The Molecular Toxicity Identification Evaluation (mTIE) Approach Predicts Chemical Exposure in Daphnia magna.

ARTICLE *in* ENVIRONMENTAL SCIENCE & TECHNOLOGY · JULY 2013

Impact Factor: 5.33 · DOI: 10.1021/es402819c · Source: PubMed

9 AUTHORS, INCLUDING:

Philipp Antczak
University of Liverpool
25 PUBLICATIONS **272** CITATIONS
SEE PROFILE

Leona D Scanlan
National Institute of Standards and Technology
7 PUBLICATIONS **67** CITATIONS
SEE PROFILE

Helen Poynton
University of Massachusetts Boston
26 PUBLICATIONS **749** CITATIONS
SEE PROFILE

Francesco Falciani
University of Liverpool
100 PUBLICATIONS **3,529** CITATIONS
SEE PROFILE

Environmental Science & Technology

Article

pubs.acs.org/est

Molecular Toxicity Identification Evaluation (mTIE) Approach Predicts Chemical Exposure in *Daphnia magna*

Philipp Antczak,[†] Hun Je Jo,[‡] Seonock Woo,[¶] Leona Scanlan,[¶] Helen Poynton,[∥] Alex Loguinov,[¶] Sarah Chan,[⊥] Francesco Falciani,*[,†,#] and Chris Vulpe*[,¶,#]

[†]Centre for Computational Biology and Modelling, Institute for Integrative Biology, University of Liverpool, L69 7ZB Liverpool, U.K.,

[‡]Yeongsan River Basin Environmental Office, Gyesuro-31, Seo-gu, Gwangju 502-862, Korea,

[¶]Nutritional Sciences and Toxicology & Berkeley Institute of the Environment, University of California, Berkeley, California 94720, United States,

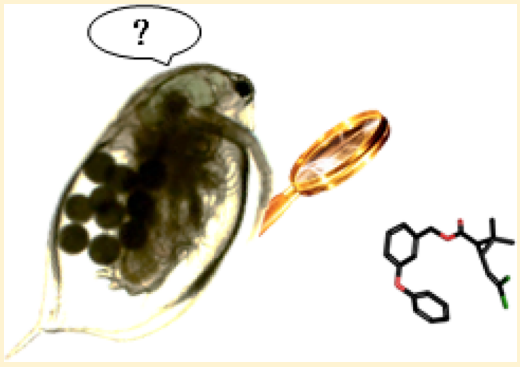
[§]Laboratory of Ecotoxicogenomics, South Sea Environment Research Division, Korea Institute of Ocean Science & Technology, Jangmok1 gil-41 Geoje-si 656-830, Korea,

[∥]Department of Environmental, Earth and Ocean Sciences, University of Massachusetts, Boston, Massachusetts 02125, United States, and

[⊥]School of Medicine, University of California, San Diego, California 92093, United States

Supporting Information

ABSTRACT: *Daphnia magna* is a bioindicator organism accepted by several international water quality regulatory agencies. Current approaches for assessment of water quality rely on acute and chronic toxicity that provide no insight into the cause of toxicity. Recently, molecular approaches, such as genome wide gene expression responses, are enabling an alternative mechanism based approach to toxicity assessment. While these genomic methods are providing important mechanistic insight into toxicity, statistically robust prediction systems that allow the identification of chemical contaminants from the molecular response to exposure are needed. Here we apply advanced machine learning approaches to develop predictive models of contaminant exposure using a *D. magna* gene expression data set for 36 chemical exposures. We demonstrate here that we can discriminate between chemicals belonging to different chemical classes including endocrine disruptors and inorganic and organic chemicals based on gene expression. We also show that predictive models based on indices of whole pathway transcriptional activity can achieve comparable results while facilitating biological interpretability.



INTRODUCTION

Freshwater habitats throughout the world are endangered due to human activity including widespread contamination with chemicals. *D. magna* are important sentinel species in toxicology due to their wide geographic distribution, central role in freshwater food webs, ability to adapt to a range of habitats, and sensitivity to anthropogenic chemicals.[1] Assessing the impact of chemical exposures on the aquatic environment currently rely on measures of acute and chronic toxicity in several species, often including *Daphnia* species. While these methods can identify that toxicity is present at a site, they do not identify the key chemicals underlying the toxicity. Current Toxicity Identification and Evaluation (TIE) methods rely on various treatment and fractionation techniques in combination with repeated toxicity tests (in an attempt to identify key toxicants). While these approaches can provide important information, they are time-consuming and expensive and often are not able to identify a cause of toxicity. An approach that is more sensitive, specific, timely, and cost-effective would greatly facilitate the TIE process. We suggest that a genomic approach, molecular TIE or mTIE, could allow for rapid contaminant evaluation in environmental samples and provide a robust, cost-effective alternative to current TIE methods.

Here we describe the development of an approach in *D. magna* for prediction of chemical class exposure, which could ultimately be used in an environmental risk assessment. We demonstrate for the first time, that it is possible to predict chemical class from the transcriptional response of adult daphnids, following exposure to sublethal concentrations of a given toxicant. We show that predictive models can be developed based on both individual gene responses to toxicants and based on whole-pathway activity indices. While both

Received: January 14, 2013
Accepted: July 22, 2013
Published: July 22, 2013

approaches provide comparable predictive accuracy, more biologically interpretable results are obtained using pathway based models.

MATERIALS AND METHODS

***Daphnia* Culturing and Toxicity Testing.** *D. magna* were obtained from Aquatic Research Organisms (Hampton, NH) and cultured at 30 animals/3L in 20 °C in modified COMBO medium[2] in an 8 h dark 16 h light cycle. Culturing medium was renewed three times a week, and Pseudokirchneriella subcapitata (formerly *Selenastrum capricornutum*) and YCT (Yeast, Cereal Leaves, Tetramin) were added for food (food concentration 30 million cells/mL - 1 mL of food added per liter). For acute toxicity test, neonates (<24 h) were exposed to chemicals using a modified U.S. Environmental Protection Agency (EPA) guideline.[3] The chemicals used for exposure in this study are shown in Table 1. Each toxicity test consisted of five or six concentrations and a control with four replicates. Solvents, such as ethanol (Fisher Sci., NJ, USA), acetone (Fisher Sci., NJ, USA), methanol (Fisher Sci., NJ, USA), and dimethyl sulfoxide (Fisher Sci., NJ, USA), were used for dissolving chemicals. Five neonates were placed in a 50 mL vessel containing 30 mL of test solution, and then the LC50 value was calculated using trimmed Spearman-Karber (TSK v1.5 U.S. EPA) and Probit (v1.5 U.S. EPA) methods after 24 h.

Table 1. Chemicals Represented in This Data Set Sorted by Classification

class	compounds
inorganic	cadmium, nickel, copper, selenium, zinc, manganese, arsenic, silver, chromium
endocrine disruptors	pyriproxyfen,[19−21] ponasterone A,[22] methyl farnesoate,[20,23] toxaphene,[24] beta-estradiol,[25,26] aroclor 1242,[19,27] 20-hydroxyecdysone,[26] methoxychlor,[28] nonylphenol[19,25]
organic	permethrin, bifenthrin, λ-cyhalothrin, diazinon, parathion, chlorpyrifos, toluene, phenol, beta-benzene hexachloride, dichlorobenzene, phenanthrene, atrazine, methyl tert-butyl ether, chloroform, acrylonitrile, bis(2-ethylhexyl)phthalate, trichloroethylene, 2-chloroethyl vinyl ether

Exposure to Chemicals, RNA Extraction, and Microarray Hybridization. Twenty daphnids (two-week old) were exposed to 1 L of COMBO medium containing 1/10 LC50 concentration of each chemical with four replicates. After 24 h exposure, total RNA was extracted with Trizol (Invitrogen, CA, USA) according to the manufacturer's protocol. No DNase treatment was performed. The quantity and quality of RNA were checked by using spectrophotometer (BioRad, USA) and agarose-formaldehyde gel electrophoresis. The extracted RNA was stored at −80 °C until the use. The custom *D. magna* oligonucleotide microarray containing 14,338 genes was manufactured by Agilent, and hybridization was performed according to the manufacture's protocol (AMAID: 023710, GPL15139).[4] Briefly, the RNA was reverse-transcribed into complementary DNA using oligo dT-promoter primer, and then, the labeled and amplified RNA was synthesized from cDNA in the presence of T7 RNA polymerase and cyanine 3-labeled CTP, followed by purification of cRNA using RNeasy Mini Kit (Qiagen, CA, USA). Nanodrop readings of all RNA and cRNA samples were used for quality control as per Agilent Labeling protocol. The yield and specific activity of cRNA were determined prior to hybridization at 65 °C during 17 h. The scanning and feature extraction were performed using GenePix 4000B (Axon, USA) and GenePix Pro 6.0 software (Axon, USA).

Data Normalization and Clustering Procedures. Raw microarray data was read and normalized against their respective controls using loess within the statistical environment R using the marray[5] and limma[6] packages. Genes that were lowly expressed (single color median log intensity across all samples <5) across all samples were removed (1816 probes). Quality of microarrays was verified through exploratory analysis. We inspected the homogeneity of the distribution of signal and noise in the arrays, the consistency of the individual samples signal distributions, and the presence of outliers. These methods were used as implemented in the Web-based tool Babelomics,[7] and all samples passed the scrutiny. Clustering of data was performed on the number of significantly differentially expressed genes at FDR < 0.05. A distance matrix was derived by calculating the Jaccard's Index of overlap which is defined as the ratio between intersect and union of two lists. This was then used as an input to a hierarchical clustering algorithm (using the R hclust function). The agglomeration method was chosen by identifying the highest correlating cophenetic matrix which resulted in "average".

CLASSIFICATION

In order to develop and validate predictive models we have employed two separate strategies. The first determines the accuracy of a model (developed using all chemicals in the set) on an independent set of experimental exposures. The second parametrizes a model on a subset of chemicals and then tests its accuracy on the remaining chemicals.

Model Development and Computing Classification Accuracy on an Independent Set of Exposures. To identify genes/pathways which were predictive of chemical class, a variable selection approach coupled with a random forest classification algorithm[8] was used as implemented in the GALGO[9] package in the statistical environment R. This procedure integrates an efficient multivariate variable selection procedure designed to optimize a small subset of predictive variables and an advanced classification algorithm that minimizes the possibility of overtraining with an in-built out-of-bag cross-validation procedure.[10] The modeling procedure was initialized using the default settings in GALGO[9] with a model size of 10. The classification accuracy was estimated as follows: Data were first split into training and test data sets, representing respectively 2/3 and 1/3 of the original data. Both training and test sets represented all chemicals but included independent biological exposures. In order to avoid over-training, models were trained with a second level split (2/3 training 1/3 test) of the training data. Up to 1000 independent models were collected, and a representative model developed using a forward selection procedure, as described in ref 9. This approach ranks the model variables (genes or pathway activity indices) on the basis of the frequency they appeared in the population. The top 50 most frequent variables are then incrementally tested, by adding each variable one by one starting with the most frequent. The model with the smaller number of variables and the higher accuracy is then selected as the final representative model.

Computing Chemical-Specific Classification Accuracy. To test whether our classification models are able to predict chemicals that have not been used to train the model itself, we employed a cross-validation (CV) procedure. In this approach, a chemical from each class is removed (2 from a two-class

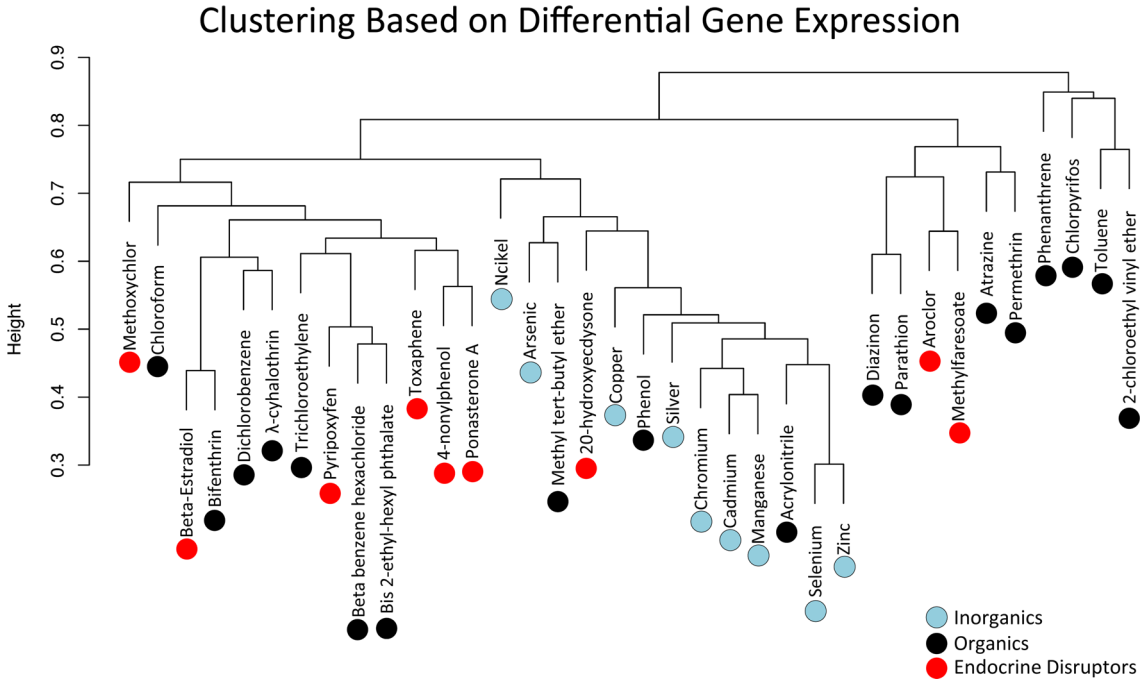


Figure 1. Hierarchical clustering of the effect of the chemicals on *D. magna*. The distance here is defined as the overlap of differentially expressed genes in each pair as defined by the Jaccard's Index of overlap. Inorganic compounds cluster in close proximity of each other, while endocrine disruptors have a more heterogeneous but still similar expression profile.

problem and 3 from a three-class problem) to generate a training set for the random Forest classifier. We then evaluated the ability of this classifier to predict the chemical class of the removed chemicals. We repeat this for every single possible combination for the 2 class (243 combinations) and 3 class (1458 combinations) problems and calculate the accuracy based on all instances where each chemical was removed. The results are then represented in the form of a barplot where each bar indicates the accuracy (across all tests) for that particular chemical (see Figures S1−S4).

Computing Indices of Pathway Activity. Due to the availability of the *D. pulex* genome all coding regions have been included in the KEGG database.[11] We first mapped, by protein blast, all sequences on the *D. magna* microarray to the *D. pulex* genome. We identified 4958 *D. magna* homologues in *D. pulex* genome ($e < 10^{-4}$). Of these, 1425 genes mapped on 114 of the 117 *D. pulex* KEGG pathways. This represents a 38% gene coverage (1425 *D. magna* genes out of 3846 *D. pulex* genes mapped on KEGG) and a 97% pathway level coverage of the *D. pulex* KEGG pathway system. To make sure that pathway indices were representative of a significant fraction of genes within a pathway we only considered pathways where 5 genes or more were available in the input data set. This reduced the list of genes to 1402 KEGG annotated genes, representing 95 distinct pathways (retention of 81% of available *D. pulex* pathway space). This provided a broad spectrum of functions represented in our data set (see Tables S3 and S4 for the list of pathways). Indices of pathway activity were then computed as the first three principal components (PC) of the gene expression profiles representative of each KEGG pathway using the prcomp function within the statistical environment R.[12] These represented at least 70% of the variance present in the data. This procedure generated a new derived data set with 285 pathway components and 144 samples. This approach was previously demonstrated to be an effective strategy to improve biological interpretability and reduce the computational space.[13]

RESULTS

Choice of Chemicals and Definition of Chemical Groups. Chemicals were selected to include a variety of organic and inorganic compounds of environmental concern. We initially classified the chemicals based on the organic or inorganic nature. Within the organic group a number of compounds have been shown to have endocrine disrupting properties.[14−17] A chemical was deemed to be an endocrine disruptor if experimental evidence of moulting or male production in *D. magna* has been previously established (Supporting Information Table S1).[18] Based on this simplified classification we decided to focus on two specific comparisons. The first comparison focuses on developing predictive models for inorganic versus all organic compounds (including endocrine disruptors), whereas the second comparison focuses on inorganic vs endocrine disruptors vs organic compounds.

The Transcriptional State of *Daphnia magna* Defines Contaminant Exposure Class. We determined the 24 h acute toxicity for all 36 chemicals in *D. magna* neonates (Table S2) and subsequently derived the gene expression profile in *D. magna* exposed to the 1/10 LC_{50} dose for each contaminant. The chosen dose represented a concentration at which most of the chemicals (35/36) showed no toxicity (below NOEC, see Supporting Information Table S2). Interestingly, despite the absence of actue toxicity we could detect a significant gene expression response. Clustering of the samples based on overlap of differential gene expression revealed a cluster that included all inorganic chemicals and four organic compounds, one of which was classified as an endocrine disruptor (Figure 1). Visual inspection of the dendrogram showed little discrimination between organic and endocrine disrupting compounds. This inability to discriminate chemical classes on

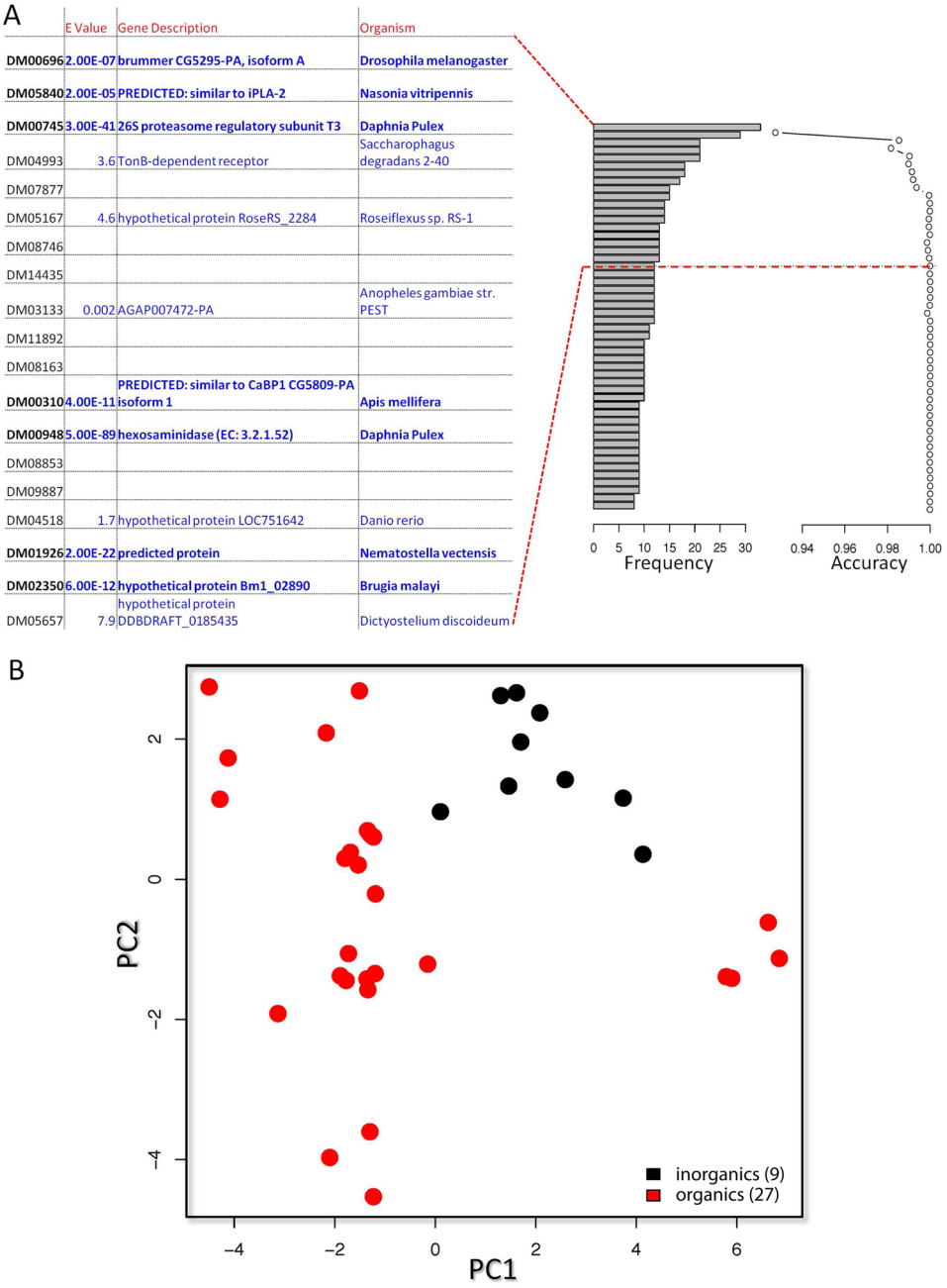
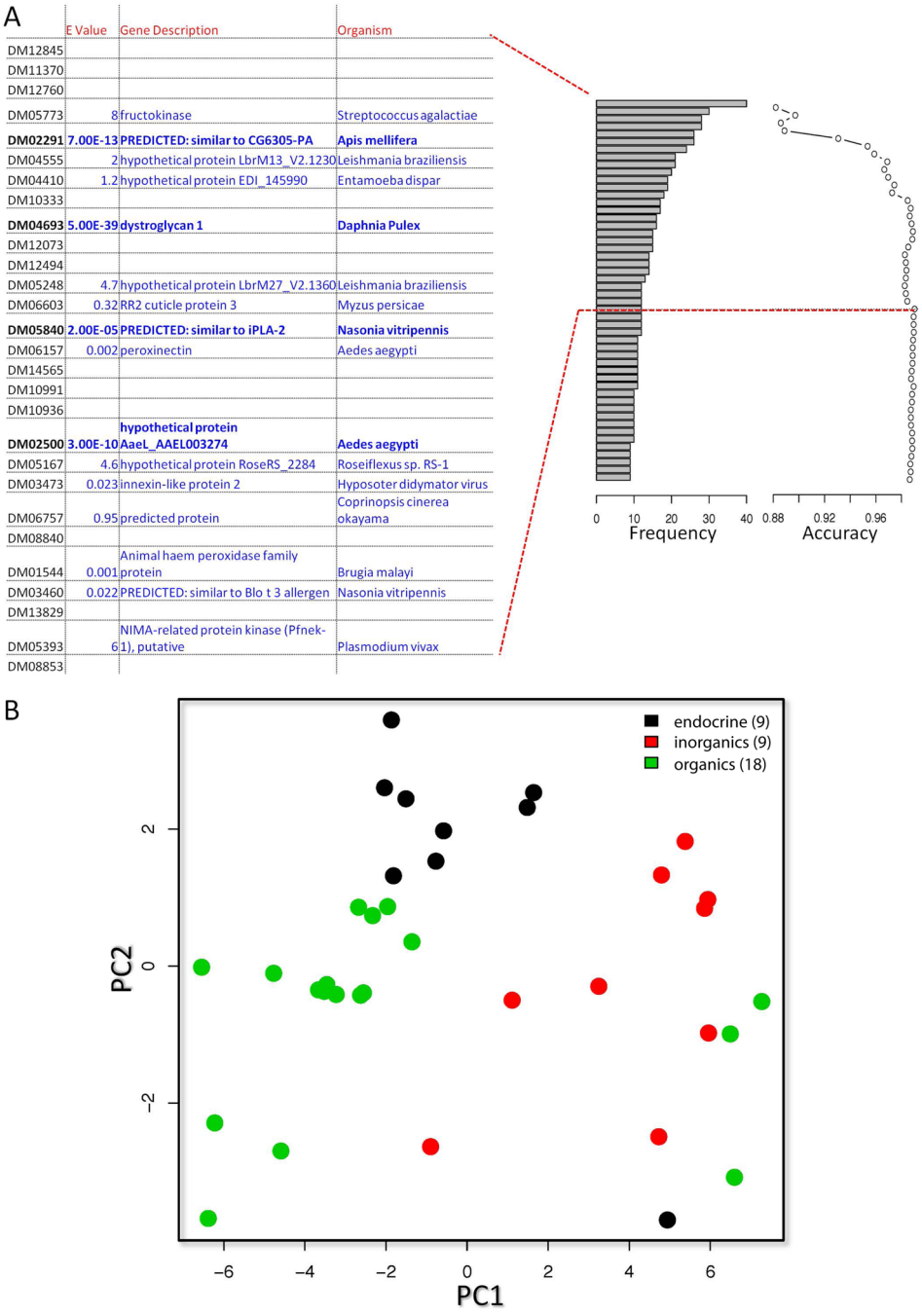


Figure 2. Gene-based model discriminating between inorganic and organic compounds. A) Shows the 19 genes identified to be 100% predictive. The graph adjacent to the table shows that the top 5 genes already account for an accuracy of 98% with the remaing 14 genes only adding little toward the final model. B) A PCA of the 19 genes shows that the majority of the variation (as summarized by PC1) is distinctive of chemical class. Note that the PCA is not designed to represent the discriminatory power of a statistical model and therefore is only used to graphically visualize the relationship between the samples while maintaining the majority of the information.

the basis of differential gene expression alone suggested the use of a more complex statistical modeling procedure to develop molecular signatures predictive of chemical class.

Gene-Level Models Predictive of Chemical Class. We built predictive models based on the transcriptional profile of subset of genes, which are predictive of chemical class. We were able to develop a single optimized model (see the Materials and Methods section for details) consisting of a subset of 14 genes whose transcriptional profiles are able to discriminate between inorganics and organics with 100% accuracy (see Supporting Information Figure S6 for a detailed breakdown of prediction of each sample). Interestingly, the three most relevant genes in this model already account for 97% of the prediction accuracy (Figure 2A). To visualize the separation between two groups (inorganic vs organic) we used these 19 genes as an input to a principal component analysis (PCA) (Figure 2B). Despite the fact that PCA is not an accurate classification tool and therefore can only partially represent the information contained in the predictive models, it clearly separated between organic and inorganic compounds. To provide a greater challenge to the identified model we extended our test to try and predict previously unseen chemicals using a CV approach. We systematically removed each chemical in the data set and tested how well our model performs. We then estimated the



	E Value	Gene Description	Organism
DM12845			
DM11370			
DM12760			
DM05773	8	fructokinase	Streptococcus agalactiae
DM02291	**7.00E-13**	**PREDICTED: similar to CG6305-PA**	**Apis mellifera**
DM04555	2	hypothetical protein LbrM13_V2.1230	Leishmania braziliensis
DM04410	1.2	hypothetical protein EDI_145990	Entamoeba dispar
DM10333			
DM04693	**5.00E-39**	**dystroglycan 1**	**Daphnia Pulex**
DM12073			
DM12494			
DM05248	4.7	hypothetical protein LbrM27_V2.1360	Leishmania braziliensis
DM06603	0.32	RR2 cuticle protein 3	Myzus persicae
DM05840	**2.00E-05**	**PREDICTED: similar to iPLA-2**	**Nasonia vitripennis**
DM06157	0.002	peroxinectin	Aedes aegypti
DM14565			
DM10991			
DM10936			
DM02500	**3.00E-10**	**hypothetical protein AaeL_AAEL003274**	**Aedes aegypti**
DM05167	4.6	hypothetical protein RoseRS_2284	Roseiflexus sp. RS-1
DM03473	0.023	innexin-like protein 2	Hyposoter didymator virus
DM06757	0.95	predicted protein	Coprinopsis cinerea okayama
DM08840			
DM01544	0.001	Animal haem peroxidase family protein	Brugia malayi
DM03460	0.022	PREDICTED: similar to Blo t 3 allergen	Nasonia vitripennis
DM13829			
DM05393	6	NIMA-related protein kinase (Pfnek-1), putative	Plasmodium vivax
DM08853			

Figure 3. Gene-based model differentiating inorganics, endocrine disruptors, and organics. A) Within the 28 genes in the model only four could be annotated. The plot adjacent to the table shows that the top 5 genes are sufficient in predicting the three classes with an accuracy of 92%. B) A PCA of the model shows that the majority of the variation is accredited to the difference between inorganic and the other two classes (PC1), whereas the separation between endocrine and organic chemicals is captured by PC2. Note that the PCA is not designed to represent the discriminatory power of a statistical model and therefore is only used to graphically visualize the relationship between the samples while maintaining the majority of the information.

degree of uncertainly in classifying the removed chemical. The class specific accuracy computed with this method remained 100%, and each chemical was always classified with 100% certainty in the correct class (Figure S1). Of the 19 genes in the model, only seven genes were homologous to genes in other species; although their function was unknown. However, most of the identified genes are predicted or hypothetical proteins, while only two genes were homologous to *D. pulex* representing 26S proteasome regulatory subunit T3 and hexosaminidase. Consequently, it is difficult to interpret the mechanistic or biologic rationale underlying the separation between inorganics and organics.

The success in separating between inorganic from organic compounds in combination with the previously observed response of *D. magna* to endocrine disruptors[14] led us to investigate whether we could further differentiate between inorganics and, within the organic class, endocrine disruptors and the group of remaining organic chemicals. We utilized the same variable selection approach as before but altered the classes to represent the three chemical groups (inorganics,

dx.doi.org/10.1021/es402819c | *Environ. Sci. Technol.* 2013, 47, 11747−11756

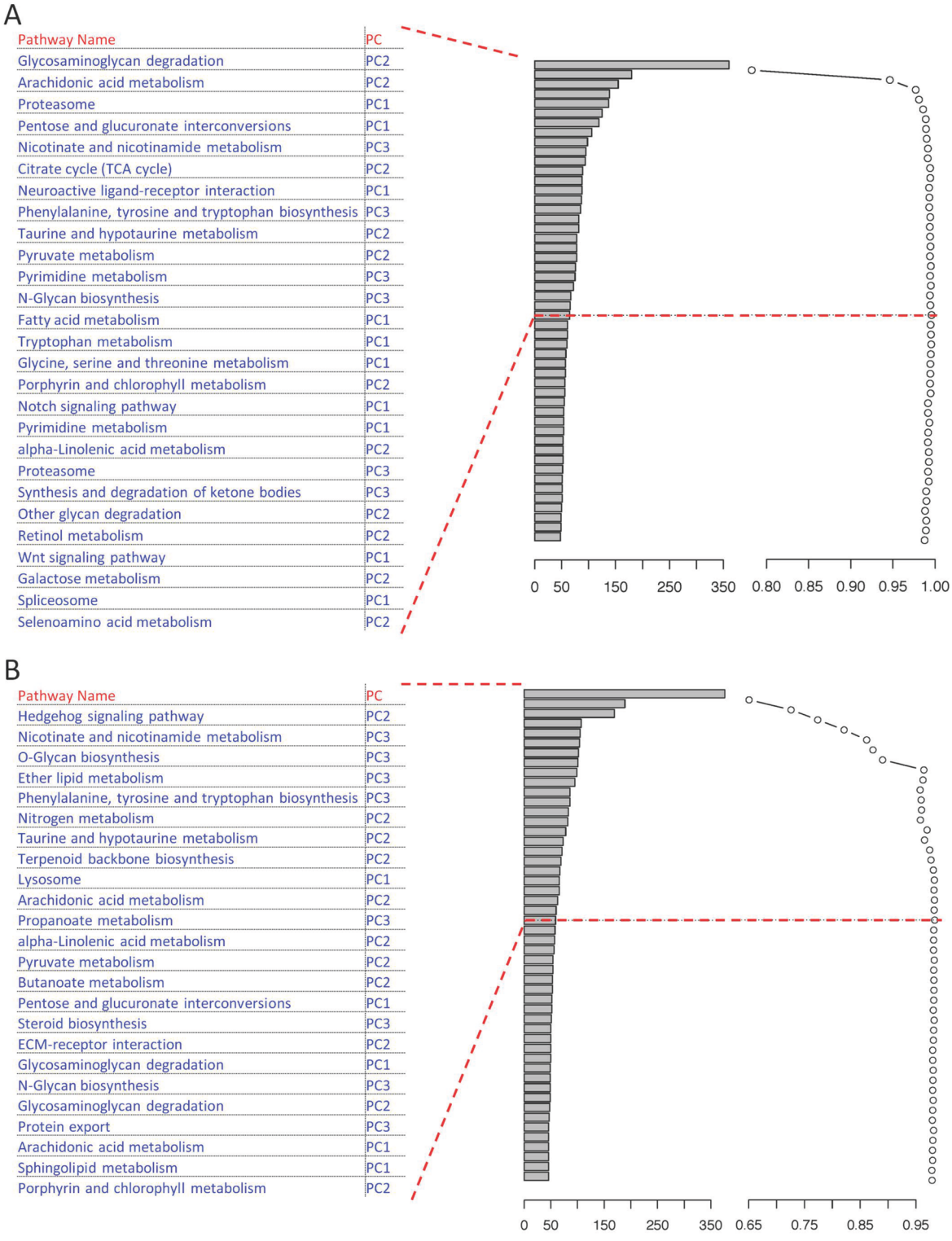


Figure 4. Pathway-based models differentiating between organics and inorganics (A) and inorganics, endocrine disruptors, and organic compounds (B). A) Here the graph shows the selection frequency for the top 50 pathways and its contribution to the final model. Interestingly, glycosaminoglycan degradation appears to be a very important in differentiating inorganics and organics (>350/1000 models contained this component). The top 3 pathways already account for a 98% accuracy. B) Similarly to the model in A, the top pathway, hedgehog signaling, is chosen >350 times in 1000 models. To achieve at least a 90% accuracy at least 5 pathways are needed.

endocrine disruptors, organics). Our approach identified a 28-gene model with a prediction accuracy of 99% using the training and test strategy mentioned above (Figure 3A and Supporting Information Figure S7). In this case, the first five genes account for approximately 92% of the accuracy. A plot of the first and second principal component of this model reveals a visible separation between all three classes (Figure 3B). To evaluate our models' ability to detect yet unseen chemical as with the organic vs inorganic model we found that copper, 20-hydroxyecdysone, and pyripoxyfen were not classified correctly

(Figure S2). Similarly to the previous model, only four genes within the model could be annotated using all species present in GenBank.

Pathway-Level Models Predictive of Chemical Class. Predictor models of contaminant class based on biological pathways have several potential advantages over gene level models. First, we suggest that contaminants with related outcomes may affect similar biological pathways but not necessarily the same genes.[29,30] Models in which genes within a pathway are considered together can take advantage of such relationships in developing predictors. In addition, pathway models have the added advantage of being biologically interpretable providing a pathway structure. This however comes at a cost where genes with no annotation are removed from the data set. In order to develop pathway level predictors, we performed a new analysis based on overall indices of pathway activity rather than individual gene expression values. We used these indices as an input to the variable selection and classification procedure outlined above to identify pathways whose activity was predictive of a contaminant class (see the Materials and Methods section for additional details). Similarly to the gene-level models, we first attempted to classify between organics and inorganics. We identified a predictive model based on the activity of 27 pathways with an accuracy of 99.6% (Figure 4A). However, the first seven pathways are sufficient to accurately predict 98.9% of sample classes. Sample specific misclassification shows small deviations of a perfect accuracy score from nickel, chromium, phenol, and acrylonitrile which total 1.5% (Supporting Information Figure S7).

Following a more stringent test, trying to predict compounds not part of the training test using a CV strategy we identified nickel, phenol, and acrylonitrile to be misclassified by our model (Figure S3). The overall accuracy, however, for all chemicals reached 92% (88% and 92% for inorganics and organics, respectively). A plot of the first and second components of a PCA of the 27 pathways (Figure S5A) shows a similar separation between inorganics and organics in sample space as the gene-level model (Figure 2 and Figure 3). The three most predictive pathways were glycosaminoglycan (GAG) degradation, arachidonic acid metabolism, and proteasome mediated degradation (Figure 4B). It is interesting to note that we identified hexosamidase in the GAG pathway and a 26 proteasome subunit as individual predictive genes using the gene level model differentiating organics from inorganics. Individual pie charts illustrating the number of differentially expressed genes for each chemical class, the contribution of each gene to the index of pathway activity and the accuracy of the model is shown in the Supporting Information (pages S40−S44). We also developed a pathway based model discriminating between inorganics, endocrine disruptors, and organic compounds. The most predictive representative model was based on a set of 24 pathways, 10 of which also appeared in metal vs nonmetal model, with an accuracy of 98.3% (Figure 4B, PCA:Figure S5B). Sample specific classification accuracy identified a misclassification of a single replicate of methylfarnesoate to be contributing to the slightly lower prediction accuracy. Small deviations also lowering the accuracy rate can be seen in samples exposed to beta-estradiol, toxaphene, bifenthrin, bis-2(ethyl-hexyl)-phthalate, MTBE, and phenol totalling a 6.4% error. Testing the model for its ability to predict specific chemicals removed from the training set, we found that some of the endocrine disruptors were misclassified (20-hydroxyecdysone, beta-estradiol, and toxaphene always misclassified, methoxychlor and pyripoxyfen misclassified 50% of the time). This gives chemical class accuracies of 100%, 56%, and 95% for inorganics, endocrine disruptors, and organic compounds, respectively (Figure S4). To distinguish between the three different classes the model consisted of nine different pathways. As in the previous analysis, individual pie charts illustrating the number of differentially expressed genes for each chemical class, the contribution of each gene to the index of pathway activity, and the accuracy of the model is shown in the Supporting Information (pages S44−S54). Not surprisingly, differential expression of gene by exposure to both inorganics and endocrine disruptors contribute to the indices of pathway activity and accuracy of the model. Of particular note, the hedgehog signaling, O-glycan (Mucin), and terpenoid synthesis pathway are identified as playing a major role in the predictive model. Together these pathway level prediction models provided similar predictive capability as gene level models while providing insight into potential mechanisms which differentiate between the exposures.

DISCUSSION

Identification of the contaminants underlying toxicity in aquatic ecosystems can represent a considerable challenge. Current toxicity identification analysis (TIE) approaches focus on the use of physical separation and various amendment strategies and subsequent monitoring of remaining toxicity. While these approaches have been the mainstay of practice in environmental monitoring, they have considerable limitations including substantial cost, require the fractionation/treatment of source water which can modify bioavailability of toxicants, and take a significant time to complete.[31] As an alternative to classic TIE, we have proposed an approach which we call molecular TIE (mTIE) to identify contaminants underlying toxicity by monitoring organismal response to a toxicant. We have previously demonstrated that *D. magna* produce distinctive patterns of gene expression in response to contaminant stress and that these patterns can be used in a diagnostic manner to understand mode of action and investigate the cause of toxicity.[32−36] In this report, we carried out gene expression profiling to a set of 36 contaminants of environmental concern of diverse chemical structure, proposed mode of action, and aquatic toxicity at a single arbitraty standardized dose chosen to minimize overt toxicity while retaining significant gene expression responses. Our results support our previous conclusions that distinct organismal responses are reflected in the gene expression profile to different contaminants.[32−36] While each gene expression profile is unique, we take advantage of similarities in expression profiles between related contaminants in order to develop predictive models which can distinguish between different groups of contaminants. We are able to identify genes (variables) through a "genetic" selection procedure (GALGO) whose expression enables differentiation between different classes of contaminants. We demonstrate that we can distinguish between two classes (inorganic vs organic) as well as three classes (inorganic vs endocrine disruptors vs organic). Interestingly, relatively few genes (up to 5) are needed to achieve high classification accuracy. While the utility of these models to distinguish between contaminant classes does not require any knowledge of the function of these genes, a biological framework to understand differences would increase interpretability of the models. Unfortunately, few of the genes in either model are annotated which makes mode of

action interpretation difficult. We therefore developed pathway driven predictive models which we demonstrate provide equivalent capability to distinguish between chemical classes. As these pathway models use only a minority of the genes present on the chips, the results suggest that there is considerable redundancy in the predictive potential of the expression data.

While we are still limited in our capability to interpret the role of the pathways in response to the contaminants classes, they do suggest potential differences in the mode of action of the different contaminants that could be further investigated with targeted analysis. For example, the pathway level models suggest that glycosaminoglycan (GAG) degradation, arachidonic acid metabolism, and proteasome mediated degradation are important in distinguishing organic vs inorganic. Inspection of the individual genes in arachidonic acid metabolism revealed that key genes which distinguish between inorganic and organic include genes predicted to encode gamma-glutamyltransferase like-3, glutathione S-transferase sigma, and glutathione peroxidase which are all involved in glutathione metabolism. Glutathione metabolism has been previously implicated in metal metabolism including in *D. pulex*[37] and *D. magna*.[32,38,39] However, these genes may not be uniquely specific to inorganic exposure but may also represent oxidative stress which in the case of our data set is predictive of inorganic exposure. The proteasome is involved in the degradation of oxidized proteins which can result from metal stress.[40] Similarly, nine pathways were identified as most predictive in differentiating between inorganic, endocrine disruptors, and other organic contaminants. Although the CV procedure we utilized reports only a 56% accuracy for the endocrine disrupting class, it is interesting to note that two of the main model components hedgehog signaling and terpenoid synthesis pathway have been previously implicated in endocrine development[41] and creation of Juvenile Hormone (JH) which plays a key role in male sex determination and response to stress.[42,43] To increase the accuracy and robustness of the model a broader range of endocrine disrupting compounds would be necessary to develop more generalized models.

To develop models of interest we utilized a well validated multivariate variable selection procedure implemented in the GALGO R package. There are several advantages with this procedure as compared to, for example, univariate techniques.[9,44] Such approaches consider each variable separately for their ability to explain a particular phenotype. The choice is usually based on ranking variables according to a statistics that is significantly associated with the difference in the classification groups. A number of single biomarkers used in environmental monitoring, such as vitellogenin, metallothionein, or CYP1A, derived from such approaches have shown to be limited to specific types of exposures and in many cases lack a strong link between exposure and biological effect (for an in-depth discussion see refs 45 and 46). We therefore propose to move away from single biomarker identification and utilize a multivariate variable selection approach which can integrate phenotypic end points. This would allow to identify genes which provide predictive power (exposure biomarkers) and simultaneously provide insight into the underlying biological response (effect biomarkers). Using such an approach we show that we are able to generate testable hypotheses of the impact of chemicals on a species.

While these results show the promise of mTIE, considerable work remains to develop this approach for routine environmental monitoring. In the immediate future we envisage that this will be achieved in at least three levels. The most important challenge is the development of larger studies representing a much broader portion of the chemical space and a more complex spectrum of doses and time points. This will allow defining additional chemical classes, for example based on relevant physical chemical features, and to develop more general models designed to work in a broader range of experimental conditions. Another aspect that is likely to play a major role in the development of future mTIE systems is the validation in the environmental setting. In this context, it is likely that the complex interaction between multiple physicochemical stressors will pose a challenge to the application of an mTIE system based on signatures developed using single chemical exposures as a reference set. Although this field is still in its infancy we already have evidence that it may be possible to identify chemical specific signatures that are informative in an environmental scenario. For example, in previous publications we have shown that in fish species, expression signatures defined by acute exposures are predictive of the status of environmental samples.[47] Moreover, using a knowledge base approach, based on information from acute exposures, it has been possible to detect specific chemical signatures in fish exposed to contaminated water.[48] On the basis of these results we believe that some degree of extrapolation will also be possible in *D. magna*. A third aspect that needs considerable development is the improvement of gene annotation in nonmodel species (in this particular case the *Daphnia magna* transcriptome). It is conceivable that due to the increasing amount of genomic data becoming available considerable progress in this area will provide additional resources for analyzing and interpreting of models such as developed in this paper. More specifically, the availability of a fully annotated *D. magna* genome will eventually allow for the development of a more interpretable model.

Ultimately, we envisage that an mTIE approach which can be robustly used in an environmental setting will need to be developed using algorithms that can integrate information between laboratory control and single and mixture exposures as well as "learn" from a collection of environmental exposure where an in-depth high quality chemistry is available.

■ ASSOCIATED CONTENT

Supporting Information

The data are available under the GEO Accession GSE43564. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +44-151-795-4558. Fax: +44-151-795-4408. E-mail: f.falciani@liv.ac.uk (F.F.). Phone: +1-510-642-1834. Fax: +1-510-643-3132. E-mail: vulpe@berkeley.edu (C.V.).

Author Contributions

#Authors wish to be considered joint senior and corresponding authors.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by a NSF (CBET-1066358) grant to C.V., by a Korea Research Foundation Grant [KRF-2008-357-

D00144] to H.J., and by an NERC Grant [NE/1028246/2] to F.F.

REFERENCES

(1) Jo, H.; Jung, J. Quantification of differentially expressed genes in *Daphnia magna* exposed to rubber wastewater. *Chemosphere* **2008**, *73*, 261−266.

(2) Kilham, S.; Kreeger, D.; Lynn, S.; Goulden, C.; Herrera, L. COMBO: a defined freshwater culture medium for algae and zooplankton. *Hydrobiologia* **1998**, *377*, 147−159.

(3) Weber, C. I. *Methods for measuring the acute toxicity of effluents and receiving waters to freshwater and marine organisms*; Environmental Monitoring Systems Laboratory, Office of Research and Development, US Environmental Protection Agency: 1993.

(4) Poynton, H.; Lazorchak, J.; Impellitteri, C.; Blalock, B.; Rogers, K.; Allen, H.; Loguinov, A.; Heckman, J.; Govindasmawy, S. Toxicogenomic responses of nanotoxicity in Daphnia magna exposed to silver nitrate and coated silver nanoparticles. *Environ. Sci. Technol.* **2012**, *46*, 6288−6296.

(5) Yee, H. Y.; Agnes, P.; Sandrine, D. Marray: Exploratory analysis for two-color spotted microarray data. R package version 1.36.0; 2009.

(6) Smyth, G. K. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*; Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., Huber, W., Eds.; Springer: New York, 2005; pp 397−420.

(7) Tárraga, J.; Medina, I.; Carbonell, J.; Huerta-Cepas, J.; Minguez, P.; Alloza, E.; Al-Shahrour, F.; Vegas-Azcárate, S.; Goetz, S.; Escobar, P.; Garcia-Garcia, F.; Conesa, A.; Montaner, D.; Dopazo, J. GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucleic Acids Res.* **2008**, *36*, W308−W314.

(8) Díaz-Uriarte, R.; De Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinf.* **2006**, *7*, 3.

(9) Trevino, V.; Falciani, F. GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics* **2006**, *22*, 1154−1156.

(10) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18−22.

(11) Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **1999**, *27*, 29−34.

(12) R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2012; ISBN 3-900051-07-0.

(13) Antczak, P.; Ortega, F.; Chipman, J.; Falciani, F. Mapping drug physico-chemical features to pathway activity reveals molecular networks linked to toxicity outcome. *PloS One* **2010**, *5*, e12385.

(14) Olmstead, A.; LeBlanc, G. Effects of endocrine-active chemicals on the development of sex characteristics of Daphnia magna. *Environ. Toxicol. Chem.* **2000**, *19*, 2107−2113.

(15) Olmstead, A. W.; Leblanc, G. A. Juvenoid hormone methyl farnesoate is a sex determinant in the crustacean Daphnia magna. *J. Exp. Zool.* **2002**, *293*, 736−739.

(16) Olmstead, A. W.; LeBlanc, G. A. Insecticidal juvenile hormone analogs stimulate the production of male offspring in the crustacean Daphnia magna. *Environ. Health Perspect.* **2003**, *111*, 919.

(17) Baldwin, W. S.; Graham, S. E.; Shea, D.; LeBlanc, G. A. Metabolic androgenization of female Daphnia magna by the xenoestrogen 4-nonylphenol. *Environ. Toxicol. Chem.* **1997**, *16*, 1905−1911.

(18) Tatarazako, N.; Oda, S. The water flea Daphnia magna (Crustacea, Cladocera) as a test species for screening and evaluation of chemicals with endocrine disrupting effects on crustaceans. *Ecotoxicology (London, England)* **2007**, *16*, 197−203.

(19) Depledge, M.; Billinghurst, Z. Ecological significance of endocrine disruption in marine invertebrates. *Mar. Pollut. Bull.* **1999**, *39*, 32−38.

(20) Iguchi, T.; Watanabe, H.; Katsu, Y. Toxicogenomics and ecotoxicogenomics for studying endocrine disruption and basic biology. *Gen. Comp. Endrocrinol.* **2007**, *153*, 25−9.

(21) Tatarazako, N.; Oda, S.; Watanabe, H.; Morita, M.; Iguchi, T. Juvenile hormone agonists affect the occurrence of male Daphnia. *Chemosphere* **2003**, *53*, 827−33.

(22) McCarthy, J. F. Ponasterone A: A new ecdysteroid from the embyros and serum of brachyuran crustaceans. *Steroids* **1979**, *34*, 799−806.

(23) Oberdörster, E.; Cheek, A. O. Gender benders at the beach: Endocrine disruption in marine and estuarine organisms. *Environ. Toxicol. Chem.* **2001**, *20*, 23−36.

(24) Soto, A. M.; Chung, K. L.; Sonnenschein, C. The pesticides endosulfan, toxaphene, and dieldrin have estrogenic effects on human estrogen-sensitive cells. *Environ. Health Perspect.* **1994**, *102*, 380−3.

(25) Hansen, P.-D.; Dizer, H.; Hock, B.; Marx, A.; Sherry, J.; McMaster, M.; Blaise, C. Vitellogenin âĂŞ a biomarker for endocrine disruptors. *TrAC, Trends Anal. Chem.* **1998**, *17*, 448−451.

(26) Hutchinson, T. H. Reproductive and developmental effects of endocrine disrupters in invertebrates: in vitro and in vivo approaches. *Toxicol. Lett.* **2002**, *131*, 75−81.

(27) Soontornchat, S.; Li, M.; Cooke, P.; Hansen, L. Toxicokinetic and toxicodynamic influences on endocrine disruption by polychlorinated biphenyls. *Environ. Health Perspect.* **1994**, *102*, 568−71.

(28) Gaido, K. W. Differential interaction of the methoxychlor metabolite 2,2-bis-(p-hydroxyphenyl)-1,1,1-trichloroethane with estrogen receptors alpha and beta. *Endocrinology* **1999**, *140*, 5746−5753.

(29) Shen, K.; Tseng, G. C. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* **2010**, *26*, 1316−1323.

(30) Manoli, T.; Gretz, N.; Gröne, H.-J.; Kenzelmann, M.; Eils, R.; Brors, B. Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics* **2006**, *22*, 2500−2506.

(31) Ankley, G.; Schubauer-Berigan, M. Background and overview of current sediment toxicity identification evaluation procedures. *J. Aquat. Ecosyst. Health* **1995**, *4*, 133−149.

(32) Poynton, H.; Varshavsky, J.; Chang, B.; Cavigiolio, G.; Chan, S.; Holman, P.; Loguinov, A.; Bauer, D.; Komachi, K.; Theil, E. Daphnia magna ecotoxicogenomics provides mechanistic insights into metal toxicity. *Environ. Sci. Technol.* **2007**, *41*, 1044−1050.

(33) Poynton, H.; Loguinov, A.; Varshavsky, J.; Chan, S.; Perkins, E.; Vulpe, C. Gene expression profiling in Daphnia magna part I: concentration-dependent profiles provide support for the no observed transcriptional effect level. *Environ. Sci. Technol.* **2008**, *42*, 6250−6256.

(34) Garcia-Reyero, N.; Poynton, H.; Kennedy, A.; Guan, X.; Escalon, B.; Chang, B.; Varshavsky, J.; Loguinov, A.; Vulpe, C.; Perkins, E. Biomarker discovery and transcriptomic responses in Daphnia magna exposed to munitions constituents. *Environ. Sci. Technol.* **2009**, *43*, 4188−4193.

(35) Poynton, H.; Lazorchak, J.; Impellitteri, C.; Smith, M.; Rogers, K.; Patra, M.; Hammer, K.; Allen, H.; Vulpe, C. Differential gene expression in Daphnia magna suggests distinct modes of action and bioavailability for ZnO nanoparticles and Zn ions. *Environ. Sci. Technol.* **2010**, *45*, 762−768.

(36) Zeis, B.; Lamkemeyer, T.; Paul, R.; Nunes, F.; Schwerin, S.; Koch, M.; Schütz, W.; Madlung, J.; Fladerer, C.; Pirow, R. Acclimatory responses of the Daphnia pulex proteome to environmental changes. I. Chronic exposure to hypoxia affects the oxygen transport system and carbohydrate metabolism. *BMC Physiol.* **2009**, *9*, 7.

(37) Shaw, J.; Colbourne, J.; Davey, J.; Glaholt, S.; Hampton, T.; Chen, C.; Folt, C.; Hamilton, J. Gene response profiles for Daphnia pulex exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. *BMC Genomics* **2007**, *8*, 477.

(38) Fan, W.; Tang, G.; Zhao, C.; Duan, Y.; Zhang, R. Metal accumulation and biomarker responses in Daphnia magna following cadmium and zinc exposure. *Environ. Toxicol. Chem.* **2009**, *28*, 305−310.

(39) Barata, C.; Varo, I.; Navarro, J.; Arun, S.; Porte, C. Antioxidant enzyme activities and lipid peroxidation in the freshwater cladoceran *Daphnia magna* exposed to redox cycling compounds. *Comp. Biochem. Physiol., Part C: Toxicol. Pharmacol.* **2005**, *140*, 175−186.

(40) Shang, F.; Taylor, A. Ubiquitin-proteasome pathway and cellular responses to oxidative stress. *Free Radical Biol. Med.* **2011**, *51*, 5−16.
(41) Cohen, M. Hedgehog signaling: endocrine gland development and function. *Am. J. Med. Genet., Part A* **2009**, *152*, 238−244.
(42) Tatarazako, N.; Oda, S.; Watanabe, H.; Morita, M.; Iguchi, T. Juvenile hormone agonists affect the occurrence of male *Daphnia*. *Chemosphere* **2003**, *53*, 827−833.
(43) Rider, C.; Gorr, T.; Olmstead, A.; Wasilak, B.; LeBlanc, G. Stress signaling: coregulation of hemoglobin and male sex determination through a terpenoid signaling pathway in a crustacean. *J. Exp. Biol.* **2005**, *208*, 15−23.
(44) Sha, N.; Vannucci, M.; Tadesse, M. G.; Brown, P. J.; Dragoni, I.; Davies, N.; Roberts, T. C.; Contestabile, A.; Salmon, M.; Buckley, C. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **2004**, *60*, 812−819.
(45) Poynton, H. C.; Wintz, H.; Vulpe, C. D. Progress in ecotoxicogenomics for environmental monitoring, mode of action, and toxicant identification. *Adv. Exp. Biol.* **2008**, *2*, 21−323.
(46) Walker, C. H. *Principles of ecotoxicology*; CRC PressI LLC: 2006.
(47) Williams, T. D.; Turan, N.; Diab, A. M.; Wu, H.; Mackenzie, C.; Bartie, K. L.; Hrydziuszko, O.; Lyons, B. P.; Stentiford, G. D.; Herbert, J. M. Towards a system level understanding of non-model organisms sampled from the environment: a network biology approach. *PLoS Comput. Biol.* **2011**, *7*, e1002126.
(48) Falciani, F.; Diab, A.; Sabine, V.; Williams, T.; Ortega, F.; George, S.; Chipman, J. Hepatic transcriptomic profiles of European flounder (*Platichthys flesus*) from field sites and computational approaches to predict site from stress gene responses following exposure to model toxicants. *Aquat. Toxicol.* **2008**, *90*, 92−101.