Pattern-Based Gene-Set Recognition for Interpreting Genome-Wide Gene Expression Profiles

Xutao Deng*[,1,2] and Charles Wang*[,1,2]

[1]*Transcriptional Genomics Core, Burns Allen Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA*

[2]*Department of Medicine, UCLA David Geffen School of Medicine, Los Angeles, CA, USA*

Abstract: Background: Accurate recognition of important gene sets from genome-wide gene expression profiles provides great insights into the underlying biological mechanisms that govern the gene expression dynamics. However, most gene set recognition algorithms rely solely on supervised sample phenotypic information, overlooking the unsupervised gene-gene expression correlations that are inherently informative in the gene expression profiles.

Results: We developed a computational framework named PAGER (Pattern Acquisition and GEne-set Recognition) for identifying gene sets showing significant supervised and unsupervised patterns. We showcased the use of PAGER in several recent expression profiling studies including cadmium treated rat primary hepatocyte toxicogenomics study and adrenal gland periodical gene expression profiling. Our results indicate that PAGER achieved better performance in discovering truly important pathways from expression profiles which were undetected using current other existing tools. These results were further corroborated by literature and cytotoxicity experiments.

Conclusions: PAGER integrated both supervised and unsupervised pattern metrics for gene set summarization. For each given gene set, PAGER provides a two-dimensional view showing its external activity and internal coherence pattern. PAGER employed statistical methods such as Relaxed Intersection-Union Tests, Stouffer's method and Fisher's method for integration of pattern significance. In addition, PAGER can be used for recognizing user-defined arbitrary gene set as demonstrated in one of our previous publications. PAGER is freely available for academic user at http://dengx.bol.ucla.edu/PAGER/PAGER.htm.

BACKGROUND

Gene set recognition, a major apparatus for interpreting gene expression profiles, has been employed in a wide range of microarray-based studies ranging from *Drosophila* heart formation to human cancer mechanisms [1-9]. Gene set recognition refers to the computational task of identifying previously curated gene sets such as regulatory pathways and gene ontology. The appealing feature of gene set recognition comes from its capability to identify signal genes as functional modules that give rise to more biologically interpretable results rather than sporadic significant genes. The basic and standard approach to gene set recognition is to use Fisher's Exact Test (FET) to compare the microarray-derived gene list against each of the pathways, gene ontology groups in standard database such as GenMAPP [10;11] and Gene Ontology [12]. This simple FET approach has been implemented in many data analysis tools [13-17]. In fact, virtually all gene set recognition algorithms involves in three basic steps: (1) each gene is assigned a score according to certain hypothesis; (2) for a specific gene set, a summarization statistic, called the enrichment score is computed from individual gene scores; and (3) a significance value is computed for each gene set enrichment score. For instance, the FET approach requires the users to retrieve the candidate gene lists from the gene expression profiles, which is essentially equivalent to assigning a binary score for each gene indicating whether the gene is of interest or not. Then for each predefined gene set the enrichment score is defined as the proportion interesting genes. Finally, this enrichment score is tested against background of all genes using FET. An obvious drawback of FET is that the dependence on gene lists requires an arbitrary cutoff value for binary scoring, which often leads to inconsistent biological interpretations [18-20]. Subramanian, Tamayo, and Mootha [18] proposed an algorithm named Gene Set Enrichment Analysis (GSEA) which used continuous scoring used for individual gene and gene sets. GSEA uses Kolmogorov–Smirnov statistic as an enrichment score for each pathway, and then significance acquisition and multiplicity adjustment procedures are performed based on the enrichment scores. Subsequently, an array of modifications, primarily targeting on different summarizing statistic and significance acquisition techniques, have been proposed [21-27].

One common feature of the above mentioned methods is their heavy dependence on sample phenotypic information such as disease status, drug treatment, age, gender, which makes them a group of supervised gene set recognition methods. This feature, however, limited their usage in gene expression profiling studies for the following three reasons: (1) the manually curated gene sets in knowledgebase represent established functional modules, and their inherent orchestrated expression patterns are a major signature of the gene sets but were overlooked by the supervised methods; (2) as the gene expression profiling studies become increas-

*Address correspondence to these authors at the 8700 Beverly Blvd., Davis Bldg., G150, Los Angeles, CA90048, USA; Tel: 1-310-423-7361; Fax: 1-310-423-2303; E-mail: charles.wang@cshs.org

8700 Beverly Blvd., Davis Bldg., G151, Los Angeles, CA90048, USA; Tel: 1-310-423-7363; Fax: 1-310-423-2303; E-mail: dengx@ucla.edu

1875-0362/08 2008 Bentham Science Publishers Ltd.

ingly complex, often involved with multiple phenotypes and unknown subgroups, the supervised scoring metric based on gene-phenotype correlations are becoming more difficult to define; and (3) some gene expression profiling studies employ unsupervised design by focusing on discovering co-regulated gene expression pattern [28-30]. Because of these drawbacks, the supervised gene set recognition methods are only effective in simple studies such as a comparison between a treatment group and control group.

The challenge hence is to define and capture the true pattern of the gene sets given the expression profiles. We distinguish the two aspects of the gene set pattern: the supervised gene set *activity* defined by the gene-phenotype correlation structure and the unsupervised gene set *coherence* defined by the gene-gene correlation structure. The coherence of each gene set not only served as a complement to the gene set activity, in some cases, it must be incorporated into gene set activity to avoid reduced recognition sensitivity and specificity [31]. For instance, a highly coherent pathway with low activity may not be important at all if the pathway genes that are constantly expressed across all samples. On the other hand, a highly active pathway with low coherence profiles may indicate its genes were activated through other pathways.

In this paper, we developed a computational framework named PAGER (Pattern Acquisition and GEne-set Recognition) which provides a two-dimensional view on gene sets defined by supervised activity and unsupervised coherence. PAGER allows one to choose activity or coherence or both for gene set recognition. For activity and coherence metrics, PAGER offers a set of expandable options based on specific study design. The major feature of PAGER hinges on its emphasis on its integration of both gene set coherence and activity for gene set recognition. Currently, the options on gene set coherence are generally not available in other software, but nevertheless important for studies employing complex design. Additionally, PAGER emphasizes a statistically justified integration method we developed before [32], in which the error is rigorously controlled and statistical power is improved over classical methods. As demonstrated in one of our published studies, PAGER was proved to be a powerful tool in identifying a set of known stem cell regulation genes in responding to the liver injury, providing evidence supporting our hypothesis that some factors released from the injured rat liver can promote bone marrow hepatic stem cells priming [33].

METHODS

The computational framework of PAGER consists of five steps: scoring, summarization, significance acquisition, integration, and adjustment. As described, although the first three steps were similar to many existing algorithms, PAGER computes both gene set activity and coherence with new summarization options. Then for each gene set, its significance on coherence and activity were combined based on statistically principled methods to identify the gene sets that are active and/or coherent. In the scoring step, the pair-wise gene-gene coherence (GGC) and gene-phenotype activity (GPA) for each gene were computed according to user selected metrics. The gene level scoring was subsequently summarized into gene set level enrichment score which consists of coherence $C(P)$ and activity $A(P)$ for each gene set P. The summarization is followed by the significance acquisition in which the p-values of $C(P)$ and $A(P)$ were obtained based on the comparison with randomized gene sets which are used as the background null distribution. For each gene set P, its p-values of coherence and activity were then com-

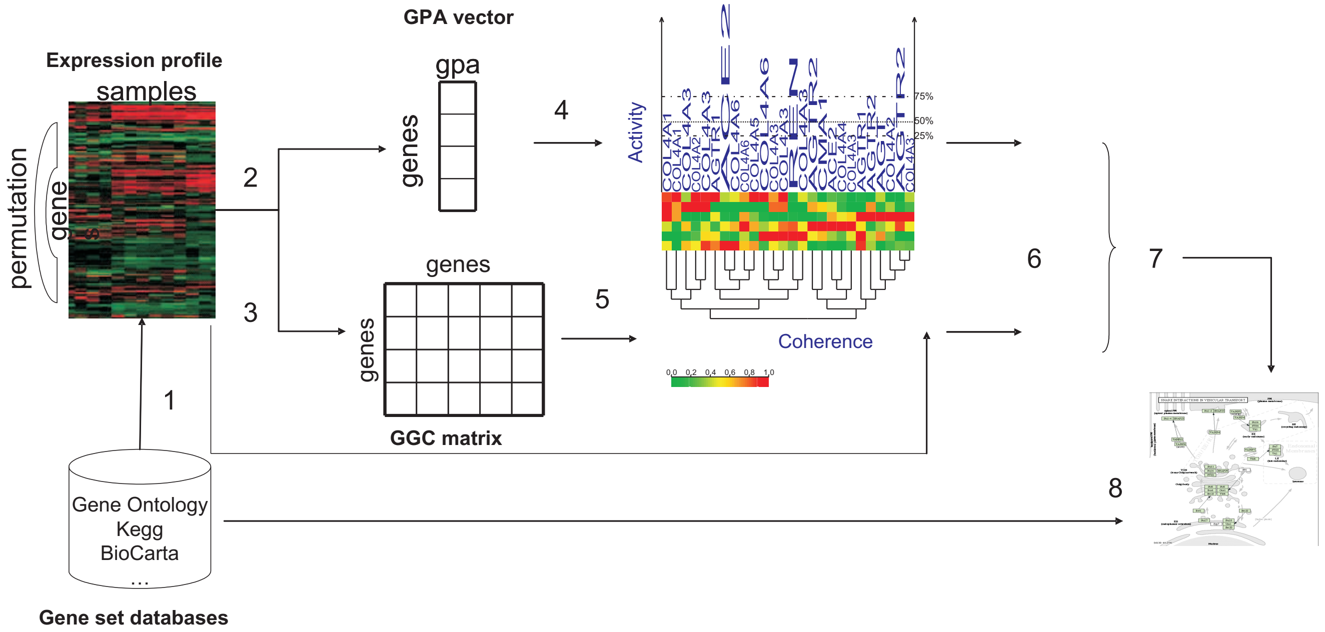


Fig. (1). Illustration of PAGER framework for gene set recognition.

1. Each gene set is superimposed onto the gene expression data 2. The GPA is computed for each gene in the gene set 3. The GGC is computed for each pair of genes 4. GPA vector is summarized into the gene set activity 5. GGC matrix is summarized into gene set coherence 6. The permuted version of the gene set is generated for significance acquisition 7. The significance of gene set activity and coherence are integrated 8. FDR is computed after all gene sets were processed for multiplicity adjustment.

bined into a single integration p-value. The p-value integration is a necessary step to determine if a gene set is significantly active and/or coherent. Then the false discovery rate (FDR) *q*-value was computed to adjust errors for testing multiple gene sets based on the p-value distributions of a given data base with multiple gene sets. The computational framework of PAGER is graphically illustrated in Fig. (**1**).

1. Scoring

PAGER can handle samples belonging to either discrete phenotypes (such as gender) or continuous phenotypes (such as age). For each gene set *P*, the GGC scoring matrix **M** and GPA scoring vector **z** were established from its individual gene expression profile represented as g_i and phenotype profile *phe*:

$\mathbf{M}_{ij} = GGC(g_i, g_j), \quad g_i, g_j \in P$, and

$$\mathbf{z}_i = GPA(g_i, phe), \quad g_i \in P \tag{1}$$

The scoring of **M** and **z** is based on the choice of coherence function $GGC(\cdot)$ and the activity function $GPA(\cdot)$. For example, the Pearson's correlation, Spearman's rank correlation, Hamming distance (after discretization), and the inverse of Euclidean distance are among the choices of the coherence function. The choice of $GPA(\cdot)$ depends highly on specific experimental design and research hypothesis. For instance, *t*-statistic or mean fold changes were suitable function for two-group comparison. The ANOVA *f*-statistic or coefficient of variance (CV) can be used for multi-group comparison. For continuous phenotypes, the GPA can be measured by Pearson's correlation or other statistic that represents the research hypotheses. For measuring gene level activity, PAGER provides the options of parametric t-statistic (without assumption of equal variance) or fold change for two-group comparison and ANOVA f-statistic or coefficient of variance for multi-group comparison. For continuous phenotypes, PAGER currently provides Pearson's correlation or Spearman's correlation.

2. Summarization

In this step, gene set level statistic for measuring coherence *C*(*P*) and Activity *A*(*P*) for each gene set *P* was computed based on **M** and **z** respectively:

$$C(P) = \Psi(\mathbf{M}), \text{ and } A(P) = \Phi(\mathbf{z}) \tag{2}$$

where $\Psi(\bullet)$ and $\Phi(\bullet)$ are summarization functions on gene level scores **M** and **z** respectively. In PAGER, the maximum spanning tree was proposed as a better metric for gene set coherence *C*(*P*). This was done by modeling each gene set *P* as a graph with each of its genes a graph vertex and each entry M_{ij} as the weight of the edge between gene pair g_i and g_j. Therefore the GGC matrix **M** was transformed into a graph adjacency matrix. The $\Psi(\mathbf{M})$ was defined as the total weights of the maximum spanning tree edges which can be computed using the well-known Prim's algorithm or Kruskal's algorithm [34]. The maximum spanning tree summarization could better capture the overall GGP structure than median or median of the elements in **M** as demonstrated in our results. It was shown that the majority of the activated gene sets has more than one conserved patterns in the expression profiles (see Results for detail). To summarize the gene set activity *A*(*P*), PAGER adopted the GSA max_mean statistic [23], GSEA Kolmogorov–Smirnov statistic [18], or mean(**z**) as options for the summarizing function $\Phi(\bullet)$.

3. Significance

The significance of gene set activity and coherence were obtained by gene randomization. In GSEA, the significance was obtained using label permutation [18]. However label permutation incurs a big cost as it requires simple study design, but large sample size to obtain a sufficient number of permutations that can lead to any significant gene sets. PAGER circumvented the sample size problem by adopting gene randomization. For each gene set *P* of size *m*, we randomly select *m* genes from the set of all genes without replacement to construct a random gene set. We draw *B* such random gene sets notated as P^1, P^2,…, P^B and compute $C(P^i)$ and $A(P^i)$ for *i*=1,…,B. It is easy to obtain p-values using the random gene sets as the background null distribution:

$$p_C = \frac{\#\{C(P^i) \geq C(P)\}}{B}, \text{ and}$$

$$p_A = \frac{\#\{A(P^i) \geq A(P)\}}{B}. \tag{3}$$

The p-values of gene set coherence and activity were notated as p_C and p_A respectively.

4. Integration

Based on the p-values p_A and p_C for each considered gene set, the integration is essentially a multiplicity error adjustment procedure. In PAGER, the integration of statistical significance can be performed in two ways: the meta-analysis approach and the Intersection-Union Test (IUT) approach. There is a subtle but important difference between two integration approaches. The integrated significant p-value using the meta-analysis approach provides the *overall* significance of activity and coherence of a given gene set; whereas the integrated significant p-value using the IUT approach indicates the gene set's activity and coherence are *both* significant. The meta-analysis methods include the well known Fisher's method [35;36] and Stouffer's method [37]. Fisher's method for computing combined p-value p_{Fisher} is computed as: $X_4^2 = -2(\log_e p_C + \log_e p_A)$ and $p_{Fisher} = p_C p_A - (p_C p_A \ln p_C p_A)$. Stouffer's method can be expressed as $Z_{Stouffer} = (Z_C + Z_A)/\sqrt{2}$, where $Z_\bullet = SD\_Normal\_Inv(p_\bullet)$, the Z-score corresponding to the tail probability of standard normal distribution.

The selection of integration method is dependent on specific research hypothesis and study design. The meta-analysis is more sensitive thus can give rise to more significant gene sets, but it is more susceptible to make false positives than the IUT approach. Alternatively, the IUT approach provides a more rigorous and conservative way against excessive false positives than meta-analysis. IUT combines activity and coherence tests into a single test which rejects the combined null hypothesis if both of the individual tests reject. In PAGER, we adopted our previously developed method for performing rigorous and powerful IUT [32], which has an adjusted IUT p-value *p'* expressed as

$$p' = \frac{(1-\pi_A)\pi_C p + (1-\pi_A)\pi_C p + (1-\pi_C)(1-\pi_A) p^2}{1-\pi_C \pi_A}, \qquad (4)$$

where $p = \max(p_C, p_A)$ is the unadjusted IUT p-value [38] and π_C and π_A are the true probabilities of alternative hypotheses. The nuisance parameters (π_C, π_A) can be estimated from all gene sets. Since there are usually hundreds of genes available for each statistical test, we can obtain crude and conservative estimates of the parameters according to the following equations [39;40]:

$$\hat{\pi}_C(\lambda) = \frac{\#\{p_C(j) < \lambda\}}{(1-\lambda)n}, \hat{\pi}_A(\lambda) = \frac{\#\{p_A(j) < \lambda\}}{(1-\lambda)n},$$

$$i = 1, 2, j = 1, \ldots, n \qquad (5)$$

where *n* is the total number of curated gene sets, λ is a chosen fixed value and $p_C(j)$ and $p_A(j)$ represent the p-values for the coherence and activity on the *j*th gene set. It was shown that this approach is more powerful than or as powerful as the unadjusted IUT approach with feature of rigorously control type 1 error at the nominal significance level [32]. Steps 1-4 were repeated to obtain the p_C and p_A for each gene set in the database.

5. Adjustment

Because hundreds of gene sets are being tested, it is crucial to perform multiplicity adjustment to avoid excessive type 1 error. In our case studies, the p-value computation was followed by false discovery rate (FDR) *q*-value adjustment using the direct FDR approach [39;40]. The FDR procedure was performed on the distributions of p_A, p_C, and p_I respectively.

RESULTS

Gene Set Database and Software

In our case studies, we use the MSigDB Version 2 Collection 2 [18]—a collection of curated gene sets from multiple sources—as the main gene set database for testing. To ensure a high quality collection of gene sets, the source database were limited to KEGG [41;42], GenMAPP [11], and Broad Institute, resulting in 394 candidate gene sets in our test database.

Due to the intensity of computation required, PAGER was released as a stand alone desktop application. PAGER was implemented in C++ and has an optional graphic user interface implemented in C#. PAGER is compatible with any public or user-edited gene sets in MSigDB gene set format

Table 1. Significant Gene Sets Identified by PAGER in the Time-Course Gene Expression Profile of *Rhesus Macaque* Adrenal Gland

Gene Set	p_C	p_A	p_I	q_C	q_A	q_I
GPCRDB_CLASS_A_RHODOPSIN_LIKE	0.001	0.001	0.001	0.026	0.041	0.006
GAMMA_HEXACHLOROCYCLOHEXANE _DEGRADATION	0.001	0.001	0.001	0.026	0.041	0.006
ACETAMINOPHENPATHWAY	0.014	0.001	0.001	0.135	0.041	0.006
ASBCELLPATHWAY	0.006	0.001	0.001	0.100	0.041	0.006
TRYPTOPHAN_METABOLISM	0.002	0.001	0.001	0.042	0.041	0.006
PEPTIDE_GPCRS	0.105	0.001	0.001	0.383	0.041	0.034
STATIN_PATHWAY_PHARMGKB	0.093	0.002	0.001	0.373	0.075	0.039
FIBRINOLYSISPATHWAY	0.035	0.007	0.001	0.222	0.188	0.037
TCAPOPTOSISPATHWAY	0.002	0.047	0.001	0.042	0.503	0.034
LYMPHOCYTEPATHWAY	0.046	0.003	0.001	0.270	0.106	0.034
GPCRDB_OTHER	0.015	0.010	0.001	0.141	0.238	0.034
EXTRINSICPATHWAY	0.098	0.001	0.001	0.377	0.041	0.034
TH1TH2PATHWAY	0.083	0.001	0.001	0.356	0.041	0.034
INFLAMPATHWAY	0.032	0.017	0.002	0.216	0.297	0.057
BIOGENIC_AMINE_SYNTHESIS	0.163	0.001	0.002	0.438	0.041	0.048
NOS1PATHWAY	0.006	0.096	0.003	0.100	0.710	0.074
BLOOD_CLOTTING_CASCADE	0.145	0.009	0.008	0.416	0.226	0.154
ERBB4PATHWAY	0.001	0.393	0.009	0.026	1.000	0.154
IL17PATHWAY	0.207	0.006	0.009	0.480	0.174	0.154
ARGININECPATHWAY	0.009	0.170	0.009	0.118	0.907	0.154
CHOLESTEROL_BIOSYNTHESIS	0.001	0.391	0.009	0.026	1.000	0.154
TYROSINE_METABOLISM	0.235	0.005	0.010	0.500	0.157	0.154

[18] and it is capable of handling gene expression samples belonging to either discrete phenotypes or continuous phenotypes. PAGER also supports standard gene-by-sample spreadsheet format or the structured SOFT format. Along with a graphic view for each gene set, the outputs of PAGER include a sortable table of p-values and q-values on gene set activity, coherence and integration. The released resources include the PAGER software and source code, the initial gene set data base, and the accompanying documentation.

Case Study 1: Time Course Gene Expression Profiling in the *Rhesus Macaque* Adrenal Gland

The data set involves expression profiles of whole adrenal glands collected at 4-hour intervals across a 24-hour period at 7 am, 11 am, 3 pm, 7 pm, 11 pm, and 3 am of 22,283 transcripts from adult female *Rhesus macaques*. The normalized data can be downloaded from the Gene Expression Omnibus (GEO) using record number GDS2110 [43]. We use Coefficient of Variance across six time points as a measure of GPA for each transcript. Pearson's correlation coefficients were used to measure GGC. The gene set summarizing statistic for gene set activity is GSA max-mean statistic, and for coherence is MST. Stouffer's method was used for integrating gene set activity and coherence. With the number of randomization set at 1000, we identified 22 significant gene sets with integrated p-value less than 0.01 and *q*-value less than 0.2 (Table **1**). Extremely small (< 5) or large (> 200) gene sets were eliminated, resulting in 382 candidate gene sets with size 5-200.

PAGER allow us to identify G-protein coupled receptors (GPCRS) related gene sets (GPCRDB_Class_A_Rhodopsin_Like Peptide,_GPCRS, GPCRDB_Other), blood-clotting related gene sets (Blood_Clotting_Cascade, Fibrinolysis_Pathway) and immune system gene sets

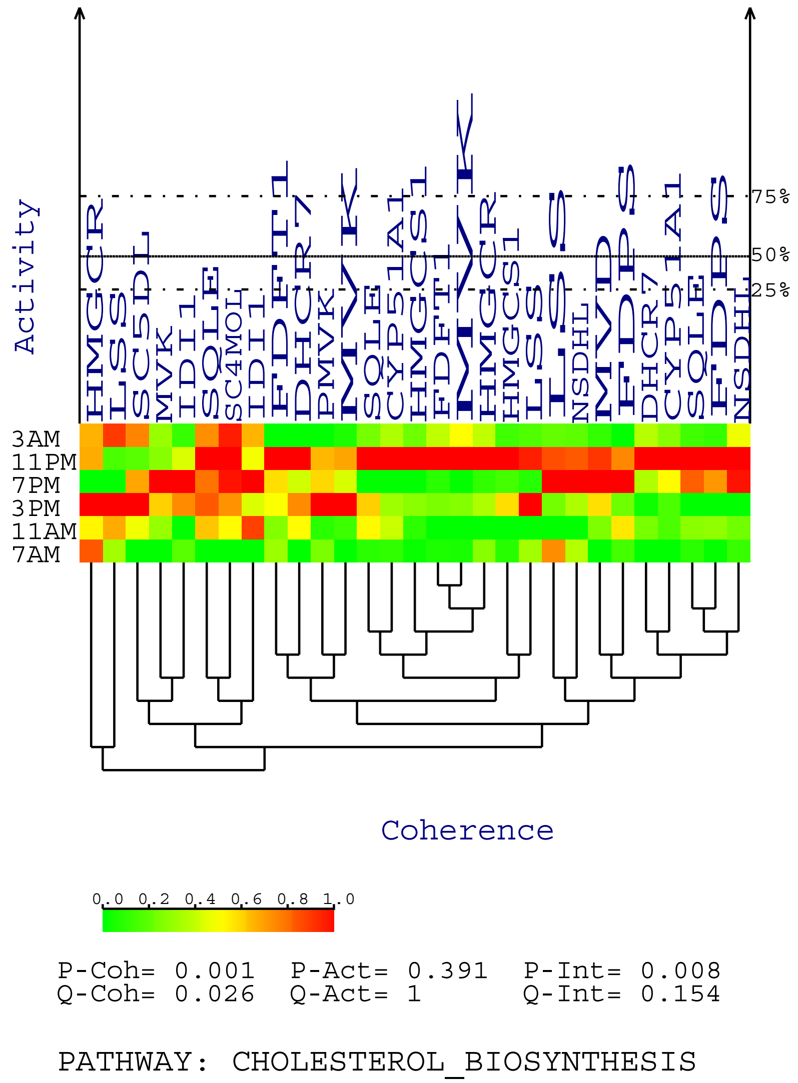


Fig. (2). Unsupervised coherence tree structure and supervised activity scores of gene set Cholesterol_Biosynthesis*.

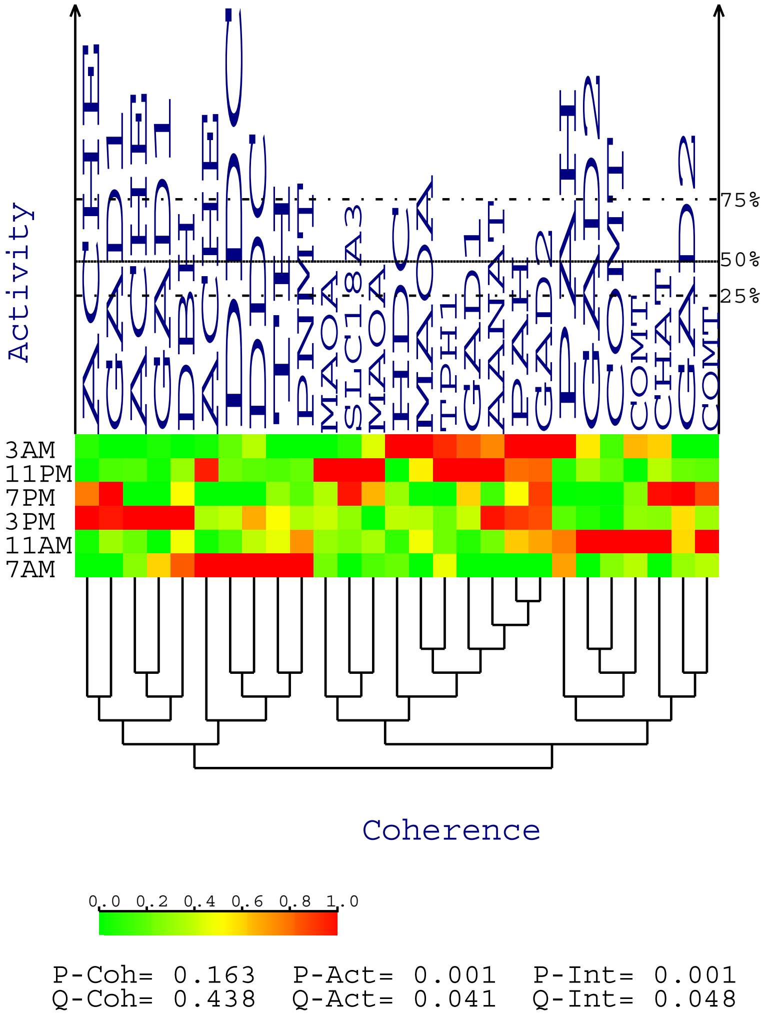
*Based on the data set GDS2110 from GEO. The Cholesterol_Biosynthesis gene set is not significant in its activity (p_A=0.391) but significant in its coherence (p_C=0.001) with peak expression at 7 PM – 11 PM for the majority of its genes. The height of each gene symbol was displayed proportionally to its activity score so that highly activated genes will stand out from the background. The 25%, 50%, and 75% quantile lines across all gene activity scores were also displayed. Gene-gene correlations were displayed in a clustered fashion and were used for the summarization of gene set coherence. The p-values and adjusted q-values for the gene set coherence, activity and integration were also displayed.

(Asbcell_Pathway, Th1Th2_Pathway) which were not reported in the original study [44]. However, their highly significant integrated p-values plus the recent findings [45;46] suggest the significance on their time-specific regulation in adrenal gland.

Our analysis also confirmed the findings in the original report [44] that Cholesterol_ Biosynthesis was periodically regulated on a daily basis. This gene set displayed a highly significant coherence (p_C = 0.001) with peak expression at 7 pm and 11 pm but its activity was not significant (p_A= 0.391). Fig. (**2**) shows the graphic output of the two-dimensional activity-coherence structure of the Cholesterol_Biosynthesis gene set. In contrast, Biogenic_Amine_Synthesis gene set displayed high activity (p_A=0.001) but nevertheless showed non-coherent (p_C=0.163) expression pattern (Fig. **3**). These cases clearly demonstrated the advantage of investigating gene set pathway pattern by looking into both coherence and activity.

For the gene set Fibrinolysis_pathway, there is a large difference in significance between MST summarization ($p_{C\text{-MST}}$ = 0.035) and mean summarization ($p_{C\text{-Mean}}$= 0.261) in summarizing the gene set coherence. This demonstrated that the selection of different summarizing statistic could lead to different test conclusions. As shown in Fig. (**4**), genes displayed their expression in different rate in this given gene set. The genes formed three clusters with their expressions peak at early, middle, or late stage of a 24 hour period. The use of MST could better capture the non-random structures of gene-gene correlation.

These examples demonstrated that using activity or coherence measure alone would miss certain patterns that are important in adrenal gland gene regulation [44]. To further elucidate this, we showed the joint distribution of p_C and p_A among the 464 gene sets. The very weak correlation between the two measures indicates that pathways with significant activity may not imply its significance in coherence and vice versa (Fig. **5**). Therefore the joint distributions of gene set activity and gene set coherence shown in Fig. (**5**) further justified the need of testing both activity and coherence for better gene set recognition. These results clearly demonstrate



PATHWAY: BIOGENIC_AMINE_SYNTHESIS

Fig. (3). Unsupervised coherence tree structure and the supervised activity scores of the gene set Biogenic_Amine_Synthesis*.

* Based on the data set GDS2110 from GEO. The gene set is not significant in its coherence (p_C=0.163) but significant in its activity (p_A=0.001) with the majority of its genes showing activity that are far above average (50% quantile) among all genes on the microarray.

the capability and advantage of PAGER by distinguish and integrate the two aspects of gene set pattern, the supervised activity and unsupervised coherence.

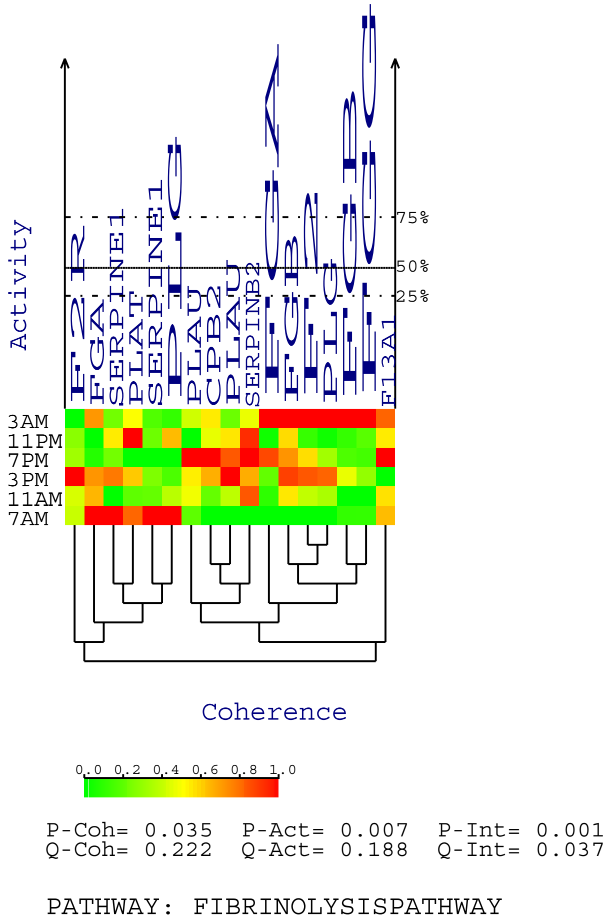


Fig. (4). Unsupervised coherence tree structure and the supervised activity scores of the gene set Fibrinolysis_Pathway*.

* Based on the data set GDS2110 from GEO. The figure shows three distinct gene expression patterns with peak expression at 7AM, 7PM, and 3AM, respectively, in a 24-hour period.

Case Study II: Rat Primary Hepatocyte Time-Course Toxicogenomics Study

To further demonstrate the flexibility and applicability of PAGER, we applied it on the rat toxicogenomics data set generated in one of our lab [47]. This study employed a more complex study design, in which gene expression profiles were monitored in primary rat hepatocytes exposed to cadmium in multi-dosage levels and at multi-time points. Briefly, primary rat hepatocytes were isolated and were exposed to cadmium acetate (0, 1.25 and 2.0 µM) for 2 h. Cells were collected at 0, 3, 6, 12 and 24 h in all three groups (0, 1.25 and 2.0 µM Cd) for mRNA expression profiling using Affymetrix RatTox U34 (RT U34) array which contains ~972 probe sets representing ~800 important toxicology-related genes. The microarray experiment was repeated with hepatocytes from 3 animals, each with 2 replicates (independent cultures) for each dose (3 doses) at each time point (5 time points), resulting in a total of 90 chips (3 animals · 2 replicates · 3 doses · 5 time points). The two replicates were averaged in our analysis.

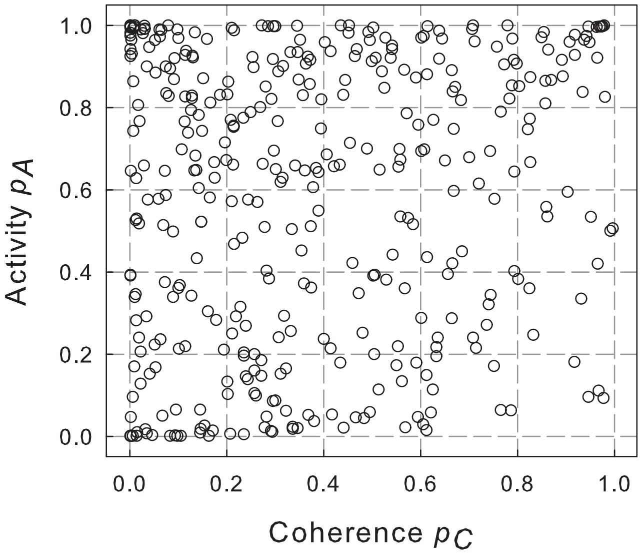


Fig. (5). Joint p-value distribution between gene set coherence and activity across 464 gene sets*.

Note: * Analysis based on the data set GDS2110 from GEO.

We were interested in identifying the gene sets that are responsive at both lower dose (1.25 µM) and higher dose (2.0 µM) cadmium treatments respectively. To measure the treatment effect, expression values were normalized as log-fold change between a treatment (1.25 and 2.0 µM) and its corresponding control (0 µM). The sum of the absolute log-fold changes across three animals and five time points was used as a measure of activity GPA for each transcript. Pearson's correlation coefficients of the log-fold changes were used to measure pairwise GGC. GSA max-mean statistic was used for summarizing activity and MST was used for summarizing coherence. Extremely small (< 5) or large (> 200) gene sets were eliminated in this analysis, resulting in 142 gene sets with size ranging from 5 to 200. Because the RT U34 contains a fairly small number of genes, the 142 gene sets cannot be assumed independent, which violates the assumption of the FDR control procedures. Therefore, we intentionally skipped the FDR step and applied the stringent integration IUT method as well as a rigorous p-value (p<0.005) to determine significance of gene sets.

Table **2** and Table **3** listed 16 and seven significant gene sets that were identified by PAGER for the higher dose (2.0 µM Cd) and low-dose (1.25 µM Cd) treatments, respectively. Many of the identified gene sets were known to be related to toxic chemical responses such as apoptosis, cell death, stress, HSP, and NF-κB regulated gene sets. In addition, all seven gene sets identified at lower dose were also identified at higher dose treatment, which indicates there is a dose-response effect due to cadmium treatment in rat primary hepatocytes. The higher dose treatment induced more drastic cell response than the lower dose treatment did, as demonstrated by the generally lower p-values in the higher dose treatment (2.0 µM Cd) compared with that in lower dose treatment (1.25 µM Cd) for each gene set. Similar to previous study, we also observed that using supervised activity alone is not sensitive enough to recognize some important

Table 2. Significant Gene Sets Identified by PAGER in Rat Primary Hepatocytes Treated with Higher-Dose Cadmium

Gene Set	p_C	p_A	p_I
P53HYPOXIAPATHWAY	0.001	0.001	0.000
PPARAPATHWAY	0.002	0.001	0.000
IL1RPATHWAY	0.006	0.002	0.001
NTHIPATHWAY	0.001	0.013	0.001
NFKBPATHWAY	0.007	0.005	0.001
HSP27PATHWAY	0.001	0.021	0.002
HIVNEFPATHWAY	0.001	0.020	0.002
41BBPATHWAY	0.001	0.021	0.002
APOPTOSIS	0.001	0.017	0.002
TOLLPATHWAY	0.001	0.026	0.003
STRESSPATHWAY	0.002	0.028	0.003
APOPTOSIS_GENMAPP	0.001	0.031	0.004
ATMPATHWAY	0.025	0.032	0.004
TIDPATHWAY	0.030	0.001	0.004
DEATHPATHWAY	0.002	0.038	0.005
GAMMA_HEXACHLOROCYCLOHEXANE_DEGRADATION	0.041	0.041	0.005

Table 3. Significant Gene Sets Identified by PAGER in Rat Primary Hepatocytes Treated with Lower-Dose Cadmium

Gene Set	p_C	p_A	p_I
NTHIPATHWAY	0.001	0.002	0.000
NFKBPATHWAY	0.003	0.001	0.000
PPARAPATHWAY	0.003	0.001	0.000
P53HYPOXIAPATHWAY	0.006	0.001	0.001
TIDPATHWAY	0.009	0.001	0.001
HIVNEFPATHWAY	0.009	0.030	0.004
APOPTOSIS	0.034	0.041	0.005

gene sets. For example in the higher dose treatment, 12 of 16 significant gene sets such as Apoptosis_GenMapp, Toll_Pathway, and Death_Pathway had no significant activity (p_A>0.005). However, PAGER showed (Table **2**) that their integrated gene set patterns were all significant (p_I <0.005). In fact, the activation of these gene sets was corroborated by cytotoxicity experiments which showed a substantial cell death by lactate dehydrogenase (LDH) leakage due to cadmium treatment [47-49]. This case study clearly demonstrated the advantage that the integration of gene set coherence in PAGER greatly contributed to recognizing and identifying these coordinated but moderately active gene sets.

Time course gene expression profiles were ideally suited for gene set analysis because they provide more dynamic information than single-time-point experiment. We performed gene set recognition along the time course and computed the p-values at each time point which allows to monitoring the gene set activity across the time course. Two gene sets showing monotonous p-value dynamics were observed. Both were cancer-related gene sets (Brentani_Death, Brentani_Transcription_Factors) that were manually curated using expressed sequence tags [50]. The cancer-related transcription factors (Brentani_Transcription_Factors) displayed high significant p-values at 0h, and the p-values continuously increase along the 24 hour course. The death gene set (Brentani_Death) showed the opposite dynamics (Fig. **6**). This indicates that the transcription factors were turned on soon after cadmium exposure, whereas the death genes were activated at later stages, which confirms our previous analysis using literature-based text mining [47].

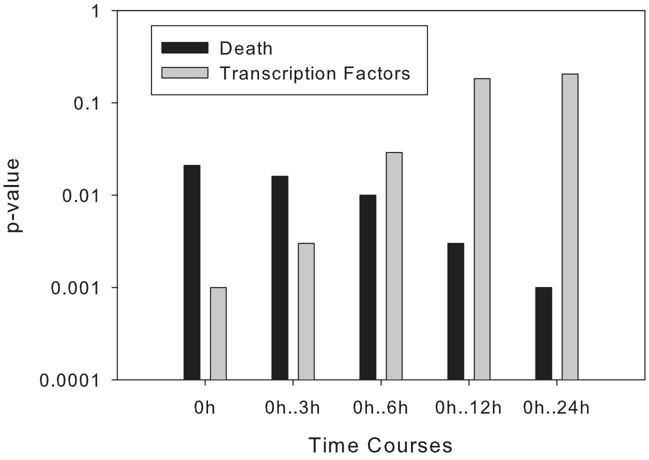


Fig. (6). The p-value dynamics of cancer related death gene set and transcription factor gene set in response to cadmium treatment*.

The cumulative effects of cadmium treatment (2.0 µM Cd) on two gene sets were illustrated by the gene set p-values computed along the time course, i.e., 0h (1 time point), 0..3h (2 time points), 0..6h (3 time points), 0..12h (4 time points), and 0..24h (5 time points).

CONCLUSIONS

As the gene set databases become increasingly comprehensive and accurate, treating genes in unity as sets can help interpret the pattern of gene expression profiles and reduce erroneous discovery. We developed a new gene set recognition framework, PAGER, which distinguish two distinct aspects of gene set pattern—activity and coherence. Therefore PAGER provides a two-dimensional view which helps in revealing the hidden patterns, especially the internal coherence for a given gene set. Another advantage of PAGER comes from its expandable framework with features and options for addressing complex expression study design. Our results showcased PAGER's flexibility and capability on physiological and toxicogenomics data sets which represent typical microarray gene expression profiles. In addition to the classical gene sets such as pathways and gene ontology, experimentally or computationally derived gene sets can be used as well. One of the limitations of PAGER and many other gene set recognition tools is the mathematical simplification of regulatory pathways as gene sets. This simplification could greatly reduce the modeling and computational complexity but it overlooks the topology and interactions between pathway genes. We are currently developing a method that can incorporate the gene interaction information to look into the detailed gene activation cascades throughout the pathways. We predict that the gene-set recognition tools will play critical role in integrative genomics and systems biology.

ADDITIONAL FILES

Additional File 1 – PAGER User's Manual

The source code, compiled program and user's manual can be accessed at http://dengx.bol.ucla.edu/PAGER/PAGER.htm. PAGER and its resources are distributed under the terms of the GNU General Public License version 2 or later.

REFERENCES

[1] B. Ballester, O. Ramuz, C. Gisselbrecht, *et al.* "Gene expression profiling identifies molecular subgroups among nodal peripheral T-cell lymphomas", *Oncogene*, vol. 25, pp. 1560-1570, Mar. 2006.

[2] K. Dybkaer, J. Iqbal, G. Zhou, *et al.* "Genome wide transcriptional analysis of resting and IL2 activated human natural killer cells: gene expression signatures indicative of novel molecular signaling pathways", *BMC Genomics*, vol. 8, pp. 230, 2007.

[3] J. K. Kulski, W. Kenworthy, M. Bellgard, *et al.* "Gene expression profiling of Japanese psoriatic skin reveals an increased activity in molecular stress and immune response signals", *J. Mol. Med.*, vol. 83, pp. 964-975, Dec. 2005.

[4] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, *et al.* "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes", *Nat. Genet.*, vol. 34, pp. 267-273, July 2003.

[5] Z. B. Nagy, P. Gergely, J. Donath, G. Borgulya, and G. Poor, "Gene Expression Profiling in Paget's Disease of Bone: Up-Regulation of Interferon Signalling Pathways in Pagetic Monocytes and Lymphocytes", *J. Bone Miner. Res.*, vol. 23, pp. 253-259, Feb. 2008.

[6] K. Omura, N. Kiyosawa, T. Uehara, *et al.* "Gene expression profiling of rat liver treated with serum triglyceride-decreasing compounds", *J. Toxicol. Sci.*, vol. 32, pp. 387-399, Oct. 2007.

[7] W. R. Swindell, "Gene expression profiling of long-lived dwarf mice: longevity-associated genes and relationships with diet, gender and aging", *BMC Genomics*, vol. 8, pp. 353, 2007.

[8] C. Wang, M. R. Chelly, N. Chai, *et al.* "Transcriptomic fingerprinting of bone marrow-derived hepatic beta2m-/Thy-1+ stem cells", *Biochem. Biophys. Res. Commun.*, vol. 327, pp. 252-260, Feb. 2005.

[9] B. Zeitouni, S. Senatore, D. Severac, C. Aknin, M. Semeriva, and L. Perrin, "Signalling pathways involved in adult heart formation revealed by gene expression profiling in Drosophila", *PLoS. Genet.*, vol. 3, pp. 1907-1921, Oct. 2007.

[10] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin, "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways", *Nat. Genet.*, vol. 31, pp. 19-20, May 2002.

[11] N. Salomonis, K. Hanspers, A. C. Zambon, *et al.* "GenMAPP 2: new features and resources for pathway analysis", BMC *Bioinformatics*, vol. 8, pp. 217, 2007.

[12] M. Ashburner, C. A. Ball, J. A. Blake, *et al.* "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium", *Nat. Genet.*, vol. 25, pp. 25-29, May 2000.

[13] F. Al-Shahrour, P. Minguez, J. Tarraga, *et al.* "FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments", *Nucleic Acids Res.*, vol. 35, pp. W91-W96, July 2007.

[14] B. R. Zeeberg, W. Feng, G. Wang, *et al.* "GoMiner: a resource for biological interpretation of genomic and proteomic data", *Genome Biol.*, vol. 4, pp. R28, 2003.

[15] S. Draghici, P. Khatri, P. Bhavsar, *et al.* "Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate", *Nucleic Acids Res.*, vol. 31, pp. 3775-3781, July 2003.

[16] D. A. Hosack, G. Dennis, Jr., B. T. Sherman, H. C. Lane, and R. A. Lempicki, "Identifying biological themes within lists of genes with EASE", *Genome Biol.*, vol. 4, pp. R70, 2003.

[17] G. Dennis, Jr., B. T. Sherman, D. A. Hosack, *et al.* "DAVID: Database for Annotation, Visualization, and Integrated Discovery", *Genome Biol.*, vol. 4, pp. 3, 2003.

[18] A. Subramanian, P. Tamayo, V. K. Mootha, *et al.* "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", Proc. *Natl. Acad. Sci. USA*, vol. 102, pp. 15545-15550, Oct. 2005.

[19] K. H. Pan, C. J. Lih, and S. N. Cohen, "Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 8961-8965, June 2005.

[20] L. Guo, E. K. Lobenhofer, C. Wang, *et al.* "Rat toxicogenomic study reveals analytical consistency across microarray platforms", *Nat. Biotechnol.*, vol. 24, pp. 1162-1169, Sept. 2006.

[21] A. Boorsma, B. C. Foat, D. Vis, F. Klis, and H. J. Bussemaker, "T-profiler: scoring the activity of predefined groups of genes using gene expression data", *Nucleic Acids Res.*, vol. 33, pp. W592-W595, July 2005.

[22] S. Draghici, P. Khatri, A. L. Tarca, *et al.* "A systems biology approach for pathway level analysis", *Genome Res.*, vol. 17, pp. 1537-1545, Oct. 2007.

[23] B. Efron and R. Tibshirani, "On testing the significance of sets of genes", Technical Report, Stanford University, Palo Alto, CA, USA, 2006.

[24] S. B. Kim, S. Yang, S. K. Kim, *et al.* "GAzer: gene set analyzer", *Bioinformatics*, vol. 23, pp. 1697-1699, July 2007.

[25] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, "Discovering statistically significant pathways in expression profiling studies", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 13544-13549, Sept. 2005.

[26] F. Al-Shahrour, R. az-Uriarte, and J. Dopazo, "Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information", *Bioinformatics*, vol. 21, pp. 2988-2993, July 2005.

[27] J. J. Goeman, J. Oosting, A. M. Cleton-Jansen, J. K. Anninga, and H. C. van Houwelingen, "Testing association of a pathway with survival using gene expression data", *Bioinformatics*, vol. 21, pp. 1950-1957, May 2005.

[28] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 14863-14868, Dec. 1998.

[29] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics", *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 9121-9126, July 2002.

[30] B. A. Novak and A. N. Jain, "Pathway recognition and augmentation by computational analysis of microarray expression data", *Bioinformatics*, vol. 22, pp. 233-241, Jan. 2006.

[31] D. M. Levine, D. R. Haynor, J. C. Castle, *et al.* "Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways", *Genome Biol.*, vol. 7, pp. R93, 2006.

[32] X. Deng, J. Xu, and C. Wang, "Improving the Power for Detecting Overlapping Genes from Multiple Gene Expression Profiles," *BMC Bioinformatics*, to appear. 2008.

[33] J. Xu, X. Deng, T. Hui, D. L. Farkas, A. A. Demetriou, and C. Wang, "Factors released from cholestatic rat livers possibly involved in inducing bone marrow hepatic stem cell priming, recruiting, homing and differentiation", *Stem Cells Dev*, vol. 17, pp. 119-132, Feb. 2008.

[34] T. H. Cormen and T. H. Cormen, *Introduction to algorithms*, 2nd ed. Cambridge, Massachusetts: MIT Press, 2001.

[35] R. A. Fisher, *Statistical Methods for Research Workers*. Edinburgh.: Oliverand Boyd, 1932

[36] R. A. Fisher, "Combining independent tests of significance", *American Statistician*, vol. 12, pp. 30, 1948.

[37] M. C. Whitlock, "Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach", *J. Evol. Biol.*, vol. 18, pp. 1368-1373, Sept. 2005.

[38] R. L. Berger, "Multiparameter hypothesis testing and acceptance sampling", Technometrics, vol. 24, pp. 295-300, 1982.

[39] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies", *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 9440-9445, Aug. 2003.

[40] J. Storey, "A direct approach to false discovery rates", *J R Statist Soc B*, vol. 64, pp. 479-498, 2002.

[41] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes", *Nucleic Acids Res.*, vol. 28, pp. 27-30, Jan. 2000.

[42] M. Kanehisa, S. Goto, M. Hattori, *et al.* "From genomics to chemical genomics: new developments in KEGG", *Nucleic Acids Res.*, vol. 34, pp. D354-D357, Jan. 2006.

[43] T. Barrett, D. B. Troup, S. E. Wilhite, *et al.* "NCBI GEO: mining tens of millions of expression profiles--database and tools update", Nucleic *Acids Res.*, vol. 35, pp. D760-D765, Jan. 2007.

[44] D. R. Lemos, J. L. Downs, and H. F. Urbanski, "Twenty-four-hour rhythmic gene expression in the rhesus macaque adrenal gland", *Mol. Endocrinol.*, vol. 20, pp. 1164-1176, May 2006.

[45] I. J. Elenkov, "Glucocorticoids and the Th1/Th2 balance", *Ann. N. Y. Acad. Sci.*, vol. 1024, pp. 138-146, June 2004.

[46] M. Kawaguchi, M. Adachi, N. Oda, F. Kokubu, and S. K. Huang, "IL-17 cytokine family", *J. Allergy Clin. Immunol.*, vol. 114, pp. 1265-1273, Dec. 2004.

[47] Y. Tan, L. Shi, S. M. Hussain, *et al.* "Integrating time-course microarray gene expression profiles with cytotoxicity for identification of biomarkers in primary rat hepatocytes exposed to cadmium", *Bioinformatics*, vol. 22, pp. 77-87, Jan. 2006.

[48] T. Hamada, A. Tanimoto, and Y. Sasaguri, "Apoptosis induced by cadmium", *Apoptosis*, vol. 2, pp. 359-367, 1997.

[49] A. Soto, B. D. Foy, and J. M. Frazier, "Effect of cadmium on bromosulfophthalein kinetics in the isolated perfused rat liver system", *Toxicol. Sci.*, vol. 69, pp. 460-469, Oct. 2002.

[50] H. Brentani, O. L. Caballero, A. A. Camargo, *et al.* "The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags", *Proc. Natl. Acad. Sci. U. S. A*, vol. 100, pp. 13418-13423, Nov. 2003.

Received: January 16, 2008 Revised: February 15, 2008 Accepted: February 20, 2008