Environmental Science & Technology

Article

pubs.acs.org/est

Analyzing High Dimensional Toxicogenomic Data Using Consensus Clustering

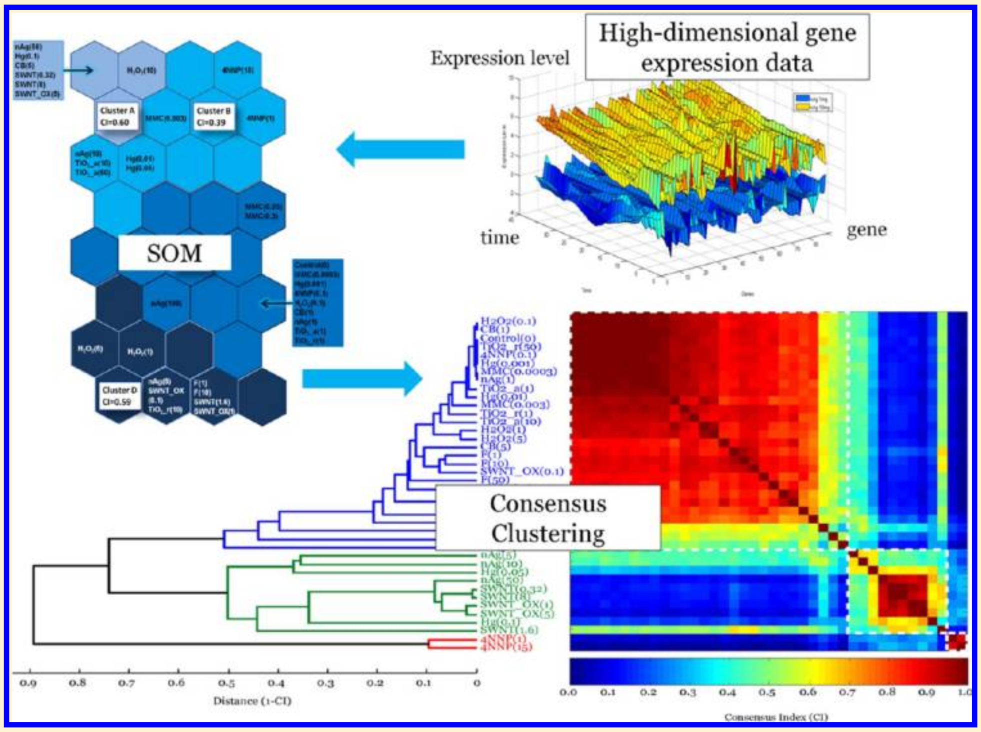
Ce Gao,[†] David Weisman,[‡] Na Gou,[†] Valentine Ilyin,[§] and April Z. Gu*[,†]

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

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

[§]Biology Department, Boston College, Chestnut Hill, Massachusetts 02467, United States

S *Supporting Information*



ABSTRACT: Rapid development of high-throughput toxicogenomics technologies has created new approaches to screen environmental samples for mechanistic toxicity assessment. However, challenges remain in the analysis, especially clustering of the resulting high-dimensional data. Because of the lack of commonly accepted validation methods, it is difficult to compare clustering results between studies or to identify the key experimental or data features that impact the clustering results. We applied consensus clustering (CC), an approach that clusters the input data repeatedly through iterative resampling, and identifies frequently occurring high-confidence clusters. We used CC to analyze a set of high dimensional transcriptomics data with temporal resolution, which were generated using our *E. coli* whole-cell array system for a diverse variety of toxicants at different dose concentrations. The CC analysis allowed us to evaluate the clustering results' robustness and sensitivity against a number of conditions that represent the common variations in high-throughput experiments, including noisy data, subsets of treatments, subsets of reporter genes, and subsets of time points. We demonstrated the value of utilizing rich time-series data and underscored the importance of careful selection of sampling times for a given experimental system. The results also indicated that temporal data compression using our proposed Transcriptional Effect Level Index (TELI) concept followed by CC largely conserved the cluster resolution. We also found that for our cellular stress response ensemble-based high-throughput transcriptomics assay platform, the size and composition of the reporter gene set are critical factors that affect the resulting coherency of clusters. Taken together, these results demonstrated that more robust consensus clustering such as CC may be valuable in analyzing high-dimensional toxicogenomic data sets.

1. INTRODUCTION

Rapid progress in high-throughput screening (HTS) technology, such as in vitro cell line-based cytotoxicity assays and toxicogenomics (e.g., gene expression microarrays) promises

Received: January 13, 2012
Revised: April 29, 2012
Accepted: June 15, 2012
Published: June 15, 2012

dx.doi.org/10.1021/es3000454 | *Environ. Sci. Technol.* 2012, 46, 8413−8421

new approaches for toxicological studies and screening. These HTS methods reduce the conventional need for laboratory animal testing and shorten the time needed for data collection and analysis.[1−3] USEPA has initiated ToxCast programs that have applied HTS approaches for prioritizing environmental chemicals in a more cost-effective manner.[4] The "omics" techniques, including genomics, proteomics, and metabolomics, can reveal toxicant perturbations of complex biological pathways, identify common patterns in mechanism of action, and assist the development of quantitative structure−activity relationship (QSAR) models.[5−7] Recent examples of omics toxicity studies include investigation of the mechanism of action,[8] analysis of toxicity-related signaling pathways,[9] and assessment of nanomaterial toxicity,[10−12] as well as development of biosensors for environmental monitoring.[13−15]

High-dimensional toxicogenomic data represent the responses of numerous transcripts, proteins, metabolites, or reporters to toxicants. These responses are functions of the specific cell line phenotypes, the specific toxicants, and their dose concentrations, as well as the time points of measurement.[16] In addition, cellular responses can be affected by confounding factors such as the media composition, degree of mechanical agitation during treatment and the homogeneity of the toxicant in the growth medium. The variations in experimental design, as well as normal biological and technical measurement noise, make interstudy comparison and biological generalization difficult.[17]

To discover biologically meaningful patterns and relationships, and to identify biological similarities in the responses to different toxicants from the complex high-dimensional omics data sets, computational data mining and statistical analysis techniques have been applied.[18] The most frequently used methods are unsupervised clustering algorithms such as hierarchical clustering and self-organizing maps (SOM).[19,20] These clustering algorithms partition the input data into nonoverlapping subsets, which represent the underlying biological similarities.[21,22] Dendrogram visualization of hierarchical clustering has been widely applied in biological studies, however, the hierarchical clustering results can be sensitive to outliers, and the visual representation neither conveys the number of distinct biological clusters, nor the clear relationships between those clusters.[23−25] Clustering by SOM is also conventionally used for discovering groups of coexpressed genes or classifying chemicals.[26] SOM is relatively resistant to noisy data and missing observations, both common in biological research.[27] However, the SOM algorithm can also be sensitive to the random process of map initialization, which makes comparisons between studies challenging.[28]

These challenges in analyzing HTS toxicogenomic data point to the pressing need of a method for validation of clustering results and assessment of cluster robustness. Current validation schemes include internal approaches, which are based solely on analysis within a data set, and external approaches, which compare clusters with known class labels.[29,30] It is argued that the internal measures might not be suitable for biological data which are subject to high noise levels.[31] To address clustering stability and validation, Monti et al. developed the consensus clustering (CC) method, which iteratively resamples the input data set and invokes a conventional clustering algorithm over the reconstructed data set.[32] Afterward, CC aggregates the ensemble of clustering outputs into a single consensus result, which has a higher confidence level than a typical single invocation of the underlying clustering algorithm. Over the past few years, several researchers have applied this approach in microarray-based gene expression research, and found that CC could identify reliable clusters and hidden patterns, and therefore provide a highly meaningful interpretation of the biological response.[33−35] Evaluation of the CC methods using a relatively large toxicogenomic data set with temporal resolution has not yet been reported.

In this study, we employed CC to analyze a high dimensional transcriptomic data set with temporal resolution, which was generated using our *E. coli* whole-cell array HTS system treated with a diverse variety of toxicants including endocrine disrupting chemicals, heavy metals, antibiotics, and nanomaterials, representing a range of possible toxicity mechanisms.[11] This HTS system contains a library of 91 GFP reporter genes representing 10 cellular stress response pathways, and it records real-time transcriptional level responses.[11,12,36] The diversity of chemical treatments is intended to explore a wide region of the reporter state space. For the CC experiments, we focused on SOM as the underlying clustering algorithm. We found that CC/SOM reliably differentiated distinct biological responses between various toxicants at different doses, separating toxicants with distinct toxic mechanisms, and uniting diverse toxicants that cause biologically similar responses. These observations support the hypothesis that robust ensemble clustering algorithms can produce meaningful results for chemical classification based on the toxic mechanisms of toxicants.[19] In addition, the CC analysis allowed us to evaluate the cluster robustness and sensitivity to a number of conditions that represent the common variations in HTS experiments, including noisy data, subsets of treatments, subsets of reporter genes and subsets of time points. The outcome of these experiments helps identify the key data features that impact the clustering results and provides information and insights for standardizing the practice in this field.

2. METHODS

2.1. Data Generation. The whole-cell array used for HTS toxicity tests was constructed through transcriptional fusions of green fluorescent protein (GFP) to 91 stress response-related promoter genes in *E.coli* K12, MG1655.[6,36,37] The selected genes cover a variety of genes involved in different known cellular stress response pathways, such as general stress, DNA damage, protein stress, redox stress, energy stress, heat shock, drug resistance, detoxification, cell killing, and other functions.[11,12] Each fusion was expressed from a low-copy plasmid, pUA66 or pUA139, that contains a kanamycin resistance gene and a fast folding *gfpmut2*, allowing for real-time measurement of the promoter activities.[6,36]

The 11 chemicals tested included 4 model chemicals with known toxic mechanisms, such as mitomycin C (MMC), mercury (Hg), hydrogen peroxide (H_2O_2), and 4-nonylphenol (4NNP). The remaining are different nanomaterials, including carbon black (CB), nano silver particles (nAg), nano titanium dioxide rutile (TiO_2_r), nano titanium dioxide anatase (TiO_2_a), fullerene soot (F), single-walled nanotube (SWNT), and oxidized single-walled nanotube (SWNT_OX). Detailed characterization information for these nanomaterials is listed in Supporting Information (SI) Tables S1 and S2. The nanomaterials were prepared in M9 medium with 1% of crude bovine serum albumin (BSA) as a dispersant.[11] The stock solutions were dispersed in a 90W sonicator for at least 15 min to increase the homogeneity of the mixture.[11] Other chemicals were dissolved in deionized water before application. For each chemical, 3−5 different dose concentrations were tested (see SI Table S1)

The protocol to measure the temporal gene expression profile was described in our previous reports.[11,12,36] In brief, the *E. coli* was cultivated in 96-well plates (Costar, Bethesda, MD, USA) in dark condition to avoid GFP photobleaching until exponential growth stage (OD_{660} ~0.1) was reached. The toxicants at specific concentrations were added into the microplate wells, and the plate was placed into a microplate reader (Synergy Multi-Mode, Biotek, Winooski, VT) for simultaneous cell growth (absorbance, OD_{660}) measurement and fluorescent readings (GFP level, EX 485 nm, EM 528 nm). Measurements were taken every 3 min over 2 hours, with a total of 36 time points for each experiment, obtained by applying 5-point moving average to the original 40 points. Two biological replicate experiments were performed for each treatment condition. Data for Hg, MMC, nAg, and TiO_2 were reported previously[11,12,36] and all other data were used for the first time in this study.

2.2. Data Preprocessing. The GFP and OD data were first corrected for background (medium control without bacteria, and bacteria control with promoter-less strains). The gene expression level was calculated as P = GFP/OD. Induction factor was calculated as the ratio of expression level between experiment groups and control groups (GFP-fused bacteria without toxicants added), $I = P_e/P_c = (\mathrm{GFP/OD})_{\mathrm{experiment}}/(\mathrm{GFP/OD})_{\mathrm{control}}$. The induction factor measures the gene expression alteration as a result of toxic effect of different toxicants. The natural log of the induction factor $\ln(I)$ was then calculated: when a certain gene is up-regulated, $\ln(I) > 0$, and when it is down-regulated, $\ln(I) < 0$.[6,29] To remove the signals below a noise floor, we filtered the inductor factor values between 0.67 and 1.5 ($-0.4 < \ln(I) < 0.4$), and changed them to 0. In our previous study we defined a toxicity end point-transcriptional effect level index (TELI) for time-series transcriptomic data, which is calculated with the altered gene expression effect level, $\mathrm{TELI}(\mathrm{gene}_p) = \int_0^t (e^{|\ln(I)|}-1)\mathrm{d}t/$ exposure time(t), where I is the induction factor of the p^{th} gene.[12]

To construct the data set to be analyzed, we defined a "treatment" as the gene expression time-series data corresponding to a specific chemical at a specific concentration, T. Each treatment is a 3276 (= 36 × 91) dimensional input row vector

$$T_n = (\ln(I_{1,1}^{(n)}), \ln(I_{1,2}^{(n)})), ...\ln(I_{1,36}^{(n)}), ...\ln(I_{p,q}^{(n)}), ...\ln(I_{91,1}^{(n)}), \ln(I_{91,2}^{(n)}), ...\ln(I_{91,36}^{(n)}) \quad (1)$$

where $\ln(I_{p,q}^{(n)})$ is the expression of the p^{th} gene at the q^{th} time point, under the n^{th} treatment (p = 1,2, ...91, q = 1,2,...36, and n = 1,2,...40). Since the control results were involved in the preprocessing, we define the control vector with all elements being 0. Then we define a "sample" as the whole or part of the overall data set that contained a total of 40 treatments.

To test the relevance of temporal pattern of the gene expression alternation on the clustering analysis, we compared the results of the full time-series with those using one selected time point data, and with those using integrated time-series data represented by the aggregated TELI value.

2.3. Ensemble Clustering and Data Validation: Consensus Clustering (CC). Consensus clustering works through iteratively resampling and clustering the input data to generate a similarity matrix (consensus matrix) of sample, which can be used to predict the number of clusters and to assess cluster stability.[38] The underlying assumption of CC is that high cluster stability, produced from perturbed input data, indicates high confidence in the resulting clustering.[35] The stability of the clustering can be evaluated from the consensus matrix, where each entry is a consensus index (CI), which are calculated as

$$\mathrm{CI}(i,j) = \frac{\sum_h M_{(h)}(i,j)}{\sum_h I_{(h)}(i,j)} \quad (2)$$

where $M_{(h)}$ (i,j) and $I_{(h)}$ (i,j) are the entries of connectivity matrix and indicator matrix in the h^{th} resampling cycle, respectively. $M_{(h)}$ (i,j) is defined to be 1 if the i^{th} and j^{th} treatments are clustered together, and 0 otherwise; $I_{(h)}$ (i,j) is defined to be 1 if the i^{th} and j^{th} treatments both appear in h^{th} resampling. The positive index, $\mathrm{CI}(i,j)$, is an indicator of similarity between treatments, and is within the range of [0,1] following the above formula. If a consensus matrix consists largely of 1's and 0's, one may infer that the sample is well clustered, because the treatments are similar within their own clusters and distinct from those outside. The algorithm of consensus clustering was implemented in MATLAB (version R2011a).

To conduct a CC analysis, a resampling scheme and a basic existing clustering algorithm are first chosen. SOM, which has been effectively used for the exploratory analysis of gene expression data, was used as the underlying clustering algorithm for this study. A SOM is a special case of neural network consisting of nodes organized on a regular typically two-dimensional grid.[39] The high-dimensional sample is projected onto the maps and clustered by proximity. For the resampling algorithm, we chose the bootstrapping with 1000 iterations, as this maintains the size of the original data set, which is necessary for stable SOM performance. Finally, a dendrogram based on average linkage hierarchical clustering, whose distance function is $D(i,j) = 1 - \mathrm{CI}(i,j)$, where $D(i,j)$ is the distance metric between the i^{th} and j^{th} treatments, was generated to show the inner structure of the blocks along the main diagonal in the heatmap. The SOM algorithm in this paper, including map creation, initialization, training, finding best-matching units for treatments on the map using U-matrix, are implemented through using SOM toolbox, a function package for MATLAB (version R2011a). The SOM topology was optimized by SOM toolkit (version 2.0) with a two-dimensional hexagonal 8 × 4 grid (32 units).[39,40] All other parameters were at the default settings.

3. RESULTS AND DISCUSSION

3.1. Consensus Clustering Based on SOM. SOM are frequently used to reveal patterns in gene expression data sets, in part, for their computational scalability to large data sets.[23] By partitioning the input data into a small set of nodes, and by representing the nodes in two dimensions, the SOM intuitively conveys the notion of a similarity neighborhood between nodes. Moreover, it is straightforward to aggregate neighboring nodes into higher-level clusters and represent these aggregate clusters as distinct colors.

We first applied SOM to our full time-series data set and found four distinct clusters, which are shown in Figure 1a. The resulting SOM patterns are mostly consistent with prior toxicological knowledge. The treatments of toxicants at their lowest dose concentrations aggregated closely around the untreated control experiments, indicating their molecular effects are close to detection limit. As dose concentrations increased, chemical-specific transcriptional level effects became more pronounced. For the same chemical, treatments with moderate and high dose levels tend to cluster closely together, as observed for SWNT (0.32 and 8 mg/L), TiO_2_a (10 and 50 mg/L), 4NNP (1 and 15 mg/L), Hg (0.01 and 0.05 mg/L), MMC (0.03 and 0.3 mg/L), H_2O_2 (1 and 5 mg/L), and fullerene (1 and 10 mg/L), indicating that there is conserved similarity in the response patterns for a given chemical at different dose levels. The four model chemicals with different known toxic mechanisms exhibited distinctive profiles and were separated into different clusters. MMC is a model genotoxicant that specifically leads to DNA damage.[36,41]

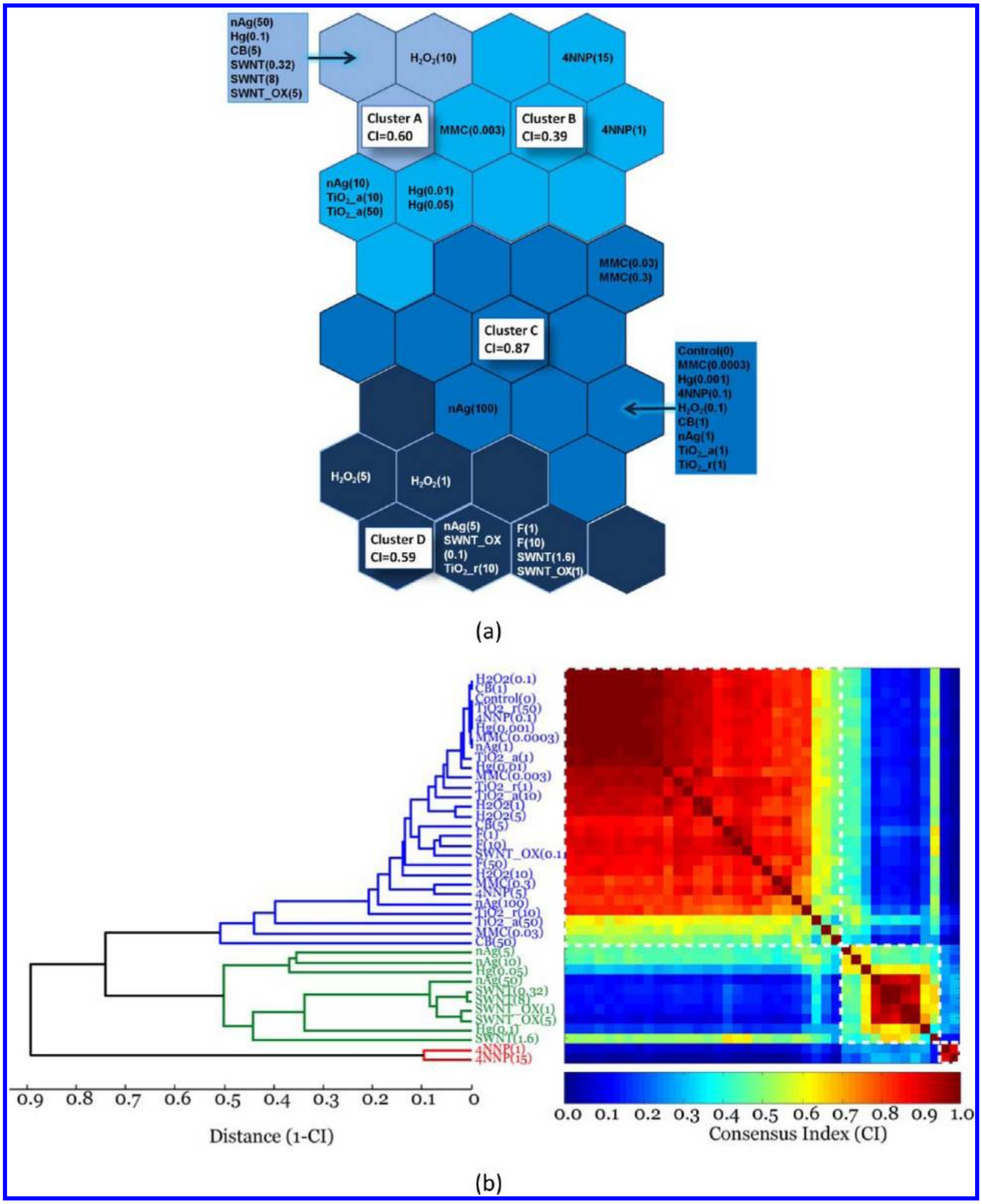


Figure 1. SOM and consensus clustering results for all treatments (concentrations shown in parentheses as mg/L). (a) Self-organizing map (SOM). Treatments are projected onto the map. The differences of gene expression pattern among the treatments are transferred into topographical distance on the map. Hexagon colors represent clusters, which are determined according to treatment distances using the U-matrix. Average CI values shown for each cluster are arithmetic mean among all pairs of treatments within the cluster. (b) Consensus clustering (CC). The consensus matrix represented as a heatmap. A dendrogram is based on the CIs, showing the inner structure in the blocks along the main diagonal. The identified clusters are labeled in different colors in the dendrogram, and dashed squares in the heatmap.

H_2O_2 is a model oxidant that causes oxidative stress via reactive oxygen species (ROS) and ultimately lead to various damages to cell.[42] Examination of the states of various stress genes of *E. coli* in exposure to mercury suggests that it leads to redox stress and DNA damage in the cells.[36] Protein stress and toxicity of 4NNP to *E. coli* strain has been reported.[43,44] The nanomaterials frequently clustered together and their distances to those model compounds indicate their toxic effects. Most nanomaterials clustered with H_2O_2, suggesting their dominant oxidative stress-related toxicity, which is consistent with literature reports.[11,12,45,46]

To quantitatively evaluate the quality and consistency of the resulting SOM, we employed CC, which iteratively resamples the input data set to recompute new SOMs, and measures the consistency of the resulting clusters. For a given pair of gene expression inputs, the consensus index (CI) represents the relative frequency that these inputs coclustered together. A CI value of 1 indicates that the two inputs always coclustered, while a CI element of 0 indicates that the inputs never coclustered. The theoretically ideal result occurs when all elements of the consensus matrix are either 1 or 0; however, in practice, resampling and reclustering produces intermediate consensus values between 1 and 0. As shown in Figure 1a, in our study, 4 distinct clusters are found and the average CI values (arithmetic mean among all pairs of treatments within the cluster) vary widely, ranging from 0.39 to 0.87. This large variance underscores the sensitivity of SOM to perturbations of the input data, and illustrates one difficulty in comparing results between high throughput screening toxicological studies. In addition to the stochastic effects of resampling, some of this variance may be explained by the inherent SOM random ordering of its input processing, coupled with inherent SOM sensitivity to initial conditions.[47]

The CI matrix, by indicating the frequencies of coclustered input pairs, can also be interpreted as a matrix of pairwise similarity scores. Since two highly similar input vectors likely cluster together frequently, their CI score tends toward 1. By the same logic, two highly dissimilar input vectors would cocluster and their CI score tends toward 0. The clustering results using this CC approach are illustrated as consensus matrix (Figure 1b). The second-level clustering reordered the rows and columns identically, such that blocks along the diagonal indicate high confidence consensus clusters. This approach revealed three distinctive clusters, labeled in different colors. The optimal cluster number was confirmed by finding the peak cumulative distribution function area as proposed by Monti et al.[32] (SI Figure S1a). The average CI value for each cluster ranged from 0.56 to 1, which was higher than those with SOM (0.39−0.87), indicating higher statistical robustness of the CC results. In addition, there are gradations in CI within each cluster. This subtlety may be lost in conventional single-pass SOM clustering, in which neighboring nodes are clustered based on the aggregate similarities of the representative node vectors (Figure 1a).[40]

In addition to the improvements in reliability, resolution, and visual presentation of the clustering results by CC approach compared to SOM results, the chemical-specific clusters are more consistent among chemicals with CC than the SOM results, and they appear in better agreement with prior understanding of the toxic mechanisms of these toxicants. For example, all SWNT treatments (except the one at lowest concentration) clustered as one group, which is consistent with the current knowledge that carbon nanotubes cause toxicity mainly via oxidative damage due to their strong ability to mediate electron transport and generate reactive oxygen species (ROS).[48] More impressively, the CI revealed more clear distinction between original SWNT versus those that had surface modification—SWNT_OX. The surface characterization changes in the SWNT_OX resulted in the subtle changes in its toxic response profiles and this phenomenon has been reported by toxicological studies.[49,50] Two nanotitanium oxides were tested, namely TiO_2_rutile (TiO_2_r) and TiO_2_anatase (TiO_2_a). TiO_2_a is known to be more toxic than TiO_2_r and they were found to lead to both oxdidative stress and DNA damage.[12] Indeed, the clustering results showed that TiO_2 clustered more closer with both DNA-damager MMC and oxidant H_2O_2. TiO_2_r has denser arrangement of atoms and higher stability, which may explain its apparent lower toxicity at similar concentrations compared to TiO_2_a at transcriptional effect level, especially in nanosize.[51] nAg, a metal nanoparticle clustered with toxic metal Hg at higher dose concentrations, indicating likely metal-specific toxic response and both of which are known to lead to oxidative stress, DNA damage, and protein stress.[12] Consistent with the SOM results, the toxicants at the lowest dose levels (near detection limit) clustered most closely with controls as expected and the treatments for the same chemical at different dose concentrations all cluster together. The chemical-sepcific response profiles became more distinctive as the dose concentration increased. Overall, the results demonstrated that the stress-response pathways ensemble-based HTS toxicity assays yield chemical-specific and concentration-sensitive transcriptomic profiles. Moreover, statistically reliable clustering such as CC based on SOM algorithm is capable of identifying classes of chemicals according to their underlying toxic mechanisms.

3.2. Impact of High Dimensional Nature of the Data: Resolution Power from Time-Series. The data sets analyzed above take advantage of the full offset of measurement time points, which presumably provides substantial information that helps distinguish the treatment clusters. To test that hypothesis, we performed three more experiments using different features extracted from the time series. The first used gene expression data at the middle time point of the testing (the 20th time point, at 57 min after measurement began). The second experiment employed the maximum gene expression signal observed for each reporter during the 2-h assay (Max), and the third experiment used a derived integrated end point-Transcriptional Effect Level Index (TELI). TELI integrates a reporter signal over time to reflect more of the temporal pattern of the reporter response than a single time point, effectively compressing a number of time points into a compact representation.[12,52] These selected data features mimic three conventional experiment designs.

In CC analysis, the treatments with high consensus index from different nodes in the dendrograms eventually merge into a common root. Higher average CI value indicates higher similarity among the treatments within the node. To visualize the quality of clustering with different subsets of time-series toxicogenomics data, we analyzed the consensus matrices and calculated the average consensus indices for each node and ordered them in an increasing fashion. Figure 2 shows the comparison of average consensus index versus ordered node number curves based on dendrograms generated with the four different data subsets described above. Lower consensus index values indicated nodes close to the root, reflecting those connecting a grouping of treatments with distinct transcriptomic profiles; while higher values indicate nodes that could represent a real cluster, a group of treatments exhibiting similar toxic response profiles. A good clustering produces nodes that have a distinct jump of the average node CI value, that is, a large difference between the intra and inter cluster values. Otherwise, nodes with similar levels of average consensus index reflect a weak consensus, a sign of lower resolution power.

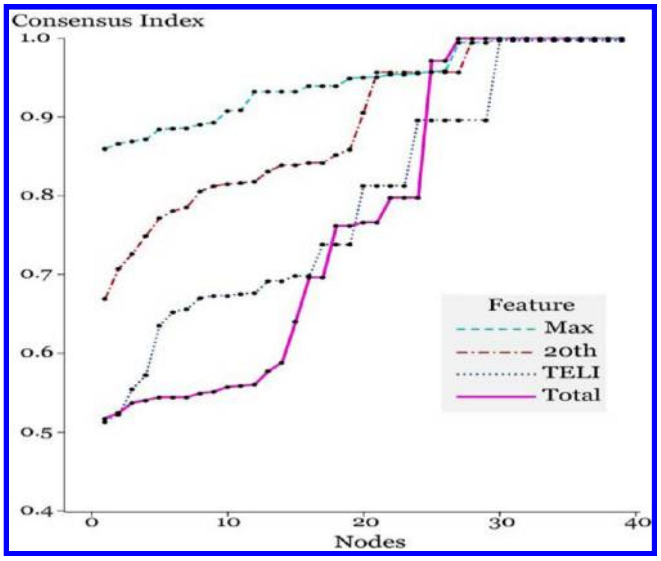


Figure 2. Average consensus index for each node in the dendrogram of consensus clustering. Results are from four different experiments that selected different time-series data subsets, including Max (data set containing the maximum gene expression signal observed for each reporter during the 2-h assay), 20th (data set that uses only gene expression data at the middle time point of the testing, the 20th time point, at 57 min of exposure), and TELI (data set consisting of derived integrated end point-transcriptional effect level index (TELI)). Total represents the full data set with all time points.

Compared to the data set with all time points, the consensus indices for data sets using peak signal or individual time points showed a dramatic deterioration in resolution between treatments (Figures 2, S2, and S3). These results indicate that the temporal differences between signal peaks contribute substantial information beyond the simple peak signal level. This finding underscores the importance of careful selection of sampling times for a given experimental system. Individual time points may not provide sufficient information to fully reflect the complex and dynamic biological response. Using maximum expression as a single feature removes any dependence on timing and synchronization, but may also sacrifice the rich biological information associated with time-series generated by cell-based HTS assays. The results also indicated that temporal data compression using our proposed TELI concept followed by CC distinguished several prominent treatment clusters, and performed qualitatively better than the other two data sets without temporal resolution. As expected, CC with the full data set performed the best among all four experiments. This result does not imply that temporal toxicogenomic data with such a high resolution (every 3 min in our assay) is necessary in all experimental systems. Rather, these experiments demonstrate the sensitivity of clustering to sampling frequency and time points, and the importance of gene response dynamics in the interpretation of omics data.

3.3. Consistency Test for Consensus Clustering. To evaluate the sensitivity of the clustering results toward the variations in experimental design such as the number of treatments, we performed SOM-based CC to analyze the nanomaterials subset of our data (24 out of total 40 treatments). If the subset clusters similarly to the full data set clustering, that would support the hypothesis that CC is resistant to treatment numbers and data size, and that clustering results can be meaningful.

The nanomaterial subset clustering results are shown in Figure 3 (the curve for determining the optimal cluster number is shown in Figure S1b). The results are largely consistent with clustering of the full data set. The low concentration treatments again clustered near the untreated control. Interestingly, the carbon nanotube treatments, SWNT and SWNT_OX, remain distinct from the remaining toxicants and the surface-oxidized nanotubes were distinguishable from their original as discussed previously. Although consisting of the same carbon element, the distinguishable toxic effects among the three carbon-based nanomaterials evaluated (carbon nanotube, fullerene, and carbon black), are suggestive of their nanostructure-dependent toxicological implications.[48] Moreover, these results showed that perturbing the input data set by taking a subset of treatments continued to produce relatively stable clusters, which indicate plausible classes of biological responses.

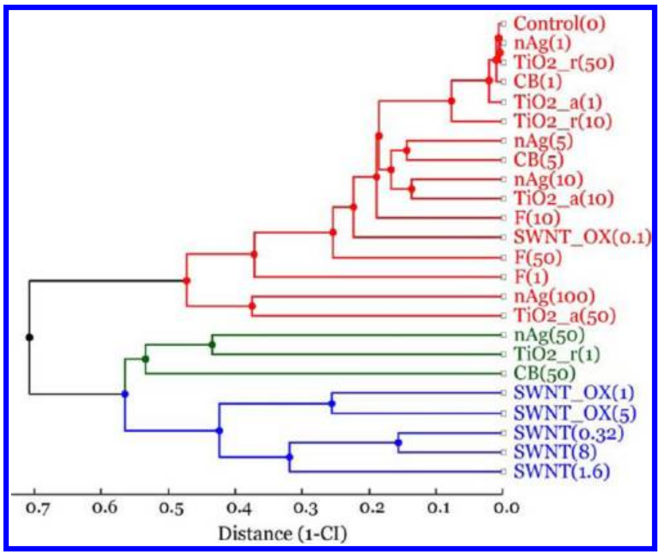


Figure 3. Dendrogram produced from consensus clustering of nanomaterials.

To further test the susceptibility of the clustering results to data noise level, we assessed the consistency of CC by adding computer-generated noise to the original data set using the MATLAB function normrnd. This experiment evaluated the robustness of CC to the variance caused by multiple laboratories performing similar experiments but lacking standardized protocols. Normally distributed random noise was generated with mean (μ) value of zero and a range of standard deviation (σ) values. The results showed that the CC could still produce relatively consistent clustering result without significantly losing resolution power at $\sigma = 0.5$, as shown in Figure 4. Although the differentiating quality was reduced as shown by the somewhat blurred clustering boundaries, the number of clusters and aggregates of chemicals were largely conserved. This result demonstrates that the CC analysis is relatively reliable and therefore may be applicable for clustering analysis with data generated from different laboratories if the noise level is not excessive.

3.4. Impact of Gene Selection: Consensus Clustering Using Reordered or Subset of Total Genes. To assess the sensitivity of the clustering results to the selection of gene reporters, we performed a CC perturbation experiment that performs bootstrap resampling of the 91 gene reporters while

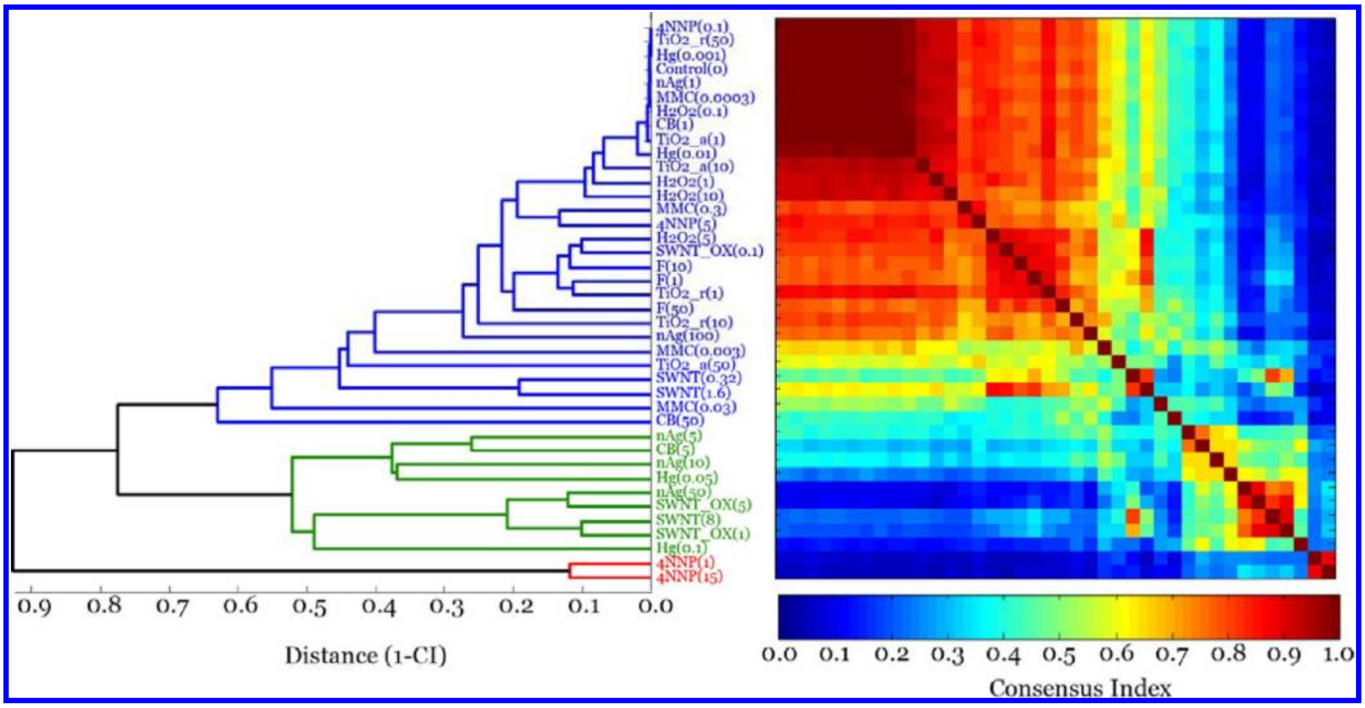


Figure 4. Consensus clustering result for data set perturbed with additional Gaussian noise having mean value (μ) as zero and standard deviation (σ) of 0.5.

maintaining the full set of treatment conditions. The resulting consensus matrix, shown in Figure 5, differed from the previous results in having broadly higher average CI values among nonclustered treatments. Individual clusters were less apparent, blurring the earlier strong distinctions between treatment classes. This loss of clustering resolution suggests that the noise introduced by gene resampling allowed only strong, common responses between treatments to cluster, while obscuring the more subtle distinctions between treatments. Choosing an appropriate and sufficient set of gene reporters remains an open research question.[53] An ideal gene reporter set would clearly discriminate between a wide variety of treatment conditions, convey large changes in signal level, have low variance as well as a low noise floor, and carry little redundancy to minimize experimental time and cost.

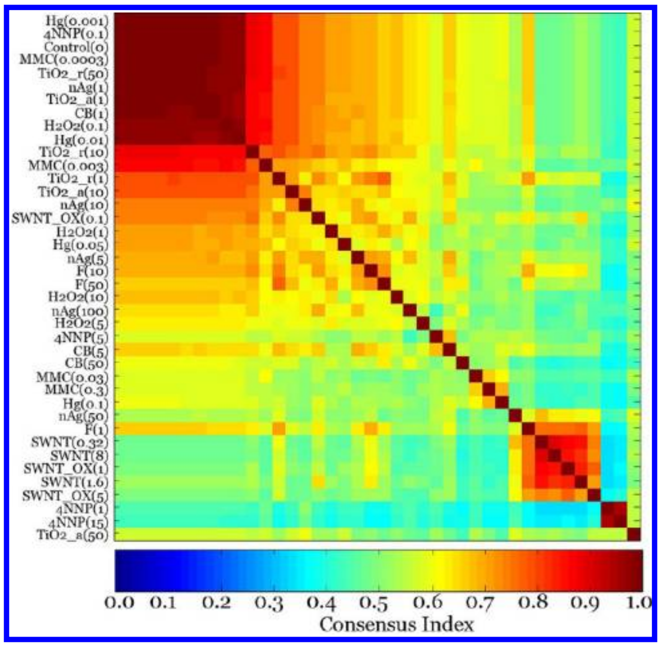


Figure 5. Consensus matrix based on the clustering result with dynamic genes library. The scheme of resampling is changed to reconstruct the treatment vectors using genes randomly selected from the original library. With this perturbed input, the pattern differences between treatments change accordingly, which in turn impact the resolving power of the clustering algorithm.

To further explore sensitivity to gene selection, we performed CC using a subset of reporters involved in redox stress. The result, shown in Figure S4, also shows a large but diffused cluster. The results from these two gene perturbation experiments suggest that the particular set of 91 reporters employed here cannot be greatly reduced while maintaining the resolution needed to distinguish between the treatments tested here. This conclusion, however, does not exclude the possibility that a different and smaller set of reporters might have the resolution necessary to clearly identify separate clusters of treatments. Nor does this result preclude using a larger set of reporters to discriminate among wider selections of toxicants. Of course, optimization in the selection of pathways and genes to gain sufficient resolution power and with minimal redundancy is yet another challenge that is beyond the scope of this study.

CC serves both as a validation algorithm for conventional clustering and as an ensemble clustering approach. This study has shown that CC performs well in reliability for our three-dimensional HTS toxicological data. These results support the hypothesis that, within reasonable limits, CC facilitates cluster comparisons between experiments with differing designs and variations. We also found that for our cellular stress response ensemble-based HTS transcriptomics assay platform, the size and composition of the reporter gene set are critical factors that affect the resulting coherency of clusters. These findings suggested a relatively low level of redundancy within the set of 91 reporters employed here to represent the stress-response pathway ensemble. We also demonstrated the value and importance of utilizing rich time-series data, which produced the highest level of cluster resolution. In summary, this study has demonstrated the value of CC, and illustrated its potential usefulness to analyze high-dimensional toxicogenomic databases.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1 listing the toxicity mechanism and dose concentrations for the chemicals and nanomaterials that were evaluated in this study; Table S2 summarizing the characterization data for all the nanomaterials; Figure S1 illustrating the confirmation of cluster number by finding the peak CDF area change as described in Monti et al; Figure S2 showing CC analysis with data set built with only redox genes. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: april@coe.neu.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Special thanks are given to the CDM/Diane and William Howard scholarship, Industrial Translational Research Initiative (ITRI) Scholarship. This study was also supported by National Science Foundation (NSF) (EEC-0926284 and CAREER CBET- 0953633) and National Science Foundation Nanoscale Science and Engineering Center (NSEC) for Highrate Nanomanufacturing (0425826). We thank Prof. Dhimiter Bello and Anoop K. Pal from University of Massachusetts at Lowell for providing the nanomaterials characterization.

■ REFERENCES

(1) Krewski, D.; Acosta, J. D.; Andersen, M.; Anderson, H.; Bailar, I. I. I. J. C.; Boekelheide, K.; Brent, R.; Charnley, G.; Cheung, V. G.; Green, J. S.; Kelsey, K. T.; Kerkvliet, N. I.; Li, A. A.; McCray, L.; Meyer, O.; Patterson, R. D.; Pennie, W.; Scala, R. A.; Solomon, G. M.; Stephens, M. Toxicity testing in the 21st century: A vision and a strategy. *J. Toxicol. Environ. Health: Part B* **2010**, *13* (2−4), 51−138.

(2) Newton, R. K.; Aardema, M.; Aubrecht, J. The utility of DNA microarrays for characterizing genotoxicity. *Environ. Health Perspect.* **2004**, *112* (4), 420−2.

(3) Elad, T.; Lee, J. H.; Belkin, S.; Gu, M. B. Microbial whole-cell arrays. *Microb. Biotechnol.* **2008**, *1* (2), 137−48.

(4) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* **2007**, *95* (1), 5−12.

(5) Simmons, S. O.; Fan, C.-Y.; Ramabhadran, R. Cellular stress response pathway system as a sentinel ensemble in toxicological screening. *Toxicol. Sci.* **2009**, *111* (2), 202−225.

(6) Van Dyk, T. K.; Wei, Y.; Hanafey, M. K.; Dolan, M.; 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.* **2001**, *98* (5), 2555−2560.

(7) Sedykh, A.; Zhu, H.; Tang, H.; Zhang, L.; Richard, A.; Rusyn, I.; Tropsha, A. Use of *in vitro* HTS-derived concentration−response data as biological descriptors improves the accuracy of QSAR models of *in vivo* toxicity. *Environ. Health Perspect.* **2010**, *119*, 3.

(8) Huang, R.; Southall, N.; Cho, M.-H.; Xia, M.; Inglese, J.; Austin, C. P. Characterization of diversity in toxicity mechanism using *in vitro* cytotoxicity assays in quantitative high throughput screening. *Chem. Res. Toxicol.* **2008**, *21* (3), 659−667.

(9) Rallo, R.; France, B.; Liu, R.; Nair, S.; George, S.; Damoiseaux, R.; Giralt, F.; Nel, A.; Bradley, K.; Cohen, Y. Self-organizing map analysis of toxicity-related cell signaling pathways for metal and metal oxide nanoparticles. *Environ. Sci. Technol.* **2011**, *45* (4), 1695−1702.

(10) George, S.; Xia, T.; Rallo, R.; Zhao, Y.; Ji, Z.; Lin, S.; Wang, X.; Zhang, H.; France, B.; Schoenfeld, D.; Damoiseaux, R.; Liu, R.; Lin, S.; Bradley, K. A.; Cohen, Y.; Nel, A. E. Use of a high-throughput screening approach coupled with in vivo zebrafish embryo screening to develop hazard ranking for engineered nanomaterials. *ACS Nano* **2011**, *5* (3), 1805−1817.

(11) 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.

(12) 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.

(13) Gu, M. B.; Mitchell, R. J.; Kim, B. C. Whole-cell-based biosensors for environmental biomonitoring and application. *Adv. Biochem. Eng./Biotechnol.* **2004**, *87*, 269−305.

(14) Lee, J. H.; Youn, C. H.; Kim, B. C.; Gu, M. B. An oxidative stress-specific bacterial cell array chip for toxicity analysis. *Biosens. Bioelectron.* **2007**, *22* (9−10), 2223−2229.

(15) Cheng Vollmer, A.; Van Dyk, T. K., Stress responsive bacteria: Biosensors as environmental monitors. In *Advances in Microbial Physiology*; Poole, Ed.; Academic Press, 2004; Vol. *49*, pp 131−174.

(16) Jiang, D.; Pei, J.; Ramanathan, M.; Tang, C.; Zhang, A. Mining coherent gene clusters from gene-sample-time microarray data. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: Seattle, WA, 2004; pp 430−439.

(17) Ioannidis, J. P. A.; Khoury, M. J. Improving validation practices in "omics" research. *Science* **2011**, *334* (6060), 1230−1232.

(18) Harper, G.; Pickett, S. D. Methods for mining HTS data. *Drug Discovery Today* **2006**, *11* (15−16), 694−699.

(19) Afshari, C. A.; Hamadeh, H. K.; Bushel, P. R. The evolution of bioinformatics in toxicology: Advancing toxicogenomics. *Toxicol. Sci.* **2011**, *120* (suppl 1), S225−S237.

(20) Daxin, J.; Chun, T.; Aidong, Z. Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowledge Data Eng.* **2004**, *16* (11), 1370−1386.

(21) D'Haeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* **2005**, *23* (12), 1499−1501.

(22) Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*; Prentice Hall: Englewood Cliffs, NJ, 1988.

(23) Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E. S.; Golub, T. R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* **1999**, *96* (6), 2907−2912.

(24) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci.* **2006**, *103* (31), 11473−11478.

(25) Bar-Joseph, Z.; Demaine, E. D.; Gifford, D. K.; Srebro, N.; Hamel, A. M.; Jaakkola, T. S. K-ary clustering with optimal leaf ordering for gene expression data. *Bioinformatics* **2003**, *19* (9), 1070−1078.

(26) Törönen, P.; Kolehmainen, M.; Wong, G.; Castrén, E. Analysis of gene expression data using self-organizing maps. *FEBS Lett.* **1999**, *451* (2), 142−146.

(27) Mangiameli, P.; Chen, S. K.; West, D. A comparison of SOM neural network and hierarchical clustering methods. *Eur. J. Operat. Res.* **1996**, *93* (2), 402−417.

(28) Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On clustering validation techniques. *J. Intell. Inform. Syst.* **2001**, *17* (2), 107−145.

(29) Gibbons, F. D.; Roth, F. P. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* **2002**, *12* (10), 1574−1581.

(30) Datta, S.; Datta, S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinform.* **2006**, *7* (1), 397.

(31) Datta, S.; Datta, S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **2003**, *19* (4), 459−466.

(32) Monti, S.; Tamayo, P.; Mesirov, J.; Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **2003**, *52* (1), 91−118.

(33) Nguyen, T. T.; Nowakowski, R. S.; Androulakis, I. P. Unsupervised selection of highly coexpressed and noncoexpressed genes using a consensus clustering approach. *OMICS* **2009**, *13* (3), 219−237.

(34) Seiler, M.; Huang, C. C.; Szalma, S.; Bhanot, G. ConsensusCluster: A software tool for unsupervised cluster discovery in numerical data. *OMICS* **2010**, *14* (1), 109−113.

(35) Wilkerson, M. D.; Hayes, D. N. ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* **2010**, *26* (12), 1572−3.

(36) 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.

(37) 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.

(38) Kim, S. Y.; Lee, J. W. Ensemble clustering method based on the resampling similarity measure for gene expression data. *Stat. Methods Med. Res.* **2007**, *16* (6), 539−564.

(39) Vesanto, J. H.; Alhoniemi, E.; ParhankangasJ. Self-organizing map in MATLAB: the SOM Toolbox. In *Proceedings of the MATLAB DSP Conference 1999*, Espoo, Finland, November 16−17, 1999; 1999; pp 35−40.

(40) Ultsch, A.; Siemon, H. P. Kohonen's self organizing feature maps for exploratory data analysis. In *Proceedings of International Neural Networks Conference, 1990*; Kluwer Academic Press, 1990; pp 305−308.

(41) Khil, P. P.; Camerini-Otero, R. D. Over 1000 genes are involved in the DNA damage response of *Escherichia coli*. *Mol. Microbiol.* **2002**, *44* (1), 89−105.

(42) Cantoni, O.; Brandi, G.; Salvaggio, L.; Cattabeni, F. Molecular mechanisms of hydrogen peroxide cytotoxicity. *Ann. Ist. Super. Sanita* **1989**, *25* (1), 69−73.

(43) Van Dyk, T. K.; Smulski, D. R.; Reed, T. R.; Belkin, S.; Vollmer, A. C.; LaRossa, R. A. Responses to toxicants of an *Escherichia coli* strain carrying a uspA'::lux genetic fusion and an *E. coli* strain carrying a grpE'::lux fusion are similar. *Appl. Environ. Microbiol.* **1995**, *61* (11), 4124−7.

(44) Soares, A.; Guieysse, B.; Jefferson, B.; Cartmell, E.; Lester, J. N. Nonylphenol in the environment: A critical review on occurrence, fate, toxicity and treatment in wastewaters. *Environ. Int.* **2008**, *34* (7), 1033−49.

(45) Reddy, A. R. N.; Reddy, Y. N.; Himabindu, V.; Krishna, D. R. Induction of oxidative stress and cytotoxicity by carbon nanomaterials is dependent on physical properties. *Toxicol. Ind. Health* **2011**, *27* (1), 3−10.

(46) Bello, D.; Hsieh, S.-F.; Schmidt, D.; Rogers, E. Nanomaterials properties vs. biological oxidative damage: Implications for toxicity screening and exposure assessment. *Nanotoxicology* **2009**, *3* (3), 249−261.

(47) de Bodt, E.; Cottrell, M.; Verleysen, M. Statistical tools to assess the reliability of self-organizing maps. *Neural Networks* **2002**, *15* (8−9), 967−978.

(48) Vecitis, C. D.; Zodrow, K. R.; Kang, S.; Elimelech, M. Electronic-structure-dependent bacterial cytotoxicity of single-walled carbon nanotubes. *ACS Nano* **2010**, *4* (9), 5471−9.

(49) Liu, S.; Ng, A. K.; Xu, R.; Wei, J.; Tan, C. M.; Yang, Y.; Chen, Y. Antibacterial action of dispersed single-walled carbon nanotubes on *Escherichia coli* and *Bacillus subtilis* investigated by atomic force microscopy. *Nanoscale* **2010**, *2* (12), 2744−50.

(50) Yang, C.; Mamouni, J.; Tang, Y.; Yang, L. Antimicrobial activity of single-walled carbon nanotubes: Length effect. *Langmuir* **2010**, *26* (20), 16013−9.

(51) Jin, C.; Tang, Y.; Yang, F. G.; Li, X. L.; Xu, S.; Fan, X. Y.; Huang, Y. Y.; Yang, Y. J. Cellular toxicity of TiO_2 nanoparticles in anatase and rutile crystal phase. *Biol. Trace Elem. Res.* **2011**, *141* (1−3), 3−15.

(52) Ahn, J.-M.; Hwang, E. T.; Youn, C.-H.; Banu, D. L.; Kim, B. C.; Niazi, J. H.; Gu, M. B. Prediction and classification of the modes of genotoxic actions using bacterial biosensors specific for DNA damages. *Biosens. Bioelectron.* **2009**, *25* (4), 767−772.

(53) Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23* (19), 2507−2517.