Environmental Science & Technology

Article

pubs.acs.org/est

Toxicity Mechanisms Identification via Gene Set Enrichment Analysis of Time-Series Toxicogenomics Data: Impact of Time and Concentration

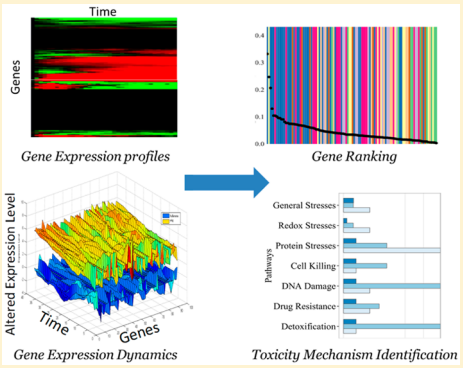
Ce Gao,[†] David Weisman,[‡] Jiaqi Lan,[†] Na Gou,[†] and April Z. Gu*,[†]

[†]Department of Civil and Environmental Engineering, Northeastern University, Boston, Massachusetts 02115, United States

[‡]Department of Biology, University of Massachusetts, Boston, Massachusetts 02125, United States

Supporting Information

ABSTRACT: The advance in high-throughput "toxicogenomics" technologies, which allows for concurrent monitoring of cellular responses globally upon exposure to chemical toxicants, presents promises for next-generation toxicity assessment. It is recognized that cellular responses to toxicants have a highly dynamic nature, and exhibit both temporal complexity and dose-response shifts. Most current gene enrichment or pathway analysis lack the recognition of the inherent correlation within time series data, and may potentially miss important pathways or yield biased and inconsistent results that ignore dynamic patterns and time-sensitivity. In this study, we investigated the application of two score metrics for GSEA (gene set enrichment analysis) to rank the genes that consider the temporal gene expression profile. One applies a novel time series CPCA (common principal components analysis) to generate scores for genes based on their contributions to the common temporal variation among treatments for a given chemical at different concentrations. Another one employs an integrated altered gene expression quantifier-TELI (transcriptional effect level index) that integrates altered gene expression magnitude over the exposure time. By comparing the GSEA results using two different ranking metrics for examining the dynamic responses of reporter cells treated with various dose levels of three model toxicants, mitomycin C, hydrogen peroxide, and lead nitrate, the analysis identified and revealed different toxicity mechanisms of these chemicals that exhibit chemical-specific, as well as time-aware and dose-sensitive nature. The ability, advantages, and disadvantages of varying ranking metrics were discussed. These findings support the notion that toxicity bioassays should account for the cells' complex dynamic responses, thereby implying that both data acquisition and data analysis should look beyond simple traditional end point responses.



INTRODUCTION

The needs in toxicity assessment of an extremely large and ever-increasing number of chemicals for their potential environmental and health risks demands for the development of mechanistic, cost-effective toxicity testing scheme and predictive models to provide toxicological information that transcends the limits of traditional toxicity assessment approach.[1,2] The advances in high-throughput toxicogenomics technologies, which allow for globally concurrent monitoring of cellular responses of numerous transcripts, proteins, or metabolites upon exposure to chemical toxicants, presents promise for achieving this goal.[2,3]

High-dimensional toxicogenomics time series data refer to those that record multiple measurements over time and those that incorporate multiple experimental factors, such as genes, conditions, and dose concentrations.[4] It is recognized that cellular or organisms' responses to toxicants are highly dynamic, and their global response profiles depend on time of measurement.[5] However, efforts in illustrating the impact of time on toxic assay results have been quite limited due to the lack of time-series toxicogenomics data. This is partially attributable to the labor-intensiveness or high-cost associated with mainstream toxicogenomics techniques such as RNA-seq or microarray technologies that prohibit measurements with high temporal resolution.[6,7] An alternative approach is the use of whole-cell arrays with transcriptional fusions of reporter genes, which allows for faster and lower-cost real-time measurement of temporal gene expressions for a large number of chemicals under various test conditions.[8,9] The high-dimensional time series gene expression data, generated by such arrays for example, call for analytic approaches that are time-factor sensitive. Most current studies simply adopt strategies extended from those of static, time-independent experiments and resort to integrated end point-like quantities,[10] which do not account for the dynamic nature of stress responses and lose temporal information by discarding all information other than end point measurements.[3,11]

Received: October 23, 2014
Revised: February 22, 2015
Accepted: March 2, 2015
Published: March 18, 2015

DOI: 10.1021/es505199f
Environ. Sci. Technol. 2015, 49, 4618−4626

Pathway analysis is one family of bioinformatic tools for toxicity mechanisms elucidation, which aims at pinpointing key functional gene groups and regulatory pathways evoked during the toxicant exposure under a given condition.[12,13] Through shifting the focus from detecting differentially expressed genes individually to discerning sets of genes that share common biological function or regulation, pathway analysis catches the expression patterns on the higher pathway level, avoids results misinterpretation due to subjective expression thresholds for individual genes, and reduces the complexity of data analysis that deals with the daunting number of genes.[12−15] Pathway analysis of high-dimensional toxicogenomics data, such as time series data, faces a great challenge, however, since most current techniques are mainly designed for the analysis of biological system snapshots.[13] The commonly used pathway analysis techniques, such as the gene set enrichment analysis (GSEA), are designed to find differentially expressed sets of genes sharing common functions or regulations.[14,15] In GSEA, genes are ranked based on a certain metric, which can simply be the expression level, or more complicated ranking methods based on various statistical analyses (i.e., Pearson's correlation, Euclidean distance, or signal-to-noise ratio).[15] A typical pathway analysis of time series experiments would analyze expression changes at different time points individually or reduce the time series to an end point-like metric, both of which bear an implicit assumption that data at multiple time points are independent.[16,17] Lacking the recognition of the inherent correlation within time series data, this approach may miss potentially important pathways or yield biased and inconsistent results that ignore dynamic patterns and time-sensitivity.

In addition to the time factor, it is recognized that the molecular toxicity response is also dose-dependent.[18] Transition of dominant function or pathway at different dose concentrations has been observed in previous studies, which provided extra mechanistic information beyond the traditional phenotypic dose-response curves.[19−22] The ability to reveal if and how the molecular toxicity response conserves or changes at various dose concentrations is therefore necessary and has not received adequate study.

In this study, we investigated the application of an improved GSEA to identify the activation of specific stress response categories and pathways for three representative chemicals based on the temporal altered gene expression data with multiple (six) dose concentrations. Temporal altered gene expression profiles were generated using a high-throughput whole-cell array of *gfp*-fused reporters, representing genes covering all known stress response pathways in *E. coli*.[10,23−25] Two score metrics were compared to rank the genes that consider the temporal gene expression profile. One employs an integrated altered gene expression quantifier-TELI (transcriptional effect level index) that integrates altered gene expression magnitude over the exposure time.[10] The other applies CPCA (common principal components analysis) to generate scores for all the genes based on their contribution to the common temporal variation among treatments for a given chemical at different concentrations. Compound-specific stress response activation profiles indicative of toxicity mechanisms were obtained using both metrics. The results were compared to determine the suitability, as well as advantages and disadvantages of applying the time-aware metric instead of the static metric. The impact of time and dose concentration on the gene enrichment analysis results were revealed and discussed.

MATERIALS AND METHODS

Toxicogenomics Time Series Data Generation. A high-throughput toxicogenomics assay was employed using the GFP-fused stress response ensemble whole-cell library of *E. coli* K12, MG1655,[23,24] with each fusion expressed from a low-copy plasmid, pUA66, which contains a kanamycin resistance gene and a fast folding *gfpmut2*, allowing for real-time measurement of gene expression level.[23,26] The selected stress response assay library covers a variety of genes (106 gene promoters) involved in different known cellular stress response pathways that are highly conserved among species,[27] and they are categorized into seven groups including general stresses, protein stresses, redox stresses, cell killing, DNA damage, drug resistance, and detoxification[28] (see Supporting Information (SI) Table S1 for a list of genes and their pathways).

Three compounds with known toxic mechanisms were evaluated for demonstration and they are mitomycin C (MMC), hydrogen peroxide (H_2O_2), and lead nitrate ($Pb(NO_3)_2$ or Pb^{2+}). For each chemical, 6 subcytotoxic concentrations (>95% survival percentage determined by a 2 h growth inhibition test) were applied for each chemical with 3 replicates, resulting in a total of 54 = 3(chemicals) × 6(concentrations) × 3(replicates) treatments. Details of chemical information are listed in SI Table S2.

The protocol to measure the temporal gene expression profile was described in our previous reports.[10,23] In brief, *E. coli* reporter strains were cultivated in 384-well microplates (Costar, Bethesda, MD, U.S.A.) in dark condition to avoid GFP photobleaching until the early exponential growth stage (OD600 ≈ 0.2) was reached. Samples of specific compounds prepared in growth media were distributed into wells via BioTek precision automated pipetting system. The plate was then placed into a microplate reader (Synergy Multi-Mode, Biotek, Winooski, VT) for simultaneous cell growth (absorbance, OD600) measurement, denoted as OD, and fluorescent readings (GFP level, excitation 485 nm, emission 528 nm), denoted as GFP. Measurements were taken every 5 min over 2 h, resulting in a total of 25 time points for every gene in every treatment.

Data Preprocessing. The raw GPF and OD data were first corrected against the average media controls (wells with growth media only) and blank media control for potential interference of chemicals, and referred as $GPF_{corrected}$ and $OD_{corrected}$. The gene expression level normalized by OD for each gene at every time point was calculated as $P = GPF_{corrected}/OD_{corrected}$ and was further adjusted by subtracting the value attributed by nonpromoter activities (promoter-less strain controls). The alteration in gene expression for a given gene at each time point due to chemical exposure relative to the control condition without any chemical exposure, also referred as induction factor I, is represented by as I = Pe /Pc, where, Pe = $(GFP/OD)_{experiment}$ is the normalized gene expression level in the experiments condition with chemical exposure, and Pc = $(GFP/OD)_{control}$ is the control condition without any chemical exposure. The natural logarithm of the induction factor $\ln(I)$ was then calculated for further data analysis, where a gene was up-regulated if $\ln(I) > 0$ and down-regulated if $\ln(I) < 0$.[29] All controls were performed in triplicates.

Gene Set Enrichment Analysis with Two Different Ranking Metric. Two ranking methods based on transcriptional effect level index (TELI) and common principal

DOI: 10.1021/es505199f
Environ. Sci. Technol. 2015, 49, 4618−4626

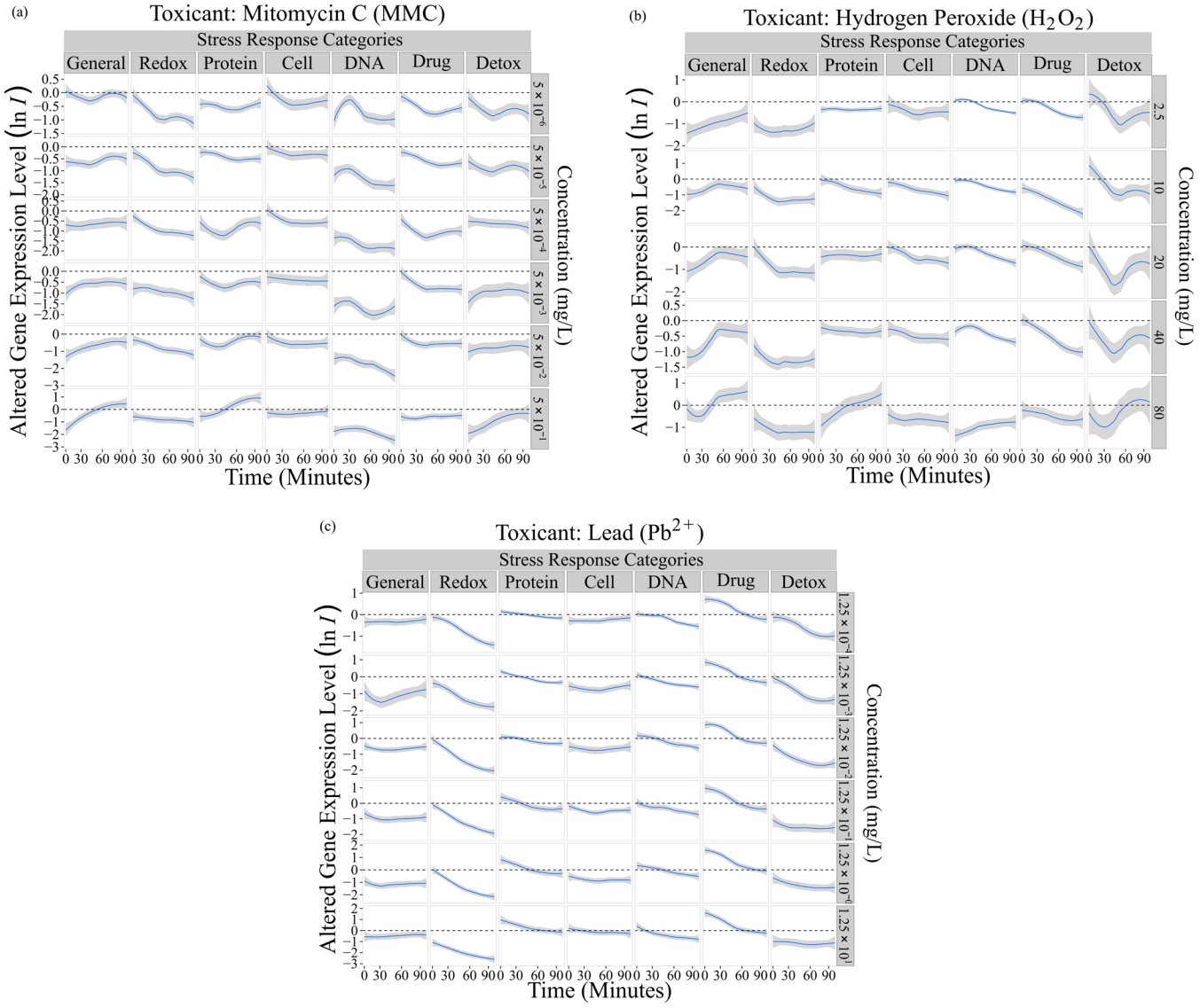


Figure 1. Exemplary temporal variations in gene expression profiles of stress response pathways (categories) upon exposure to various dose concentrations of (a) MMC, (b) H_2O_2, and (c) Pb^{2+}. Each curve represents the mean temporal variations (measured as ln(I), I is the altered gene expression level) profile of all the genes (3 replicates) in a specific stress response pathway, with 95% confidence intervals indicated by the gray bands. x-axis top: various stress response pathways and categories; x-axis bottom: exposure time in minutes. The temporal altered gene expression data at 5 min intervals were smoothed using moving average of every five successive measurements and therefore time zero corresponds to 20 min exposure time; y-axis left: temporal variations expressed in altered gene expression level as ln(I). Pathway abbreviation: General (general stresses), Redox (Redox stresses), Protein (protein stress), Cell (cell killing), DNA (DNA damage), Drug (drug resistance), and Detox(detoxification). (See the list of genes in SI Table S1.)

component analysis (CPCA) were applied for gene set enrichment analysis (GSEA), and they are described as follows.

Pathway Analysis: Gene Set Enrichment Analysis (GSEA). Gene Set Enrichment Analysis (GSEA) is a statistical procedure to determine whether a predefined set of genes is over-represented toward the top or bottom of a ranked gene list.[15] In this study, different gene lists were generated using either CPCA score or TELI as the ranking metric (see detailed description of CPCA and TELI in following sections). For a given list, the enrichment score (ES) for a specific gene set or pathway is calculated by walking down the ranked gene list, and increase the running-sum statistic if the gene from the pathway is encountered and decrease if otherwise. The ES is the maximum deviation from 0 encountered in the walk corresponding to a weighted Kolmogorov−Smirnov-like statistic.[30] The statistical significance of the ES was estimated by permutation test: the gene list is permutated 1000 times, and then an *ES* for each permutation was calculated to generate a null distribution for the ES. The p value for original ES was then calculated in relative to the null distribution. Because the ES for multiple pathways were calculated, the multiple comparison errors were adjusted using false-discovery rate (BH method) with adjusted $p < 0.05$ as significance cutoff point.[15,31]

Ranking Metric Based on Common Principal Component Analysis. We proposed to implement Common Principal Component Analysis (CPCA) for determination of ranking metrics based on the their contribution to the temporal variance in altered gene expression level over the exposure time.[32,33] A treatment was defined as one specific test for one

DOI: 10.1021/es505199f
Environ. Sci. Technol. 2015, 49, 4618−4626

particular chemical at a given concentration. The raw data for one treatment is a matrix for T time points (rows) and G genes (columns). PCA uses linear orthogonal transformation to convert the genes into a new set of uncorrelated variables, called principal components (PC). The transformation is designed such that the first PC points to the direction with the highest possible variance in the data space formed by the raw data, and each succeeding PC points to the direction with the largest remaining variance, while under the constraint of being orthogonal to all the preceding PCs. Each PC is a linear combination of the original genes, namely $P_j = \sum_{i=1}^{G} l_{ij} g_i$, where g_i denotes the i^{th} gene in the set of G genes, P_j denotes the j^{th} PC, and l_{ij}, called loading, represents the weight or contribution of the i^{th} gene to the j^{th} PC. Only the top PCs are retained for the purpose of dimension reduction and filtering out system noise, and they represent most of the variance of the raw data. We determined the number of PCs retained by setting the threshold to be 70%, namely, the first K PCs whose sum of variances is larger than 70% of the total variance are retained. For each gene, its contribution to the total variance can be measured using its loadings on these PCs, namely $S_i = \sum_{j=1}^{K} l_{ij}^2$, where s_i is the metric of the contribution of the i^{th} gene, K is total PCs retained. Since the variance discussed here indicates the temporal variation, the contribution of genes in their ranking metric was based on the assumption that higher temporal variation indicates higher transcriptional level alterations.

CPCA is a generalization of PCA to generate a set of common principal components (CPC) that agrees most closely to the data from multiple experiments. It is designed in such a way that the first CPC points to direction closest to all first PCs from each individual treatment, and so forth. Like PCs, each CPC is also a linear combination of the original genes, namely $q_j = \sum_{i=1}^{G} w_{ij} g_i$, where q_j denotes the j^{th} CPC, and w_{ij}, also called loading, represents the weight or contribution of the i^{th} gene to the j^{th} CPC. Mathematically, CPCs can be calculated as the eigenvectors of matrix, namely $H = \sum_{i=1}^{N} L_i L_i'$, where N is the total number of treatments, L_i is the matrix for the i^{th} treatment containing all the loadings for the retained PCs. Since the CPCs present the directions of common temporal variance, each gene's contribution to total temporal variance can be calculated using the new loadings, namely $t_i = \sum_{j=1}^{J} w_{ij}^2$, where t_i is the metric indicating the contribution of the i^{th} gene, J is the total number of CPCs and determined by the minimum number of PCs retained among all treatment. In this study, the genes' contribution to temporal variance, designated as CPCA score, was used as the ranking metric to differentiate genes with varying temporal activities.

Transcriptional Effect Level Index. In previous studies, we defined a toxicity end point-transcriptional effect level index (TELI) for quantifying gene expression data, which incorporates the number, magnitude, and the cumulative temporal pattern of genes with altered expression[3,24] and is calculated as follows:

$$\text{TELI}_{(i)} = \frac{\int_{t=0}^{T} e^{|\ln(I)|}\, dt}{T}$$

where $i = 1,...,G$ indexes one of the G genes, $t = 1,...,T$ indexes one of the T time points. More detailed description and discussion of TELI can be found in our previous reports.[10] In this study, we applied TELI as an alternative ranking matric for GSEA analysis.

Software. The calculation of TELI and CPCA scores were implemented using MATLAB (MathWorks, MA, version 2013a). GSEA algorithm and visualization of this study is implemented using R (version 2.15.3) together with package ggplot2 (version 0.9.3).

RESULTS AND DISCUSSION

Chemical-Specific and Temporally Dynamic Stress Response Gene Expression Profiles. Both visual and statistical examinations of temporal altered expression profiles of various genes from different stress response pathways revealed the chemical-specific and dose-dependent patterns (Figure 1, statistical analysis results in SI Tables S3−S5). Figure 1 displays the average temporal profile of altered gene expression level for gene assembles indicative of seven stress response categories in *E. coli* for three chemicals, namely MMC, H_2O_2, and Pb^{2+}, across 6 concentrations. The gray bands show the 95% confidence interval of the altered expression level among all the genes in a given pathway among the three replicates. The relatively narrow gray bands suggested that genes selected in each specific stress response pathway (or category) seemed to have high coexpression tendency, which agrees with prior report of the high proportion of coregulations of genes under environmental stresses.[34]

For the three chemicals tested, most stress response pathways exhibited distinct temporal patterns among the different chemicals, and yet, for a specific chemical, most seemed to have conserved temporal trends among all concentrations with dose-dependent magnitude changes (Figure 1). The various temporal patterns for stress response pathways included a relatively constant expression level (such as cell killing pathway when exposed to MMC, Figure 1a); short impulses (such as detoxification pathway induced by H_2O_2, Figure 1b); and a monotonic increasing/decreasing trend spanning over a large expression level range (such as drug resistance pathway in response to Pb^{2+}, Figure 1c). These patterns were consistent with those observed in earlier studies, where both short impulse-like and long sustained gene expression patterns were found in response to environmental stimuli depending on the function of the specific pathway.[35] Although, there were some that showed changing temporal patterns as dose concentration increased, such as those in protein damage and detoxification pathways in response to MMC, which showed transitional pattern changes at higher doses. These patterns are likely reflections of cellular requirements for an immediate remedy or a gradual recovery seeing the severity of the stress, as well as balances among efforts/energy needed for homeostasis and stress responses. The chemical-specific yet conserved temporal trends among genes in a given stress response pathway suggest that these conserved patterns can potentially be chemical-specific indicators for further toxicity mechanism evaluation.

The observed dose concentration-dependent changes in the magnitude of altered gene expression level suggest possible dose-response relationship, which is the central dogma of toxicology.[36] As cellular stress responses to environmental perturbation involve coordinating gene expression in both magnitude and timing, it is desirable to have a means to quantify how temporal pattern changes according to chemical dosage.[37] However, there is not yet a consensus on dose-dependent temporal patterns and molecular toxicity end points-based dose response relationship.[18] A number of previous studies suggested the existence of a relationship between dose

DOI: 10.1021/es505199f
Environ. Sci. Technol. 2015, 49, 4618−4626

and molecular toxicity end points at a single gene or an ensemble of stress response genes.[3,10,18,21,22,24,25,27,36,38−41] In this study, all chemical concentrations applied were at subcytotoxic levels, where the temporal stress responses were likely at the homeostasis stage. The dose-dependent patterns of altered gene expression profiles indicated variations in both magnitude and nature of the molecular stress response systems in response to a toxicant at varying concentrations. The results also demonstrated that time-course experiments could potentially be used to delineate prototypical temporal activation and coregulation of genes.

The above observation that aggregated expression dynamics at the pathway level could be consistent, chemical-specific, and concentration-dependent highlighted the suitability of pathway analysis for toxicity mechanisms studies. The common temporal expression pattern of genes within a particular pathway can be more informative and reliable, as it reflects more cohesive biological effects and avoids the possible inconsistent and likely error-prone results among individual genes.[13] For example, cellular processes often involve sets of genes acting in concert rather than the significant expression of only one gene.[15] For time series toxicogenomics data, the inconsistency in altered gene expression for individual gene among replicates may result from both inherent experimental system error and lack of proper data processing processes, such as of temporal gene expression profile alignment.[42,43]

Impact of Time on GSEA Analysis for Toxicity Mechanisms Identification. Most conventional GSEA analyses have been performed with static gene expression profiling data, which captured the snapshots among the dynamic transcriptional systems with the underlying assumption of stable cellular states. The dynamic temporal differential expression profiles of stress response genes in response to a toxicant as shown in our study suggests that application of GSEA directly to isolated time points would neglect the important time factor and lead to inconsistent conclusions. To demonstrate the caveat, we simulated and compared the GSEA results for three single-time points using our time series data. Altered expression levels of genes after the array being exposed to MMC samples (0.5 μg/L) at 3 time points (30, 65, and 100 min) were isolated and used to identify enriched pathways, respectively. Figure 2 illustrates the statistical significance of 7 stress categories at different times calculated by GSEA. MMC is a known DNA-damage agent and is often used as a model genotoxic compound in environmental toxicity studies.[23] GSEA identified the significant enrichment of DNA damage and repair pathway, but only in the middle of the experiment at 65 min, whereas at the other two time points, other stress responses were more significant. These results demonstrate the dynamic nature of cellular stress responses and highlight the importance of exposure time and the necessity to take temporal patterns into account for the identification of toxicity mechanisms. The temporal variability may be as meaningful as the expression level at a single time point.

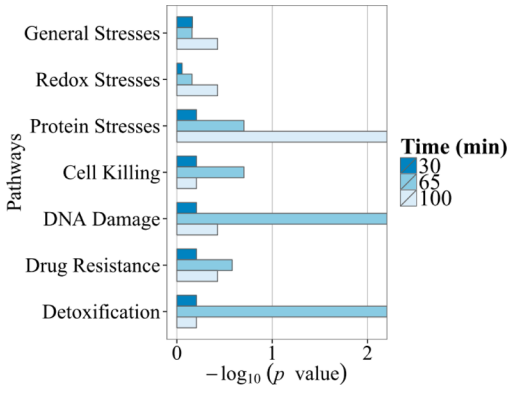


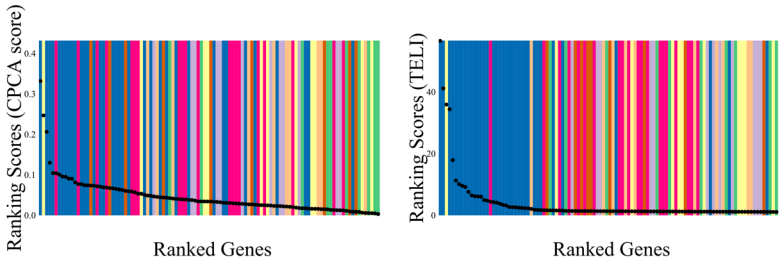
Figure 2. Comparison of GSEA results based on toxicogenomics data at three different time points upon exposure to MMC (0.5 ng/L) (30, 65, and 100 min upon exposure). Genes were ranked by mean altered gene expression levels based on triplicate treatments. x axis bottom: displays p value in negative logarithmic scale.

GSEA of Time Series Toxicogenomics Data-Comparison of Two Score Metrics. To incorporate the time impact into the GSEA, we proposed and employed two gene ranking metrics, namely CPCA score and TELI score. As previously described, CPCA analysis can be applied to rank genes based on their contribution to the temporal variation among different treatments. Another molecular end point-TELI was also applied for gene ranking, which measured the accumulative gene expression level alternation normalized to exposure time, and it represented the average expression level within the experiment period. Figure 3 shows the comparison of GSEA ranking results with both metrics. On the basis of the gene ranking results, the statistical significances of enrichment for 7 stress response pathways are summarized in Table 1.

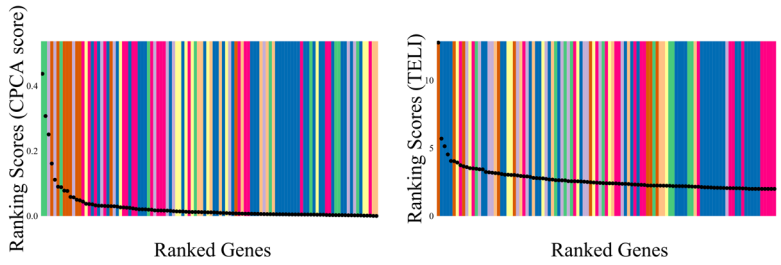
Assessment of the pathway enrichment analysis results against the prior knowledge of the toxicity mechanism of the three model chemicals allows us to evaluate the performance and reliability of the two approaches for gene ranking. For treatments with MMC, DNA damage and repair pathway-related genes were significantly ($p < 0.01$) enriched based on both CPCA and TELI based GSEA analysis, which is consistent with known toxic mechanism of MMC as a genotoxicant. For H_2O_2, CPCA-based GSEA analysis seemed to point to enriched general stress and detoxification response that are consistent with known toxic effects of H_2O_2 as a strong oxidant with wide and multiple cellular impacts. TELI-based analysis that reflects more of the averaged cumulative altered gene expression changes, yield only detoxification category to be significantly enriched. There was also discrepancy between GSEA results for Pb^{2+} with either CPCA or TELI ranking method. CPCA score pinpointed significant activation of drug resistance pathway, whereas results based on TELI did not identify any significantly altered stress response pathway (based on $p < 0.0.5$), although it indicated stronger redox stress response activities than others ($p = 0.07$).[44]

CPCA-based GSEA gives more weight to genes that have large temporal variation during exposure time, and less weight to those that have low level but sustained level of expression. For example, a gene that exhibits a sharp impulse response would be ranked higher than one that has lower but sustained altered expression level over the exposure time period. This "high temporal variance high ranking" assumption is somewhat ambiguous. It is recognized that the temporal regulation processes, the expression pattern, and gene/pathway involvement could be too complex to be represented by a simple variance. More drastic and higher magnitude of expression change has been often assumed to indicate likely involvement and responsiveness of a gene to external stimulus of toxicant. However, genes that exhibited sustained but lower level of altered expression may also be vital for certain pathway activities depending on the gene function and its maximum level of fold change (i.e., the maximum level of alteration may

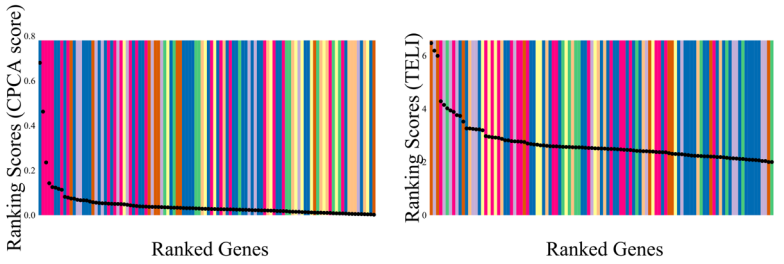
DOI: 10.1021/es505199f
Environ. Sci. Technol. 2015, 49, 4618−4626



(a) GSEA ranking results for MMC (0.5 μg/L) based on CPCA score (left) and TELI (right)



(b) GSEA ranking results for H_2O_2 (10mg/L) based on CPCA score (left) and TELI (right)



(c) GSEA ranking results for Pb^{2+} (0.125 μg/L) based on CPCA score (left) and TELI (right)

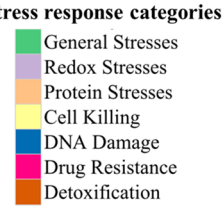


Figure 3. Comparison of GSEA gene ranking results for altered gene expression results for (a) MMC at 0.5 μg/L, (b) H_2O_2 at 10 mg/L, and (c) Pb^{2+} at 0.125 μg/L using two different ranking metrics CPCA score (left) and TELI (right). Genes are positioned based on their ranking score in nondecreasing order from left to right based (score values for each gene are listed in SI Tables S6–S11). Each vertical strip represents a gene, with color-code indicating its associated stress response pathway and the dot displaying its ranking score. The results are based on an average of triplicate treatments. The color codes for each stress response categories or pathways labels are shown in the legend at the bottom.

Table 1. Comparison of GSEA Results (p-Values) Using Two Different Ranking Metrics, Namely CPCA Score and TELI, Respectively[a]

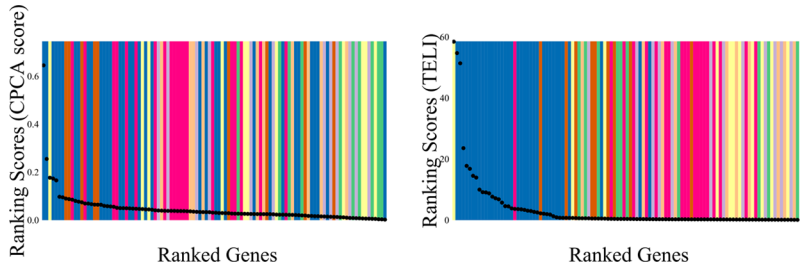
	MMC		H_2O_2		Pb^{2+}	
pathways	CPCA	TELI	CPCA	TELI	CPCA	TELI
general stress	1.00	1.00	0.00	0.61	1.00	0.88
redox stress	1.00	1.00	0.20	0.07	0.49	0.07
protein stress	1.00	1.00	0.99	0.61	1.00	0.88
cell killing	0.13	0.12	0.99	0.17	1.00	0.47
DNA damage	0.00	0.00	0.99	0.25	0.08	0.99
drug resistance	1.00	1.00	0.81	0.91	0.01	0.47
detoxification	1.00	1.00	0.02	0.00	0.49	0.16

[a]Data were based on time series stress response altered gene expression profiling data after exposure to MMC (0.5 μg/L), H_2O_2 (10 mg/L) and Pb^{2+} (0.125 μg/L). p-values are corrected for multiple comparisons using false-discovery rate method.[15]

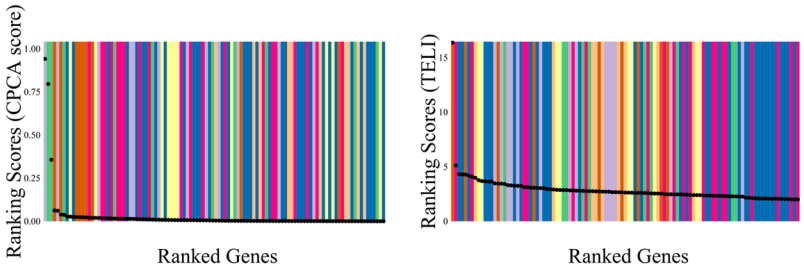
vary for different genes).[13] The overall results indicated that CPCA-based GSEA appear to identify some toxicity mechanisms that were missed by TELI-based GSEA, suggesting the likely higher importance of temporal dynamics of gene expression than the accumulative magnitude alone, at least in the case of stress responses. It should be pointed out that the CPCA applied here also extracted common variation information from multiple samples such as replicates treatments, which helped to further reduce systematic experiment errors.

Impact of Concentrations on Toxicity Mechanisms Identification. Toxic response of a given chemical as revealed by toxicogenomics data can vary with dose concentrations, as previously discussed. The molecular toxicological responses appeared to be dose-dependent (SI Tables S3–S5), therefore toxicant concentration would impact the GSEA results for toxicity mechanism identification.[19,27,36,40] Here for each

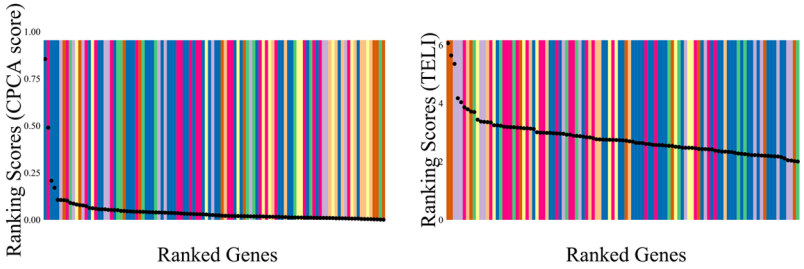
DOI: 10.1021/es505199f
Environ. Sci. Technol. 2015, 49, 4618–4626



(a) GSEA ranking results for MMC based on CPCA score (left) and TELI (right) .



(b) GSEA Ranking results for H_2O_2 based on CPCA score (left) and TELI (right) .



(c) GSEA Ranking results for Pb^{2+} based on CPCA score (left) and TELI (right) .

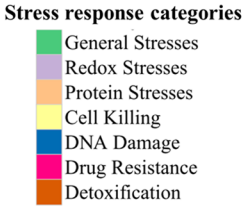


Figure 4. Comparison of GSEA analysis gene ranking results with consideration of all six dose concentrations for (a) MMC, (b) H_2O_2, and (c) Pb^{2+} using two different ranking metrics CPCA score (left) and TELI (right) for each chemical. Genes are positioned in nondecreasing order from left to right based on the metrics. Each vertical strip represents a gene, with color indicating its associated pathway and dot displaying ranking score. The results are based on results (triplicate treatments) of all six dose concentrations for each chemical. The color codes for each stress response category or pathway label are shown in the legend at the bottom.

chemical, we compare the results between GSEA using data from samples with a single concentration and that using collective data including samples from all doses. By inspecting the difference, we could see whether dose is an important factor influencing the analysis outcome.

Figure 4 and Table 2 show the GSEA results and statistical significances analysis of seven stress response pathways. For MMC, GSEA that considered the common genes among all dose concentrations led to similar conclusions as that with one single concentration. Both TELI-based and CPCA-based GSEA identified DNA damage as main molecular effect of MMC, which is consistent with the known toxicity mechanism of MMC, a DNA-damaging model compound. TELI-based GSEA also isolated significantly enriched pathway activity in cell killing. For H_2O_2, the CPCA-based GSEA analysis using data that incorporated various dose concentrations, however, yielded a different outcome from those based on one single dose concentration that indicated the significant activities of genes involved in general stress and detoxification to mostly general stress and redox stress. Both CPCA-based GSEA analysis considering multiple dose concentrations suggested activation of redox stress (p = 0.07), the known toxic mechanism of oxidant H_2O_2, which was not detected by analysis with one single dose concentration (Table 1). TELI-based analysis with multiple concentrations led to a similar outcome as that with a single concentration. For Pb^{2+}, the CPCA-GSEA identified dominant and significant enrichment of pathways that changed from mainly drug resistance at single concentration to DNA-damage for multiple concentrations, suggesting possible dose-depending shifts in molecular response patterns. The results for TELI-based GSEA with data incorporating multiple concentrations, pointed toward significant redox stress and detox-

DOI: 10.1021/es505199f
Environ. Sci. Technol. 2015, 49, 4618−4626

Table 2. Comparison of GSEA Results with Consideration of Six Dose Concentrations Using Different Ranking Metrics, CPCA Score and TELI[a]

	MMC		H_2O_2		Pb^{2+}	
pathways	CPCA	TELI	CPCA	TELI	CPCA	TELI
general stress	1.00	1.00	0.03	0.68	0.50	0.52
redox stress	1.00	1.00	0.07	0.07	0.40	0.04
protein stress	1.00	1.00	0.99	0.34	1.00	0.21
cell killing	0.50	0.00	0.80	0.47	0.52	0.52
DNA damage	0.00	0.00	0.99	0.14	0.01	0.99
drug resistance	0.62	1.00	0.80	0.62	0.40	0.08
detoxification	0.21	1.00	0.34	0.00	0.40	0.02

[a]Data were based on time series data after exposure to MMC, H_2O_2 and Pb^{2+} at six different dose concentrations for each and in triplicates. *p*-Values are corrected for multiple comparisons using false-discovery rate method.

ification response activities, which were not identified with a single concentration. These results are generally consistent with prior toxicological knowledge, as Pb^{2+} is known to cause oxidative stresses that can consequently lead to cellular level multiple responses, including DNA damage and detoxification.[41]

The above results demonstrated that concentration affects molecular toxic response profiles and the pathway activation revealed by GSEA could yield different outcome depending on the ranking metric employed, as well as on the toxicity nature of the chemical. This is consistent with the previous statistical analysis that revealed the impact of both time and dose on gene expression levels (SI Tables S3−S5). The impact of concentration on pathway activation and toxicity mechanism identification may be more pronounced for chemicals that exhibit dose-dependent toxicity mechanisms, such Pb^{2+}. Others may exhibit more chemical-specific and not so concentration-sensitive toxicity effects such as MMC. Designing and implementing particular techniques, such as CPCA, presents the ability to extract and identify genes that show more common behavior among all the dose concentrations and indicate the conserved chemical-specific toxic effects. However, it can be argued that it overly focuses on the commonality and may overlook the dose-dependent specificity of toxicity nature. Although dose-dependent toxicity is the central dogma of toxicology at the phenotypic level, whether this holds true for molecular toxicity as revealed by toxicogenomics studies remains largely unknown. However, there were some successful cases that applied molecular end points to address molecular dose−response relationships[10,18,25]

In conclusion, we employed CPCA-based gene set enrichment analysis with two ranking metrics against high-dimension time series toxicogenomics data, with the aim to evaluate the impacts of time and concentration on the determination of significance of pathway activation and identification of chemical toxicity mechanisms. Our results demonstrated that both time and dose concentration impact the altered gene expression profiles, therefore the consequent GSEA outcome. Comparison of the two gene ranking methods and metric indicated that choice of ranking score matric may lead to inconsistent GSEA results and toxicity evaluation conclusions as the results of differences in the underlying assumptions, logic, and the weight given to the genes. Employment of a ranking metric that has improved ability to capture the temporal dynamics in gene expression pattern, as well as the consideration of dose-dependent toxic responses, as demonstrated in this study, is expected to potentially lead to more accurate identification of toxicity mechanisms of a chemical.

■ ASSOCIATED CONTENT

Supporting Information

Table S1 lists the information for the genes included in the *E. coli* stress assay library for toxicogenomic assessment. Table S2 lists the toxicity mechanism and dose concentrations for the chemicals that were evaluated in this study. Table S3−S5 show repeated measures ANOVA test on the impacts of concentration and time factors on the gene expression levels. Figure S1 illustrates the different ranking results for each chemical at different concentrations. The ranking score information for Figure 3 has been summarized in Tables S6−S11. The ranking score information for Figure 4 has been summarized in Tables S12−S17. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 617-373-3631; fax: 617-373-4419; e-mail: april@coe.neu.edu (A.Z.G.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF, EEC-0926284, CAREER CBET-0953633, and CBET-1440764), National Institute of Environmental Health Sciences (NIEHS) (PROTECT P42ES017198) and CDM/Diane and Bill Howard Scholarship.

■ REFERENCES

(1) Krewski, D.; Acosta, D., Jr; Andersen, M.; Anderson, H.; Bailar, J. C., III; Boekelheide, K.; Brent, R.; Charnley, G.; Cheung, V. G.; Green, S., Jr Toxicity testing in the 21st century: a vision and a strategy. *J. Toxicol. Environ. Health, B* **2010**, *13* (2−4), 51−138.

(2) Hayes, K. R.; Bradfield, C. A. Advances in toxicogenomics. *Chem. Res. Toxicol.* **2005**, *18* (3), 403−414.

(3) Gao, C.; Weisman, D.; Gou, N.; Ilyin, V.; Gu, A. Z. Analyzing high dimensional toxicogenomic data using consensus clustering. *Environ. Sci. Technol.* **2012**, *46* (15), 8413−8421.

(4) Bar-Joseph, Z. Analyzing time series gene expression data. *Bioinformatics* **2004**, *20* (16), 2493−2503.

(5) Bar-Joseph, Z.; Gitter, A.; Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **2012**, *13* (8), 552−564.

(6) National Research Council (U.S.) Committee on Applications of Toxicogenomic Technologies to Predictive Toxicology. *Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment*; National Academies Press: Washington, D.C., 2007.

(7) Jayapal, M. Integration of next-generation sequencing based multi-omics approaches in toxicogenomics. *Front. Genet.* **2012**, *3*, 88.

(8) Melamed, S.; Elad, T.; Belkin, S. Microbial sensor cell arrays. *Curr. Opin. Biotechnol.* **2012**, *23* (1), 2−8.

(9) Timothy, S.; O'Connor, F.; Lan, J.; North, M.; Loguinov, A.; Zhang, L.; Smith, M. T.; Gu, A. Z.; Vulpe, C. Genome-wide functional and stress response profiling reveals toxic mechanism and genes required for tolerance to benzo[*a*]pyrene in *S. cerevisiae*. *Front. Genet.* **2012**, *3*, 316.

(10) Gou, N.; Gu, A. Z. A new Transcriptional Effect Level Index (TELI) for toxicogenomics-based toxicity assessment. *Environ. Sci. Technol.* **2011**, *45* (12), 5410−5417.

DOI: 10.1021/es505199f
Environ. Sci. Technol. 2015, 49, 4618−4626

(11) Schliep, A.; Costa, I. G.; Steinhoff, C.; Schonhuth, A. Analyzing gene expression time-courses. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2005**, *2* (3), 179−193.
(12) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37* (1), 1−13.
(13) Khatri, P.; Sirota, M.; Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **2012**, *8* (2), e1002375.
(14) Mootha, V. K.; Lindgren, C. M.; Eriksson, K.-F.; Subramanian, A.; Sihag, S.; Lehar, J.; Puigserver, P.; Carlsson, E.; Ridderstråle, M.; Laurila, E.; Houstis, N.; Daly, M. J.; Patterson, N.; Mesirov, J. P.; Golub, T. R.; Tamayo, P.; Spiegelman, B.; Lander, E. S.; Hirschhorn, J. N.; Altshuler, D.; Groop, L. C. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **2003**, *34* (3), 267−273.
(15) Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (43), 15545−15550.
(16) Calvano, S. E.; Xiao, W.; Richards, D. R.; Felciano, R. M.; Baker, H. V.; Cho, R. J.; Chen, R. O.; Brownstein, B. H.; Cobb, J. P.; Tschoeke, S. K.; Miller-Graziano, C.; Moldawer, L. L.; Mindrinos, M. N.; Davis, R. W.; Tompkins, R. G.; Lowry, S. F. Inflamm; Host Response to Injury Large Scale Collab. Res, P., A network-based analysis of systemic inflammation in humans. *Nature* **2005**, *437* (7061), 1032−1037.
(17) Grigoryev, Y. A.; Kurian, S. M.; Avnur, Z.; Borie, D.; Deng, J.; Campbell, D.; Sung, J.; Nikolcheva, T.; Quinn, A.; Schulman, H.; Peng, S. L.; Schaffer, R.; Fisher, J.; Mondala, T.; Head, S.; Flechner, S. M.; Kantor, A. B.; Marsh, C.; Salomon, D. R. Deconvoluting post-transplant immunity: cell subset-specific mapping reveals pathways for activation and expansion of memory T, monocytes and B cells. *PLoS One* **2010**, *5* (10), e13358.
(18) Allen, B. C.; Kavlock, R. J.; Kimmel, C. A.; Faustman, E. M. Dose-response assessment for developmental toxicity. II. Comparison of generic benchmark dose estimates with no observed adverse effect levels. *Toxicol. Sci.* **1994**, *23* (4), 487−495.
(19) Daston, G. P. Gene expression, dose-response, and phenotypic anchoring: applications for toxicogenomics in risk assessment. *Toxicol. Sci.* **2008**, *105* (2), 233−234.
(20) Andersen, M. E.; H, J. C., III; Bermudez, E.; Willson, G. A.; Thomas, R. S. Genomic signatures and dose-dependent transitions in nasal epithelial responses to inhaled formaldehyde in the rat. *Toxicol. Sci.* **2008**, *105* (2), 368−383.
(21) Ahlborn, G. J.; Nelson, G. M.; Ward, W. O.; Knapp, G.; Allen, J. W.; Ouyang, M.; Roop, B. C.; Chen, Y.; O'Briend, T.; Kitchin, K. T.; Delker, D. A. Dose response evaluation of gene expression profiles in the skin of K6/ODC mice exposed to sodium arsenite. *Toxicol. Appl. Pharmacol.* **2008**, *227* (3), 400−416.
(22) Mezentsev, A.; Amundson, S. A. Global gene expression responses to low- or high-dose radiation in a human three-dimensional tissue model. *Radiat. Res.* **2011**, *175* (6), 677−88.
(23) Onnis-Hayden, A.; Weng, H.; He, M.; Hansen, S.; Ilyin, V.; Lewis, K.; Gu, A. Z. Prokaryotic real-time gene expression profiling for toxicity assessment. *Environ. Sci. Technol.* **2009**, *43* (12), 4574−4581.
(24) Gou, N.; Onnis-Hayden, A.; Gu, A. Z. Mechanistic toxicity assessment of nanomaterials by whole-cell-array stress genes expression analysis. *Environ. Sci. Technol.* **2010**, *44* (15), 5964−5970.
(25) Gou, N.; Yuan, S.; Lan, J.; Gao, C.; Alshawabkeh, A. N.; Gu, A. Z. A quantitative toxicogenomics assay reveals the evolution and nature of toxicity during the transformation of environmental pollutants. *Environ. Sci. Technol.* **2014**, *48* (15), 8855−8863.
(26) Zaslaver, A.; Bren, A.; Ronen, M.; Itzkovitz, S.; Kikoin, I.; Shavit, S.; Liebermeister, W.; Surette, M. G.; Alon, U. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nat. Methods* **2006**, *3* (8), 623−628.
(27) Peddada, S. D.; Lobenhofer, E. K.; Li, L.; Afshari, C. A.; Weinberg, C. R.; Umbach, D. M. Gene selection and clustering for time-course and dose−response microarray experiments using order-restricted inference. *Bioinformatics* **2003**, *19* (7), 834−841.
(28) Keseler, I. M.; Mackie, A.; Peralta-Gil, M.; Santos-Zavaleta, A.; Gama-Castro, S.; Bonavides-Martínez, C.; Fulcher, C.; Huerta, A. M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Muñiz-Rascado, L.; Ong, Q.; Paley, S.; Schröder, I.; Shearer, A. G.; Subhraveti, P.; Travers, M.; Weerasinghe, D.; Weiss, V.; Collado-Vides, J.; Gunsalus, R. P.; Paulsen, I.; Karp, P. D. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* **2013**, *41* (Databaseissue), D605−D612.
(29) Dyk, T. K. V.; Wei, Y.; Hanafey, M. K.; Maureen, Dolan; Reeve, M. J. G.; Rafalski, J. A.; Rothman-Denes, L. B.; LaRossa, R. A. A genomic approach to gene fusion technology. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (5), 2555−2560.
(30) Hollander, M.; Wolfe, D. A. *Nonparametric Statistical Methods*; Wiley: New York, 1999.
(31) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **1995**, 289−300.
(32) Jolliffe, I. *Principal Component Analysis*; Wiley Online Library: 2005.
(33) Raychaudhuri, S.; Stuart, J. M.; Altman, R. B. In *Principal components analysis to summarize microarray experiments: application to sporulation time series*, Pac. Symp. Biocomput., 2000; NIH Public Access: 2000; p 455.
(34) Chung, H. J.; Bang, W.; Drake, M. A. Stress response of *Escherichia coli*. *Compr. Rev. Food Sci. Food Safety* **2006**, *5* (3), 52−64.
(35) Yosef, N.; Regev, A. Impulse control: temporal dynamics in gene transcription. *Cell* **2011**, *144* (6), 886−896.
(36) Altshuler, B. Modeling of dose-response relationships. *Environ. Health Perspect.* **1981**, *42*, 23−27.
(37) Chechik, G.; Koller, D. Timing of gene expression responses to environmental changes. *J. Comput. Biol.* **2009**, *16* (2), 279−290.
(38) Nevozhay, D.; Adams, R. M.; Murphy, K. F.; Josić, K.; Balázsi, G. Negative autoregulation linearizes the dose−response and suppresses the heterogeneity of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (13), 5123−5128.
(39) Kærn, M.; Elston, T. C.; Blake, W. J.; Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **2005**, *6* (6), 451−464.
(40) Burgoon, L. D.; Zacharewski, T. R. Automated quantitative dose-response modeling and point of departure determination for large toxicogenomic and high-throughput screening data sets. *Toxicol. Sci.* **2008**, *104* (2), 412−418.
(41) Flora, G.; Gupta, D.; Tiwari, A. Toxicity of lead: A review with recent updates. *Interdiscip. Toxicol.* **2012**, *5* (2), 47−58.
(42) Aach, J.; Church, G. M. Aligning gene expression time series with time warping algorithms. *Bioinformatics* **2001**, *17* (6), 495−508.
(43) Lin, T.-h.; Kaminski, N.; Bar-Joseph, Z. Alignment and classification of time series gene expression in clinical studies. *Bioinformatics* **2008**, *24* (13), i147−i155.
(44) Goering, P. L. Lead-protein interactions as a basis for lead toxicity. *Neurotoxicology* **1992**, *14* (2−3), 45−60.

DOI: 10.1021/es505199f
Environ. Sci. Technol. 2015, 49, 4618−4626