's project was initiated on June 1, 2020 and continued through August 7, 2020. His deliberate objective was to identify heavy metal biomarkers and identify pathway-based mechanisms affected by exposure to heavy metals using model-based systems engineering.

3. MATERIALS AND METHODS

This section describes the use of supervised Random Forest (RF) analysis for the identification of biomarkers as well as the use of pathway-based analysis as a measure of risk factor. The optimized model was used to classify heavy metal treatment, followed by biomarker gene selection for liver tissue. The study design was formulated as a machine-learning problem and the process is presented in the detailed procedures as shown in Figure 1.

Experimental Setup

All the programs used the operating system listed in Table 1 with Windows 10 Pro. The configuration of the code and description can be found in Table 2, and was developed in conjunction with Pacific Northwest National Laboratory.

Table 1. Summary of CPU Descriptive Table Used to Run the Samples

Name	Description
Lenovo ThinkPad X1	13-inch, Mid 2018
Processor	Intel ® Core ™ i7/-85500U CPU @ 1.80 GHz
Memory	8GB 1600 MHz DDR3
Graphics	Intel HD Graphics 4000 1536MB
Serial Number	R90V36H8
System Type	64-bit operating system x64- based processor

All resources, supplementary material and developed code can be found in Appendix A, Table 4.

Table 2. Demographic Summary of Zebrafish Heavy Metal Studies

Compound	Dataset	*n*	Conc. (ppm)	Tissue type	ID (s)
Arsenic	3048	12	15.0	A	[22]
Arsenic	30062	12	15.0	A	[23]
Cadmium	41622	10	30.0	A	[23],[24]
Cadmium	41623	10	30.0	A	[23],[24]
Mercury	18861	12	200.0	A	[25]
Control	Combined	60	-	A	[22],[36,][37],[25]

A. Liver (hepatocytes).
Gene transcript IDs platforms - GPL2715.
IDs. Study references.

Workflow diagram

The general workflow diagram of this study is displayed in Figure 1. We first processed five datasets in which the mRNA microarray expressions were selected accordingly. Then we applied the feature selection methods on the mRNA and mapped our candidate genes using bioinformatics. In the same process, we selected common genes across platforms and identified the prognostic biomarkers using Random Forest analysis.

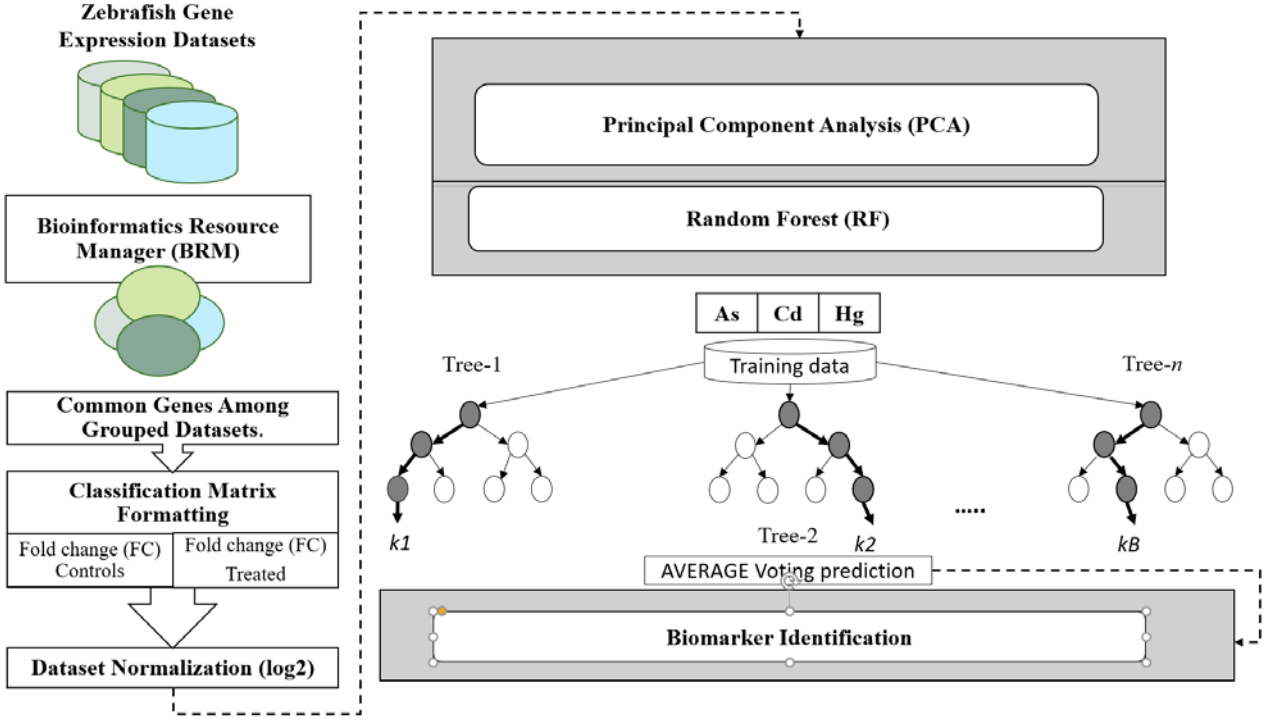

Figure 1. Flowchart of Random Forest Algorithm that Identifies Heavy Metal Biomarkers using Zebrafish Gene Expression Data.

Data preprocessing

First, we downloaded the zebrafish gene expression datasets using the Gene Expression Omnibus (NCBI GEO) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc= GSE30062, GSE3048, GSE41622, GSE41623, GSE18861, respectively [26]. Among the 5 datasets used, we chose Systems Biogenesis based on the GPL2715 platform. All the datasets were downloaded in the .txt format. Once we began our download, the preliminary analysis led to the classification and class toxicity profiles. A total of 56 combined heavy metal (As, Cd and Hg) treated samples were processed along with 60 controls and used for biomarker identification. Three different heavy metals including As, Cd, and Hg respectively, were analyzed.

Bioinformatic Analysis

For downstream bioinformatic analysis, the Bioinformatics Resource Manager (BRM) (https://cbb.pnnl.gov/brm/) was used where a workflow in which zebrafish identifiers were converted and merged. This can be found in a summarized form in Figure 2. We utilized a platform that has a familiar get started menu and provides access to most data imports and retrieval options. BRM is an environment for data management, mining, integration and functional annotation of high throughput biological data and managed by Pacific Northwest National Laboratory (PNNL) [27]. It also performs retrievals for batch annotations, cross reference of species and retrieval of micro RNA (miRNA) data necessary for system biology research [27].

Mixed identifiers (Entrez Gene IDs) were introduced from human and zebrafish. Integrating BRM to identify the subset genes of regulated exposure zebrafish was important in reducing the amount of redundant gene identifiers mapped. Both BRM and DAVID software recognized mixed identifiers once the *danio rerio* (zebrafish) retrieval option was selected to obtain orthologue mapping for most of the data [28]. Each dataset was primarily formatted to a delimited file format before mapping the spreadsheets.

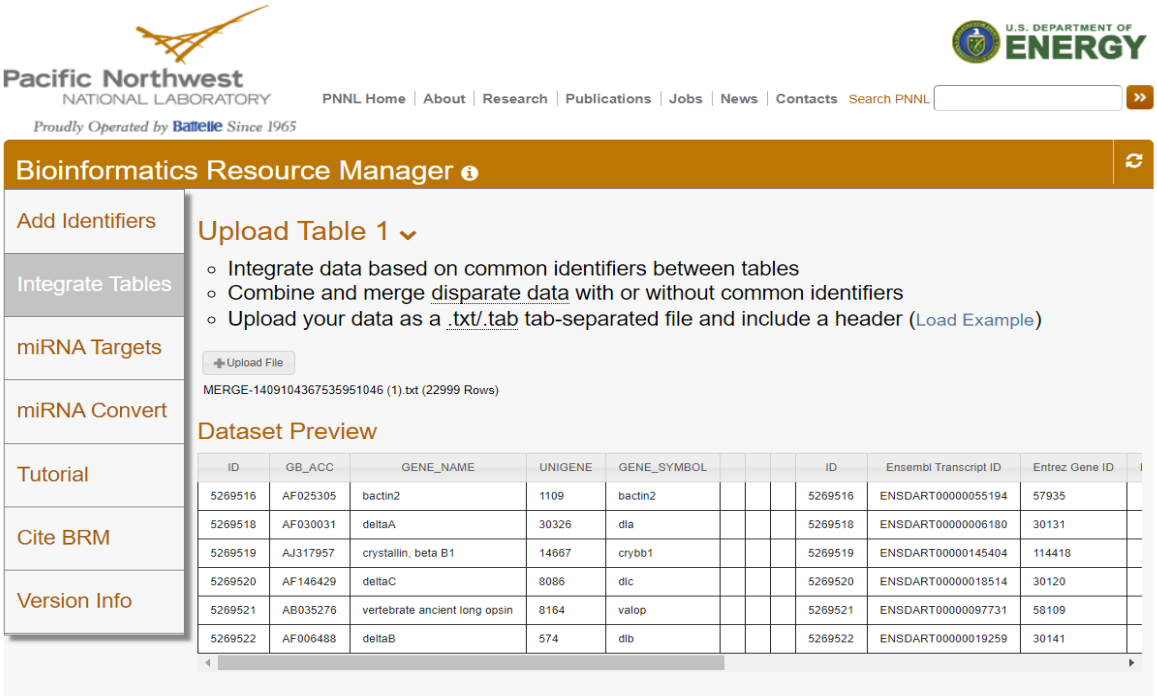

Figure 2. Bioinformatics Resource Manager integrated, merged and identified common ENTREZ IDs from nucleotide GENBANK Accession Numbers. This technique formatted a total of 11 spreadsheets in which a total of 2,914 genes were mapped according to their assigned probe identifier.

Fold Change

In this section, the means to measure epigenetic regulation is described, which is to calculate the difference between samples or fold change. Fold change is a method that measures how much a quantity differs in fitting from the beginning to the end. As an example, an expression value at 40 and a terminal value at 80 describes a fold change of 2, equivalently two more times. It can also be described as a proportion distinction between the final and the primary output over the fundamental value.

The fold change process is often practiced in the interpretation of RNA-Seq and microarray gene expression data, in which the level of intensity or counts is estimated determining the variation in expression level. If the primary value such is A and the final value is B, the fold change variation can be computed as B/A-1, or equivalently, as (B-A)/A [29]

A primary downside to this approach is that the data may be biased and may cause avoidance of certain differentially expressed genes with large variations (B-A) but small ratios (A/B), which may elicit a miss rate of large concentration [29]. To circumvent our problem, we applied the log2 transformation to represent the fold change variation.

Normalization of Datasets

All the preprocessing, processing and statistical analysis were performed using R software. The transformation procedure taking the different gene expression data across all platforms and samples and grouping each sample together for screening, was calculated using a predefined data-formatting package. The ($\log_2$(expression ratios)) was calculated, respectively. The idea behind log2 transformations serves directly to fit overexpressed and under expressed values. As an example, if we assume there are 45 counts per read in the healthy control and 90 read counts in the treatment for gene A, this would mean a fold change of 2. However, if this exercise were reversed, the fold change expression value will be 0.5 representing under expression. Having different unit values for up-regulation and down-regulation allows for a uniform normalization procedure, treating regulation equally. This also allows for a continuous mapping space. Finally, the data corresponding to the healthy control zebrafish was grouped and averaged, respectively. Several packages and libraries were created and implemented.

RStudio setup

This step-in biomarker identification is crucial for model development. The feature selection method used in this report was developed in the R Software and programming language (version 1.3.1056). The other platforms for gene matching and further analysis were implemented using Microsoft Excel 2010. Each package listed in Table 3, Table 4, Table 5 and Table 6 controls individual formatting and downstream analysis of GSE3048, GSE30062, GSE41622, GSE41623, and GSE18861, respectively.

Table 3. Features and short descriptions of random forest implementations.

R Software packages	Description
Library (Random Forest).	Breiman and Cutler´s Random Forest for Classification and Regression
Library (ggplot).	Maps variables to aesthetics
Library (pcaMethods).	Principal Component Analysis validation and visualization of results
Library (tidyverse).	Designed for data formatting and structure
Library (dplyr).	Data manipulation package picking variables based on their (names)
Library (impute).	Imputation for microarray data
Library (RColorBrewer).	Heatmap visualization using color palettes for graphics

After the subsets of expression data are down to an individual contaminant (As, Cd, and Hg), we make sure that each of the order of the samples in the data are the same as in the metadata. To make sure, we pulled the meta info along with metadata associated with the correct samples.

Table 4. Step by Step Downstream Analysis Using Rstudio

	Description	Script to use:
1.	Install packages from.	See Table 3
2.	Set your working directory to.	*setwd("C:/Users/jumor/Desktop/Pacific Northwest Nat. Laboratory/R/Random Forest/Lisa")*
3.	Obtain sample ids from experiments studying As, Cd, and Hg.	*cd_meta = subset(meta, Experiment_HeavyMetal == "Cd")* *cd_ids = cd_meta$SampleID*
4.	Select each subset in the data to As, Cd, Hg study.	*cd_data = data[,cd_ids]*
5.	Set the row names in the dataset to gene names.	*rownames(cd_data) = data$Entrez.Gene.ID*
6.	Confirm the order of the samples matches with the metadata and pull the meta info with associated samples.	*ord_vec = match(cd_meta$SampleID, names(cd_data))*
7.	Switch the order of the data columns to match the metadata.	*cd_data_ord = cd_data[,ord_vec]*
8.	Transpose the data so the rows are samples and genes are columns.	*cd_t_data = t(as.matrix(cd_data_ord))*
9.	Remove any genes with NA values and only select the genes with non-.NA values.	*rmv_ids = which(apply(!is.na(cd_t_data), 2, sum)==0)*

Unsupervised Principal Component Analysis (PCA) Setup

In this step, we used Principal Component Analysis to reduce the number of genes down to the transformed variables that are a weighed sum of gene abundances. During this process, the algorithm does not know anything about which samples belong to which groups, but it accounts for as much variability in the data as possible.

Table 5. Principal Component Analysis (PCA) model input

	Description of the PCA process	Script used:
1.	This PCA version allows for missing values.	*pca_res = pcaMethods::pca(cd_t_data_final, method = "ppca")*
2.	Combine the sample information with the pca scores.	*pca_results = data.frame(cd_meta, PC1 = pca_res@scores[,1], PC2 = pca_res@scores[,2])*
3.	You can see that the control samples and treatment groups separate based on the new transformed variables.	*ggplot(data = pca_results, aes(x = PC1, y = PC2, color = Treatment)) +* *geom_point(size = 3) +* *theme_bw()*
4.	Look at the percentage of variability in the data that is accounted for by the new transformed variables	*pca_res@R2*100*
5.	Add this information to the plot (visualization).	*ggplot(data = pca_results, aes(x = PC1, y = PC2, color = Treatment)) +* *geom_point(size = 3) +* *theme_bw() +* *xlab("PC1 (34.8%)") +* *ylab("PC2 (15.9%)")*
6.	Look at which genes are contributing to the new variables	plot(1:ncol(cd_t_data_final),

	the most looking at the "loadings".	abs(pca_res@loadings[,1]), xlab = "Gene", ylab = "Absolute PC1 Loading Values")
7.	write out a csv of genes and their loadings.	*pca_loads = data.frame(Gene_ID = colnames(cd_t_data_final), PC1_Loading = pca_res@loadings[,1], Abs_PC1_Loading = abs(pca_res@loadings[,1]), PC2_Loading = pca_res@loadings[,2], Abs_PC2 = abs(pca_res@loadings[,2]))* *write.csv(pca_loads, file = "cd_pca_loadings.csv", row.names = F)*

Supervised Random Forest (RF) Setup

The Random Forest analysis general functioning of the algorithm is depicted in Table 6. In the original RF method, each tree is used as a standard classification tree that uses so-called Decrease of Gini Impurity as a splitting criterion and selects each predictor randomly from selected subsets [17]. Since there are several variants in RF, we describe each step in the table below.

Table 6. Random Forest model setup

	Description of the Random Forest process	**Script to use:**
1.	Initiate Random Forest analysis with As, Cd, Hg versus control samples.	*rf_res = randomForest(x = t_cd_imputed, y = as.factor(cd_meta$Treatment))*
2.	Make predictions using a validation strategy that holds some of the samples out a time and then tries to predict their group.	*pred = data.frame(Predicted = rf_res$predicted, Truth = cd_meta$Treatment)*
3.	Calculate the accuracy of the RF model.	*pred$Predicted == pred$Truth*
4.	Count the number of times this is true	*sum(pred$Predicted == pred$Truth*
5.	Divide by the number of samples to get accuracy.	*sum(pred$Predicted == pred$Truth)/nrow(pred)*
6.	Determine which genes are most important in our prediction by looking at the variable of importance metric.	*var_imp = rf_res$importance*
7.	Pull the genes that are important based by filter all genes with an importance level of 0.	*imp_res = data.frame(Gene = row.names(var_imp)[which(var_imp != 0)], Importance = var_imp[which(var_imp != 0)])*
8.	Plot the variables of importance, select top 30 #1.	*varImpPlot(rf_res, n.var = 30, main = 'Cd Subset Results')*
9.	Rank each gene based on importance, make the list so rank 1 is the highest level in importance.	*imp_res$Rank = nrow(imp_res) - rank(imp_res$Importance, ties.method = "random")*
10.	Plot variables of importance #2 [after ranking].	*lot(x = imp_res$Rank, y = imp_res$Importance, cex.main=1.5, xlab='Gene rank',ylab='Variable importance',cex.lab=1.5, pch=16,main='Cd subset_results')*
11.	Write out/ discard importance results with level of importance as 0.	*full_imp_res = data.frame(Gene = row.names(var_imp), Importance = var_imp)* *names(full_imp_res) = c("Gene", "Importance")* *full_imp_res$Rank = nrow(full_imp_res) - rank(full_imp_res$Importance, ties.method = "random")*

12. Determine top 25 genes (ranked).	*top25_data = t_cd_imputed[,gn.imp]*
13. Install heatmap packages and group data to visualize the top 25 most important genes.	*install.packages("RColorBrewer") library(RColorBrewer) my_group <- as.numeric(as.factor(cd_meta$Treatment)) colSide <- brewer.pal(2, "Set1")[my_group]*
14. Visualize results in a heatmap.	*heatmap(t(top25_data), Colv = NA, scale = "row", ColSideColors = colSide)*

Candidate biomarker genes analysis via Gene Ontology and KEGG Pathway Analysis

In this report, we compared different datasets, determined spatial variability among heavy metals and identified the candidate biomarker genes for As, Cd, and Hg using Random Forest analysis. First, the genes will be evaluated using the Database for Annotation, Visualization and Integrated Discovery (DAVID) to comprehensively study our selected genes giving biological meaning behind our large list of genes. Nevertheless, Gene Ontology (GO) will allow us to develop a comprehensive computational model of biological systems. These will range from the organismal to the molecular level, across many model species. Mapping the information on the function of genes is particularly useful for any large-scale molecular biology and genetics experiments in biomedical research. In addition, by further identifying many processes, researchers can generate their hypothesis of interest.

Secondly, we applied the gene clustering and ontology assignments into functional groups and performed the GO enrichment analysis to determine whether the heavy metal genes determined were significant. The p-value <0.05 should be considered statistically significant when defining the GO term enrichment analysis.

4. RESULTS AND ANALYSIS

Unsupervised Principal Component Analysis.

Zebrafish gene expression datasets were grouped and evaluated among control and treatment [As, Cd, and Hg] conditions. Our preliminary analysis revealed in Figure 3, illustrates the transformed variables in the grouped datasets accounted for. This analysis was helpful with respect to visually understanding the variability in the data. More, the algorithm reduced the number of genes to their transformed values using the weight sum of gene abundances. About 38 percent of the variance was passed onto Principal Component 1 (PC1) and about 15.9 percent of the variance was passed onto PC2. Overall, we see that the first component was effective in separating the data from control versus the treatment types. This analysis was successful in clustering similar samples based on their condition type. In general, the algorithm was successful at creating new components and in turn, a simpler description of the system was visualized.

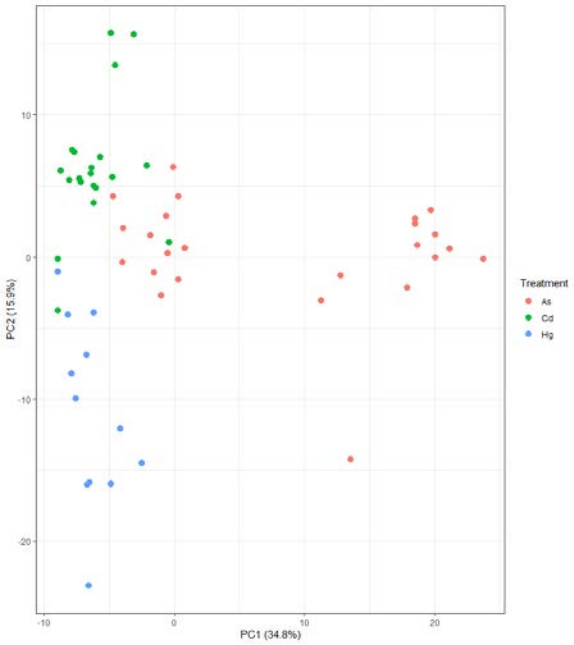

Figure 3. Results from Grouped unsupervised Principal Component Analysis (PCA). The Percentage of Variability Among arsenic, cadmium and mercury was determined to be 34.8% in PC1 and 15.9 % in PC2. This Process allowed for the Reduction in Genes to Their Transformed Values Using the Weight Sum of Gene Abundances.

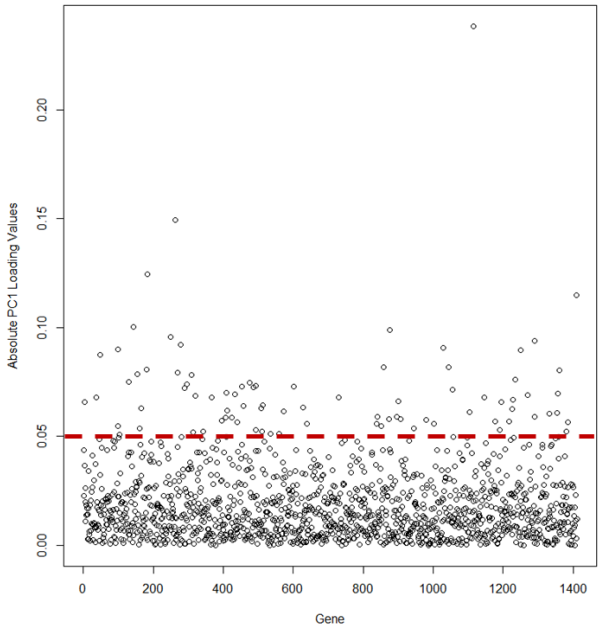

Figure 4. Contributing Genes from PC1 using Absolute Values for New Variable Identification. Dataset GSE3048, GSE30062, GSE41622, GSE41623 and GSE18861 were used in this analysis. Gene Loadings with A Cut-Off Criterion of < 0.05 were selected for RF biomarker identification.

Random Forest Analysis

We used the RF algorithm for [56 observations - 1411 predictors] and supported GSE41622 / GSE41623 [NA] data using the k-nearest neighbor (K-NN) classification algorithm. We trained the data and classified the results for As, Cd, and Hg using ordinance classification responses based on the variable class and observations [17]. Table 7 shows the classification results run times during training and accuracy performance. The genes were ranked according to variables of importance and further visualized selecting the top 25 most important genes using the Mean decrease in Gini [17]. The RF algorithm made predictions using a validation strategy that holds some of the samples out at a time and tries to predict their group.

Figure 5, Figure 7 and Figure 9 show the classification of genes selected during the RF classification process. The top 25 most important genes for As, Cd and Hg, identified using the RF algorithm, can be seen in Figure 6, Figure 8, and Figure 10, respectively. With respect to model accuracy and performance, the arsenic dataset was ranked highest in performance with a 96% accuracy. Mercury was second, reaching a successful training with 95% accuracy. Finally, the lowest accuracy among all was cadmium, reaching a 31% success model performance accuracy.

Table 7. Random Forest Model Training and Accuracy

Random Forest dataset training	Accuracy (%)	Truth Statements	Time (s)
As	96%	54	0.48
Cd	31%	10	3.88
Hg	95%	53	0.55

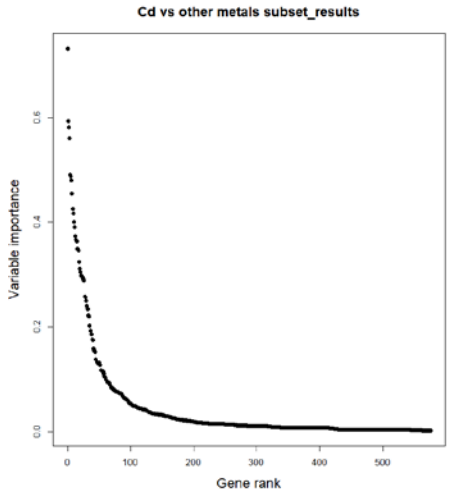

Figure 5. Cadmium gene subset list ranked according to variable of importance. The higher the ranking of importance the more predictive power each gene is considered.

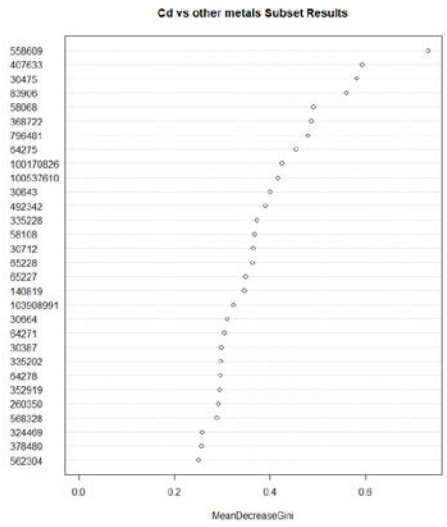

Figure 6. Cadmium subset gene list including the most important genes ranked according to the Mean Decrease Gini ordinance.

Table 8. Top 25 cadmium biomarkers identified using Random Forest (RF) analysis

Entrez ID	Gene Symbol	Gene name
30387	PSMB5	Proteasome (prosome, macropain) subunit, beta type, 5
30475	ERH	Enhancer of rudimentary homolog (Drosophila)
30643	Acat2	Acetyl-coa acetyltransferase 2
30664	GSK3A	Glycogen synthase kinase 3 alpha
30712	snap25a	Synaptosome-associated protein 25a
58068	Pc	Pyruvate carboxylase
58108	FTH1	Ferritin, heavy polypeptide 1
64271	epha4a	Eph receptor a4a
64275	supt5h	Suppressor of Ty 5 homolog (S. Cerevisiae)
64278	psmb7	Proteasome (prosome, macropain) subunit, beta type, 7
65227	Rgl2	Ral guanine nucleotide dissociation stimulator-like 2
65228	aspn	Asporin (LRR class 1)
83906	myl9l	Myosin, light polypeptide 9, like; bing3 like gene
140819	ptprf	Protein tyrosine phosphatase, receptor type, F
260350	HAS2	Hyaluronan synthase 2; similar to hyaluronan synthase 2
324469	wu:fc30c06	Wu:fc30c06
335202	wu:fk88f07	Wu:fk88f07
335228	wu:fk92d04	Wu:fk92d04
352919	selt1a	Selenoprotein T, 1a
368722	pdzk1ip1l	PDZK1 interacting protein 1, like
378480	pdgfaa	Platelet-derived growth factor alpha a
407633	si:ch211-191d7.6	Bat2-like protein
492342	mllt10	Myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog,

		Drosophila); translocated to, 10
558609	si:dkey-73n10.1	Si:dkey-73n10.1

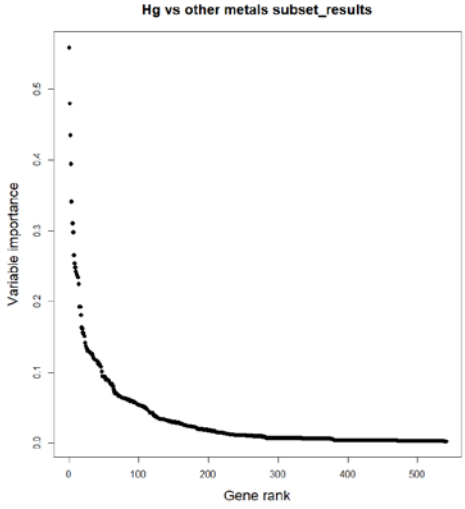

Figure 7. Mercury subset gene list ranked according to variable of importance. The higher the importance the more predictive power each gene contains.

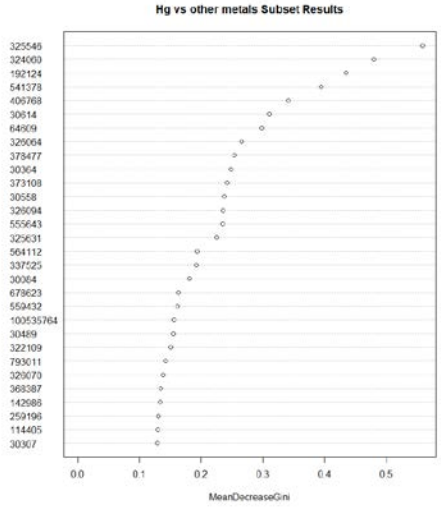

Figure 8. Mercury subset gene list including the most important genes ranked according to the mean Decrease Gini ordinance.

Table 9. Top 25 mercury biomarkers identified using Random Forest (RF) analysis

Entrez ID	Gene Symbol	Gene Name
30084	nme2b.2	Nucleoside diphosphate kinase-Z2
30307	jak2a	Janus kinase 2a
30364	fzd8c	Frizzled homolog 8c
30489	bfb	Complement component bfb
30558	PAX9	Paired box gene 9
30614	fxr	Farnesoid X-activated receptor
64609	atp1a2a	Atpase, Na+/K+ transporting, alpha 2a polypeptide
114405	cx27.5	Connexin 27.5
142986	ptk2.1	Protein tyrosine kinase 2.1
192124	Glra3	Glycine receptor, alpha 3
259196	MALT1	Mucosa associated lymphoid tissue lymphoma translocation gene 1
322109	wu	Wu:fb50c11
325631	usp24	Ubiquitin specific peptidase 24
373108	XIAP	X-linked inhibitor of apoptosis
378477	ANKRD6	Ankyrin repeat domain 6
406768	zgc	Zgc:55262
541378	Tufm	Zgc:110766
555643	hnrnpm	Heterogeneous nuclear ribonucleoprotein M
559432	si	Si:dkeyp-11g8.2
564112	MAGI2	Membrane associated guanylate kinase, WW and PDZ domain containing 2
678623	SLC11A2	Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 2

793011	rxrab	Retinoid x receptor, alpha b
100535764	-	-

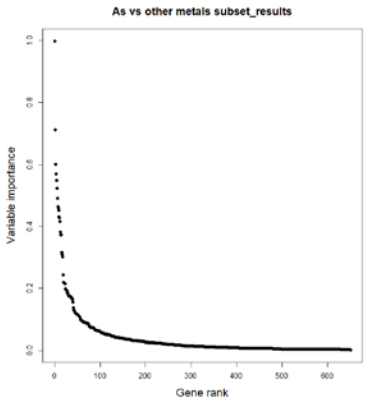

Figure 9. Arsenic subset gene list ranked according to variable of importance. The higher the gene is to 1 the more predictive power each gene contains.

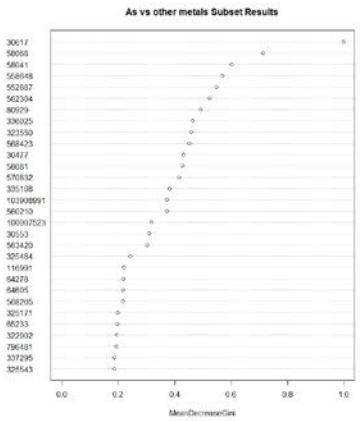

Figure 10. Arsenic subset gene list including the most important genes ranked according to the Mean Decrease Gini ordinance.

Table 10. Top 25 arsenic biomarkers identified using Random Forest (RF) analysis

Entrez ID	Gene Symbol	Gene Name
30477	TCP1	T-complex polypeptide 1
30553	SOD1	Superoxide dismutase 1, soluble
30617	Urod	Uroporphyrinogen decarboxylase
58041	HSPE1	Heat shock 10 protein 1 (chaperonin 10)
58068	Pc	Pyruvate carboxylase
58081	baxa	Bcl2-associated X protein, a
64278	psmb7	Proteasome (prosome, macropain) subunit, beta type, 7
64605	htatip2	HIV-1 Tat interactive protein 2
65233	six4.2	Sine oculis homeobox homolog 4.2
80929	id	Id:ibd2048
116991	UGDH	UDP-glucose dehydrogenase
322002	wu	Wu:fb40f06
323550	wu	Wu:fc02e03
325171	wu	Wu:fc57d08
325484	si	Si:dkeyp-86b9.2
325543	wu	Wu:fc85f10
335108	wu	Wu:fk69e07
336025	wu	Wu:fj43f12
337295	wu	Wu:fk14c11
552887	wu	Wu:fb63c04
558648	si	Si:ch211-15i6.2
560210	hsp70l	Heat shock cognate 70-kd protein, like; MCM5 minichromosome maintenance deficient 5 (S. Cerevisiae); heat shock cognate 70-kd protein; zgc:174006; similar to heat shock protein 8

562304	zgc	Similar to cytochrome P450, family 2, subfamily J, polypeptide 2; cytochrome P450 monooxygenase; similar to LOC562304 protein
563420	si	Si:ch211-245h14.1

Table 11. Arsenic results of Biological Functions (BF) gene set enrichment analysis performed on the top 25 ranked genes achieved through the Random Forest learning methods. Significant GO TERM enrichment processes p < 0.05.

Pathway ID	Biological Process (BP) Enriched Pathway	Count in gene set	*p*-value	FDR
GO:0010038~	Response to metal ion	4	1.03E-05	0.0107
GO:0010035~	Response to inorganic substance	4	2.42E-05	0.0253
GO:0046686~	Response to cadmium ion	3	1.28E-04	0.1338
GO:0051597~	Response to methylmercury	2	0.0145	14.2414
GO:0009410~	Response to xenobiotic stimulus	2	0.0203	19.3575
GO:0010033~	Response to organic substance	2	0.0681	52.2562

Table 12. Arsenic results of a pathway-based gene set analysis performed on the top 25 ranked genes achieved through the Random Forest learning methods.

Entrez ID	Gene Symbol	KEGG PATHWAY
116991	UGDH	dre00040: Pentose and glucuronate interconversions
		dre00053: Ascorbate and aldarate metabolism,
		dre00500: Starch and sucrose metabolism
		dre00520: Amino sugar and nucleotide sugar metabolism,
58081	baxa	dre04115: p53 signaling pathway
		dre04210: Apoptosis,
560210	hsp70l	dre03030: DNA replication,
		dre03040: Spliceosome,
		dre04010: MAPK signaling pathway,
		dre04110: Cell cycle,
		dre04144: Endocytosis,
64278	psmb7	dre03050: Proteasome,
58068	Si:ch211-15i6.2	dre00020: Citrate cycle (TCA cycle),
		dre00620: Pyruvate metabolism,
325484	Si:dkeyp-86b9.2	dre04130: SNARE interactions in vesicular transport,
30617	Urod	dre00860: Porphyrin and chlorophyll metabolism,

Table 13. Cadmium results of Biological Functions (BF) gene set enrichment analysis performed on the top 25 ranked genes achieved through the Random Forest learning methods. Significant GO TERM enrichment processes p < 0.05.

Pathway ID	Biological Process (BP) Enriched Pathway	Count in gene set	p-value	FDR
GO:0009987~	Cellular process	13	0.0242	26.7502
GO:0001947~	Heart looping	2	0.0521	49.2405
GO:0016051~	Carbohydrate biosynthetic process	2	0.0616	55.3135
GO:0003007~	Heart morphogenesis	2	0.0895	69.5271

GO:0035239~	Tube morphogenesis	2	0.0971	72.6053

Table 14. Cadmium results of a pathway-based gene set analysis performed on the top 25 ranked genes achieved through the Random Forest learning methods.

Entrez ID	Gene Symbol	KEGG PATHWAY
30643	Acat2	dre00071: Fatty acid metabolism,
		dre00072: Synthesis and degradation of ketone bodies,
		dre00280: Valine, leucine and isoleucine degradation,
		dre00310: Lysine degradation,
		dre00380: Tryptophan metabolism,
		dre00620: Pyruvate metabolism,
		dre00640: Propanoate metabolism,
		dre00650: Butanoate metabolism,
		dre00900: Terpenoid backbone biosynthesis,
58108	FTH1	dre00860: Porphyrin and chlorophyll metabolism,
83906	myl9l	dre04270: Vascular smooth muscle contraction,
		dre04510: Focal adhesion,
		dre04530: Tight junction,
		dre04810: Regulation of actin cytoskeleton,
378480	pdgfaa	dre04010: MAPK signaling pathway,
		dre04510: Focal adhesion,
		dre04540: Gap junction,
		dre04810: Regulation of actin cytoskeleton,
30387	PSMB5	dre03050: Proteasome,
64278	psmb7	dre03050: Proteasome,
58068	Pc	dre00020: Citrate cycle (TCA cycle),
		dre00620: Pyruvate metabolism,
30712	snap25a	dre04130: SNARE interactions in vesicular transport,

Table 15. Mercury results of Biological Functions (BF) gene set enrichment analysis performed on the top 25 ranked genes achieved through the Random Forest learning methods. Significant GO TERM enrichment processes $p < 0.05$.

Pathway ID	Biological Process (BP) Enriched Pathway	Count in gene set	p-value	FDR
GO:0010740~	Positive regulation of protein kinase cascade	2	0.0137	15.4435
GO:0045944~	Positive regulation of transcription from RNA polymerase II promoter	2	0.0206	22.2516
GO:0010627~	Regulation of protein kinase cascade	2	0.0240	25.4486
GO:0009967~	Positive regulation of signal transduction	2	0.0507	46.7403
GO:0010647~	Positive regulation of cell communication	2	0.0507	46.7403
GO:0035162~	Embryonic hemopoiesis	2	0.0524	47.8493
GO:0045893~	Positive regulation of transcription, DNA-dependent	2	0.0557	49.9990
GO:0051254~	Positive regulation of RNA metabolic process	2	0.0573	51.0407
GO:0006357~	Regulation of transcription from RNA polymerase II promoter	2	0.0704	58.6333

GO:0010628~	Positive regulation of gene expression	2	0.0960	70.4849
GO:0045941~	Positive regulation of transcription	2	0.0960	70.4849
GO:0010740~	Positive regulation of protein kinase cascade	2	0.0137	15.4435
GO:0045944~	Positive regulation of transcription from RNA polymerase II promoter	2	0.0206	22.2516
GO:0010627~	Regulation of protein kinase cascade	2	0.0240	25.4486
GO:0009967~	Positive regulation of signal transduction	2	0.0507	46.7403

Table 16. Mercury results of a pathway-based gene set analysis performed on the top 25 ranked genes achieved through the Random Forest learning methods.

Entrez ID	Gene Symbol	KEGG PATHWAY
64609	atp1a2a	dre04260: Cardiac muscle contraction,
30307	jak2a	dre04630: Jak-STAT signaling pathway,
		dre04920: Adipocytokine signaling pathway,
373108	XIAP	dre04120: Ubiquitin mediated proteolysis,
		dre04210: Apoptosis,
		dre04510: Focal adhesion,
		dre04621: NOD-like receptor signaling pathway,
192124	Glra3	dre04080: Neuroactive ligand-receptor interaction,
555643	hnrnpm	dre03040: Spliceosome,
564112	MAGI2	dre04530: Tight junction,
30084	nme2b.2	dre00230: Purine metabolism,
		dre00240: Pyrimidine metabolism,
142986	ptk2.1	dre04012: ErbB signaling pathway,
		dre04370: VEGF signaling pathway,
		dre04510: Focal adhesion,
		dre04810: Regulation of actin cytoskeleton,
678623	SLC11A2	dre04142:Lysosome,
406768	zgc	dre04210: Apoptosis,

5. CONCLUSION

In the present study, control and treated zebrafish samples were analyzed with respect to identifying discriminating markers associated with heavy metal exposure conditions. The response results from our experiment demonstrated that the liver proves to be a sensitive organ/indicator of metal toxicity in adult zebrafish. Finally, we proved that Random Forest (RF) analysis was effective in discriminating heavy metal biomarkers.

First, the algorithms for unsupervised Principal Component Analysis (PCA) reduced the number of genes to their transformed values using the weight sum of gene abundances. For the grouped analysis, about 38 percent of the variance was passed onto Principal Component 1 (PC1) and about 15.9 percent of the variance was passed onto PC2. We also see that the first component was effective in separating the data from control versus the treatment types. PCA effectively clustered similar samples based on their sample treatment types.

The use of supervised a Random Forest classification algorithm classified the grouped series of 56 observations - 1411 predictors, and supported GSE41622 / GSE41623 [NA] data using the k-nearest neighbor (K-NN) classification algorithm. We evaluated the model performance while during training data. We confirmed that GSE41622 and GSE41623 did not achieve optimal accuracy in model performance; however GSE3048, 30062 and GSE18861 achieved 95 and 96 percentile model performances, respectively.

We used the Database for Annotation, Visualization and Integrated Discovery (DAVID) software for a more systematic functional interpretation for arsenic, cadmium, and mercury biomarkers. Importantly we chose DAVID because it considers the functional gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways pipeline. The functional annotation and enrichment analysis performed with DAVID are described in Tables 10-16. In order to investigate the enriched biological functions and KEGG pathways between arsenic, cadmium and mercury, DAVID analysis was performed separately using the biomarkers identified in Table 8-Table 10.

First, we grouped the arsenic subset of genes and computed the GO term enrichment analysis. We identified that among the topmost Biological Processes (BP), response to inorganic substance (GO: 0010035) or chemicals and responses to metal ions (GO: 0010038) were significantly enriched processes. Using the same subset of genes, KEGG Pathway terms enrichment was assessed, however no enrichment could be found at *p-value* <0.05. The top-25 upmost ranked genes at an FDR<10^{-8} are presented in Table 12.

The cadmium subset of genes reflected few enriched processes. Noteworthy, the significant process in the cadmium markers was cellular processes (GO: 009987). Many genes participated

in the enrichment process of cadmium. It was estimated that the Pyruvate metabolism (6.5E-2) and Proteasome signaling (9.0E-2) pathways were enriched.

Lastly, mercury-enriched processes were identified in positive regulation of protein kinase cascade (GO:0010740), positive regulation of transcription from RNA polymerase II promoter (GO: 0045944), and regulation of protein kinase cascades (GO: 0010627). Like arsenic, enrichment analysis using the top 25 ranked genes was assessed, however no enrichment could be found at p-value <0.05. The top-25 upmost ranked genes at an FDR<10^{-8} are presented in Table 16.

In conclusion, Random Forest analysis determined the biomarkers for arsenic, cadmium and mercury and associated novel pathways derived from the gene expression data. In addition, our results show that the zebrafish transcriptome responds to treatment in a sensitive manner. The identification of heavy metal biomarkers in response to As, Cd, and Hg proves to be a fruitful molecular approach to strengthen traditional environmental measurements, detailing a comprehensive pathway framework to monitor accumulated metals in surface waters.

6. REFERENCES

1. Halverson N. Final Report on the Aquatic Mercury Assessment Study. 2008;(September). Available from: http://www.osti.gov/servlets/purl/939852-ax1Udk/
2. Dobson S, Howe PD, Floyd P. Mono- and disubstituted methyltin, butyltin, and octyltin compounds. IPCS Concise Int Chem Assess Doc. 2006;(73).
3. Singh R, Gautam N, Mishra A, Gupta R. Heavy metals and living systems: An overview. Indian J Pharmacol [Internet]. 2011 May [cited 2018 Nov 20];43(3):246–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21713085
4. He L, Gao B, Luo X, Jiao J, Qin H, Zhang C, et al. Health risk assessment of heavy metals in surface water near a uranium tailing pond in Jiangxi Province, South China. Sustain. 2018;10(4).
5. The A. discovery and development Zebrafish as a model for translational neurobiology : Implications for drug discovery and development.
6. Yang Q, Salim L, Yan C, Gong Z. Rapid Analysis of Effects of Environmental Toxicants on Tumorigenesis and Inflammation Using a Transgenic Zebrafish Model for Liver Cancer. Mar Biotechnol (NY) [Internet]. 2019 Jun [cited 2019 Sep 9];21(3):396–405. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30852708
7. Peterson SM, Zhang J, Webr G, Freeman JL. Global gene expression analysis reveals dynamic and developmental stage-dependent enrichment of lead-induced neurological gene alterations. Environ Health Perspect. 2011 May;119(5):615–21.
8. Streisinger G. ICPEMC working paper 4/2 extrapolations from species to species and from various cell types in assessing risks from chemical mutagens. Mutat Res Genet Toxicol. 1983 Jan 1;114(1):93–105.
9. Mathavan S, Lee SGP, Mak A, Miller LD, Murthy KRK, Govindarajan KR, et al. Transcriptome Analysis of Zebrafish Embryogenesis Using Microarrays. PLoS Genet [Internet]. 2005;1(2):e29. Available from: http://dx.plos.org/10.1371/journal.pgen.0010029
10. Bambino K, Chu J. Zebrafish in Toxicology and Environmental Health. In: Current Topics in Developmental Biology. Academic Press Inc.; 2017. p. 331–67.
11. Raghavachari N. Microarray technology: Basic methodology and application in clinical research for biomarker discovery in vascular diseases. Methods Mol Biol [Internet]. 2013 [cited 2020 Sep 27];1027:47–84. Available from: https://link.springer.com/protocol/10.1007/978-1-60327-369-5_3
12. Niu C, Jiang M, Li N, Cao J, Hou M, Ni D-A, et al. Integrated bioinformatics analysis of As, Au, Cd, Pb and Cu heavy metal responsive marker genes through Arabidopsis thaliana GEO datasets. 2019 [cited 2019 Oct 15]; Available from: http://doi.org/10.7717/peerj.6495
13. Martin EM, Fry RC. Environmental Influences on the Epigenome: Exposure-Associated DNA Methylation in Human Populations. Annu Rev Public Heal [Internet]. 2018 [cited 2020 Feb 5];39:309–33. Available from: https://doi.org/10.1146/annurev-publhealth-
14. Zapp Sluis M, Boswell KM, Chumchal MM, Wells RJD, Soulen B, Cowan JH. Regional variation in mercury and stable isotopes of red snapper (Lutjanus campechanus) in the northern gulf of Mexico, USA. Environ Toxicol Chem. 2013;32(2):434–41.
15. Lam SH, Winata CL, Tong Y, Korzh S, Lim WS, Korzh V, et al. Transcriptome kinetics of arsenic-induced adaptive response in zebrafish liver. Physiol Genomics [Internet].

2006;27(3):351–61. Available from: http://physiolgenomics.physiology.org/cgi/doi/10.1152/physiolgenomics.00201.2005
16. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature [Internet]. 2013 Apr 17 [cited 2018 Oct 30];496(7446):498–503. Available from: http://www.nature.com/doifinder/10.1038/nature12111
17. Boulesteix A-L, Janitza S, Kruppa J, König IR. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics pre-review versio [Internet]. 2012 [cited 2020 Sep 27]. Available from: http://www.stat.uni-muenchen.de
18. Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, et al. Pathway analysis using random forests classification and regression. Bioinformatics [Internet]. 2006 Aug 15 [cited 2020 Aug 25];22(16):2028–36. Available from: http://bioinformatics.med.yale.edu/pathway-analysis/rf.htmhttp://bioinformatics.med.yale.edu/pathway-analysis/rf.htm
19. Ram M, Najafi A, Shakeri MT. Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest. J Pathol Iran J Pathol. 2017;12(4):339–47.
20. Edwards PG, Gaines KF, Bryan AL, Novak JM, Blas SA. Are U, Ni, and Hg an Environmental Risk within a RCRA/CERCLA Unit on the U.S. Department of Energy's Savannah River Site? Hum Ecol Risk Assess. 2014;20(6):1565–89.
21. Goyer R, Golub M, Choudhury H, Hughes M, Kenyon E, Stifelman M. Issue Paper on the Human Health Effects of Metals [Internet]. 2046 [cited 2019 May 22]. Available from: https://www.epa.gov/sites/production/files/2014-11/documents/human_health_effects.pdf
22. Siew HL, Winata CL, Tong Y, Korzh S, Wen SL, Korzh V, et al. Transcriptome kinetics of arsenic-induced adaptive response in zebrafish liver. Physiol Genomics. 2006 Nov 27;27(3):351–61.
23. Zhang X, Ung CY, Lam SH, Ma J, Chen YZ, Zhang L, et al. Toxicogenomic Analysis Suggests Chemical-Induced Sexual Dimorphism in the Expression of Metabolic Genes in Zebrafish Liver. PLoS One [Internet]. 2012 Dec 18 [cited 2020 Aug 27];7(12). Available from: https://pubmed.ncbi.nlm.nih.gov/23272195/
24. Ung CY, Lam SH, Zhang X, Li H, Zhang L, Li B, et al. Inverted Expression Profiles of Sex-Biased Genes in Response to Toxicant Perturbations and Diseases. PLoS One. 2013 Feb 14;8(2).
25. Ung CY, Lam SH, Hlaing MM, Winata CL, Korzh S, Mathavan S, et al. Mercury-induced hepatotoxicity in zebrafish: In vivo mechanistic insights from transcriptome analysis, phenotype anchoring and targeted gene expression validation. BMC Genomics. 2010 Mar 30;11(1).
26. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res [Internet]. 2012 Nov 26 [cited 2019 May 16];41(D1):D991–5. Available from: http://academic.oup.com/nar/article/41/D1/D991/1067995/NCBI-GEO-archive-for-functional-genomics-data
27. Tilton SC, Tal TL, Scroggins SM, Franzosa JA, Peterson ES, Tanguay RL, et al. Bioinformatics resource manager v2.3: An integrated software environment for systems biology with microRNA and cross-species analysis tools. BMC Bioinformatics [Internet].

2012 Nov 23 [cited 2020 Aug 5];13(1):311. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-311
28. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane H, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol [Internet]. 2003 [cited 2019 Nov 11];4(9):R60. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-4-9-r60
29. Srizon AY, Hasan MAM. Prognostic Biomarker Identification for Pancreatic Cancer by Analyzing Multiple mRNA Microarray and microRNA Expression Datasets. 5th Int Conf Comput Commun Chem Mater Electron Eng IC4ME2 2019. 2019;(July).

APPENDIX A.

Table 17. Heavy metal dataset demographics according to time, GSE Accession, group and dose.

SampleID	GSE Accession	TimePoint	Group	Treatment	Concentration
GSM67011	GSE3048	24	Treated	As	15 µg/L
GSM67012	GSE3048	24	Treated	As	15 µg/L
GSM67013	GSE3048	24	Treated	As	15 µg/L
GSM67014	GSE3048	48	Treated	As	15 µg/L
GSM67015	GSE3048	48	Treated	As	15 µg/L
GSM67016	GSE3048	48	Treated	As	15 µg/L
GSM67017	GSE3048	8	Treated	As	15 µg/L
GSM67018	GSE3048	8	Treated	As	15 µg/L
GSM67019	GSE3048	8	Treated	As	15 µg/L
GSM67020	GSE3048	96	Treated	As	15 µg/L
GSM67021	GSE3048	96	Treated	As	15 µg/L
GSM67022	GSE3048	96	Treated	As	15 µg/L
GSM67023	GSE3048	24	Control	Control	
GSM67024	GSE3048	24	Control	Control	
GSM67025	GSE3048	24	Control	Control	
GSM67026	GSE3048	48	Control	Control	
GSM67027	GSE3048	48	Control	Control	
GSM67028	GSE3048	48	Control	Control	
GSM67029	GSE3048	8	Control	Control	
GSM67030	GSE3048	8	Control	Control	
GSM67031	GSE3048	8	Control	Control	
GSM67032	GSE3048	96	Control	Control	
GSM67033	GSE3048	96	Control	Control	
GSM67034	GSE3048	96	Control	Control	
GSM744231	GSE30062	8	Treated	As	15 µg/L
GSM744232	GSE30062	8	Treated	As	15 µg/L
GSM744233	GSE30062	8	Treated	As	15 µg/L
GSM744234	GSE30062	24	Treated	As	15 µg/L
GSM744235	GSE30062	24	Treated	As	15 µg/L
GSM744236	GSE30062	24	Treated	As	15 µg/L
GSM744237	GSE30062	48	Treated	As	15 µg/L
GSM744238	GSE30062	48	Treated	As	15 µg/L
GSM744239	GSE30062	48	Treated	As	15 µg/L
GSM744240	GSE30062	96	Treated	As	15 µg/L
GSM744241	GSE30062	96	Treated	As	15 µg/L
GSM744242	GSE30062	96	Treated	As	15 µg/L
GSM744105	GSE30062; GSE41622	8	Control	Control	
GSM744106	GSE30062; GSE41622	8	Control	Control	
GSM744107	GSE30062; GSE41622	8	Control	Control	
GSM744108	GSE30062; GSE41622	24	Control	Control	
GSM744109	GSE30062; GSE41622	24	Control	Control	
GSM744110	GSE30062; GSE41622	24	Control	Control	
GSM744111	GSE30062; GSE41622	48	Control	Control	
GSM744112	GSE30062; GSE41622	48	Control	Control	
GSM744113	GSE30062; GSE41622	48	Control	Control	
GSM744114	GSE30062; GSE41622	96	Control	Control	
GSM744115	GSE30062; GSE41622	96	Control	Control	
GSM744116	GSE30062; GSE41622	96	Control	Control	
GSM1020289	GSE41622	8	Treated	Cd	30 µg/L
GSM1020290	GSE41622	8	Treated	Cd	30 µg/L

GSM1020291	GSE41622	24	Treated	Cd	30 µg/L
GSM1020292	GSE41622	24	Treated	Cd	30 µg/L
GSM1020293	GSE41622	48	Treated	Cd	30 µg/L
GSM1020294	GSE41622	48	Treated	Cd	30 µg/L
GSM1020295	GSE41622	48	Treated	Cd	30 µg/L
GSM1020296	GSE41622	96	Treated	Cd	30 µg/L
GSM1020297	GSE41622	96	Treated	Cd	30 µg/L
GSM1020298	GSE41622	96	Treated	Cd	30 µg/L
GSM1020299	GSE41623	8	Treated	Cd	30 µg/L
GSM1020300	GSE41623	8	Treated	Cd	30 µg/L
GSM1020301	GSE41623	24	Treated	Cd	30 µg/L
GSM1020302	GSE41623	24	Treated	Cd	30 µg/L
GSM1020303	GSE41623	48	Treated	Cd	30 µg/L
GSM1020304	GSE41623	48	Treated	Cd	30 µg/L
GSM1020305	GSE41623	48	Treated	Cd	30 µg/L
GSM1020306	GSE41623	96	Treated	Cd	30 µg/L
GSM1020307	GSE41623	96	Treated	Cd	30 µg/L
GSM1020308	GSE41623	96	Treated	Cd	30 µg/L
GSM744057	GSE41623	8	Control	Control	
GSM744058	GSE41623	8	Control	Control	
GSM744059	GSE41623	8	Control	Control	
GSM744060	GSE41623	24	Control	Control	
GSM744061	GSE41623	24	Control	Control	
GSM744062	GSE41623	24	Control	Control	
GSM744063	GSE41623	48	Control	Control	
GSM744064	GSE41623	48	Control	Control	
GSM744065	GSE41623	48	Control	Control	
GSM744066	GSE41623	96	Control	Control	
GSM744067	GSE41623	96	Control	Control	
GSM744068	GSE41623	96	Control	Control	
GSM467510	GSE18861	8	Treated	Hg	200 µg/L
GSM467511	GSE18861	8	Treated	Hg	200 µg/L
GSM467512	GSE18861	8	Treated	Hg	200 µg/L
GSM467513	GSE18861	24	Treated	Hg	200 µg/L
GSM467514	GSE18861	24	Treated	Hg	200 µg/L
GSM467515	GSE18861	24	Treated	Hg	200 µg/L
GSM467516	GSE18861	48	Treated	Hg	200 µg/L
GSM467517	GSE18861	48	Treated	Hg	200 µg/L
GSM467518	GSE18861	48	Treated	Hg	200 µg/L
GSM467519	GSE18861	96	Treated	Hg	200 µg/L
GSM467520	GSE18861	96	Treated	Hg	200 µg/L
GSM467521	GSE18861	96	Treated	Hg	200 µg/L
GSM467498	GSE18861	8	Control	Control	
GSM467499	GSE18861	8	Control	Control	
GSM467500	GSE18861	8	Control	Control	
GSM467501	GSE18861	24	Control	Control	
GSM467502	GSE18861	24	Control	Control	
GSM467503	GSE18861	24	Control	Control	
GSM467504	GSE18861	48	Control	Control	
GSM467505	GSE18861	48	Control	Control	
GSM467506	GSE18861	48	Control	Control	
GSM467507	GSE18861	96	Control	Control	
GSM467508	GSE18861	96	Control	Control	
GSM467509	GSE18861	96	Control	Control	