OMICS A Journal of Integrative Biology
Volume 15, Numbers 7 and 8, 2011

DOI: 10.1089/omi.2010.0066

Original Article

A Topology-Preserving Selection and Clustering Approach to Multidimensional Biological Data

Hai Fang,[1,2,*] Yanzhi Du,[2] Lu Xia,[1] Junmin Li,[1] Ji Zhang,[1,2] and Kankan Wang[1,2]

Abstract

Multidimensional genome-wide data (e.g., gene expression microarray data) provide rich information and widespread applications in integrative biology. However, little attention has been paid to the inherent relationships within these natural data. By simply viewing multidimensional microarray data scattered over hyperspace, the spatial properties (topological structure) of the data clouds may reveal the underlying relationships. Based on this idea, we herein make analytical improvements by introducing a topology-preserving selection and clustering (TPSC) approach to complex large-scale microarray data. Specifically, the integration of self-organizing map (SOM) and singular value decomposition allows genome-wide selection on sound foundations of statistical inference. Moreover, this approach is complemented with an SOM-based two-phase gene clustering procedure, allowing the topology-preserving identification of gene clusters. These gene clusters with highly similar expression patterns can facilitate many aspects of biological interpretations in terms of functional and regulatory relevance. As demonstrated by processing large and complex datasets of the human cell cycle, stress responses, and host cell responses to pathogen infection, our proposed method can yield better characteristic features from the whole datasets compared to conventional routines. We hence conclude that the topology-preserving selection and clustering without *a priori* assumption on data structure allow the in-depth mining of biological information in a more accurate and unbiased manner. A Web server (http://www.cs.bris.ac.uk/~hfang/TPSC) hosting a MATLAB package that implements the methodology is freely available to both academic and nonacademic users. These advances will expand the scope of omics applications.

Introduction

With the development of genomics and the emerging of other -omics in recent years, bioscience research enters the era of shifting the emphasis from identifying individual molecules to exploring interconnected relationships among these biological molecules (Collins et al., 2003; Hood et al., 2004). Microarray-based high-throughput technologies provide a routine in the context of network-level understanding (Fang, et al., 2010; Imbeaud and Auffray, 2005; Muller et al., 2008; Rapaport et al., 2007). It allows the simultaneous measurement of temporal changes in gene expression at the genome scale, producing massive amounts of data that can be abundant in information underlying a given biological process. However, such high-dimensional data are also inherent with problematic traits, such as a large number of missing values and small sample size versus huge gene volume, thus limiting the power of conventional data mining tools to effectively characterize the inherent data structure. Even worse, few existing methods fully take into account the topological structure of data and those methods without such considerations will unavoidably tend to be biased and ineffective (Kohonen, 2001).

We view the topological structure of gene expression data as inherent relationships within the data itself. Without loss of generality, the more similar expression patterns the genes exhibit, the closer space they occupy. To intuitively describe these relationships, we conceptually express multidimensional microarray data in terms of a vector space model. This model considers expression values (typically, a log-2 transformed ratio relative to a control) of a given gene across N-related samples as coordinates of the gene in an N-dimensional hyperspace. Accordingly, the set of G genes in the primary expression matrix correspond to data clouds in the

[1]State Key Laboratory of Medical Genomics, Sino-French Research Center for Life Sciences and Genomics, Ruijin Hospital affiliated to Shanghai Jiao Tong University School of Medicine (SJTU-SM), Shanghai, People's Republic of China.
[2]Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and SJTU-SM, Shanghai, People's Republic of China.
*Present address: Department of Computer Science, University of Bristol, Woodland Road, Bristol BS8 1UB, UK.

hyperspace. Data points around the origin of the hyperspace are more likely to correspond to genes with no change in expression or random variation, whereas those located far away from the origin could be genes with an observable expression pattern. In other words, the spatial properties of the data clouds can be considered as a proxy for the topological structure of gene expression data. Although the concept of topology should also involve their interconnections (Basener, 2006), we here ignore such description; actually we rely on an artificial neuronal network [i.e., self-organizing map (SOM); see below] to capture their connectivity. One of the most prominent characteristics for the topological structure in this context is that, featured and informative data tend to be on the fringes of the hyperspace, whereas randomized and artificial data are always centered on the origin of the hyperspace. As shown in the main text (see also Supplemental Fig. 2), the data clouds resulting from the observed data matrix tend to be further away from the origin of the hyperspace than those resulting from the randomized matrix. Therefore, this topological structure of the data provides a basis for the recognition and selections of biologically meaningful genes. One advantage of such topological preservation is to perform exploratory analysis of large and complex multidimensional data, particularly without *a priori* assumption of data structure.

Among various dimensionality reduction techniques to capture such topological structure of the expression data (Lee and Verleysen, 2007), is a SOM. The SOM has been shown to be competent in preserving local and global topological properties (Kohonen, 2001; Tamayo et al., 1999; Xiao et al., 2003). This unsupervised nonlinear algorithm configures output vectors into a topological preservation of the input high-dimensional data in a nonlinear manner, producing a SOM map in which genes with the same or similar expression patterns can be mapped to the same or nearby map nodes. For the sake of human-centered visualization, the topology of SOM usually refers to the lattice structure on the two-dimensional map grid, and is trained by following the structure of the input data. More importantly, the SOM algorithm quantizes the input data [vector quantization (VQ)] and carries out a nonlinear topological preserving projection [vector projection (VP)] in an interactive manner, allowing smoother neighborhood kernels to define the extent of regularization that VP exerts on the VQ (Kohonen, 2001). In addition to the visual benefits provided by the ordered map (Bi et al., 2009; Fang et al., 2008; Xiao et al., 2003), the output matrix of the SOM with an appropriate neighborhood kernel may be more powerful than the expression matrix for exploratory analyses. Depending on the specific questions being addressed, neighborhood kernels can be geared to tasks such as the topology-preserving preprocessing and characterization of microarray data. Another powerful method is singular value decomposition (SVD), which reveals promising potentials in the recognition of biologically meaningful features from microarray data (Alter et al., 2000; Holter et al., 2000). It allows the linear transformation of expression data from a genes×samples hyperspace to a greatly reduced eigengenes×eigensamples space, capturing some characteristic variables that represent essential patterns of temporal changes in gene expression. Although this method is powerful for recognizing dominant expression patterns, the effectiveness of SVD appears to be largely dependent on the choice of data preprocessing (Holter et al., 2000). Such a linear method, if directly applied to complex microarray data, may lead to loss of information. Therefore, it is appealing to first apply a nonlinear method (i.e., SOM) for data preprocessing, followed by a linear method (i.e., SVD) for dominant pattern recognition.

Here, we have introduced an approach termed as topology-preserving selection and clustering (TPSC). It integrates SOM for data pre-processing and SVD for pattern recognition, thus allowing a topology-preserving selection of regulated genes based on sound foundations of statistical inference. Furthermore, we have complemented the approach with an SOM-based two-phase gene clustering, resulting in the categorization of genes in a topology-preserving manner. By incorporating functional/regulatory enrichment analyses, these gene clusters have revealed characteristic features relevant to a given biological process. Potential utilities of these approaches are illustrated by processing large and complex data including datasets of the human cell cycle, various stress responses, and host cell responses to pathogen infection, resulting in the identification of characteristic features in a more accurate and meaningful manner from the whole datasets.

Methods

Procedures to TPSC

Figure 1 summarizes the procedures to implement topology-preserving gene selection and clustering. Also, this figure can serve as a rational recipe for data mining. At the core of TPSC are the sequential applications of SOM using two different neighborhood kernels to complete gene selection and gene clustering, respectively. The brief summary is described below, and details on the underlying algorithms can be found in the rest of this section.

First, the table-format multidimensional microarray data are prepared from public database such as Stanford SMD (Demeter et al., 2007) or NCBI GEO (Barrett et al., 2009). A conventional gene expression matrix is usually tabulated in matrix form, that is, an expression matrix of genes (rows) against different experimental samples (columns).

Second, hybrid SOM–SVD is applied for topology-preserving gene selection. The tabulated gene expression matrix (as input matrix) is subjected to nonlinear transformation using the SOM algorithm with the Epanechikov (EP) neighborhood kernel, with emphasis on vector quantization (VQ). The resultant output matrix (i.e., nodes in rows× samples in columns) serves as an intermediate format for pattern recognition by SVD, which is sequentially followed by dominant eigenvector selection, SVD subspace projection, and distance statistic construction, significant node assessment using the false discovery rate (FDR) procedure for multiple hypothesis testing, and finally the selection of significant nodes and their corresponding genes as defined by the best-matching node (BMN). Those primary gene expression data selected through hybrid SOM–SVD form the characteristic matrix, which proceeds to further analysis.

Third, a SOM-based two-phase gene clustering method is utilized to cluster and visualize the characteristic matrix in a topology-preserving manner. In the first phase, SOM training with a Gaussian neighborhood kernel is applied to better preserve the topology (VP), which can be visualized in component plane presentations (CPPs). In the second phase, the

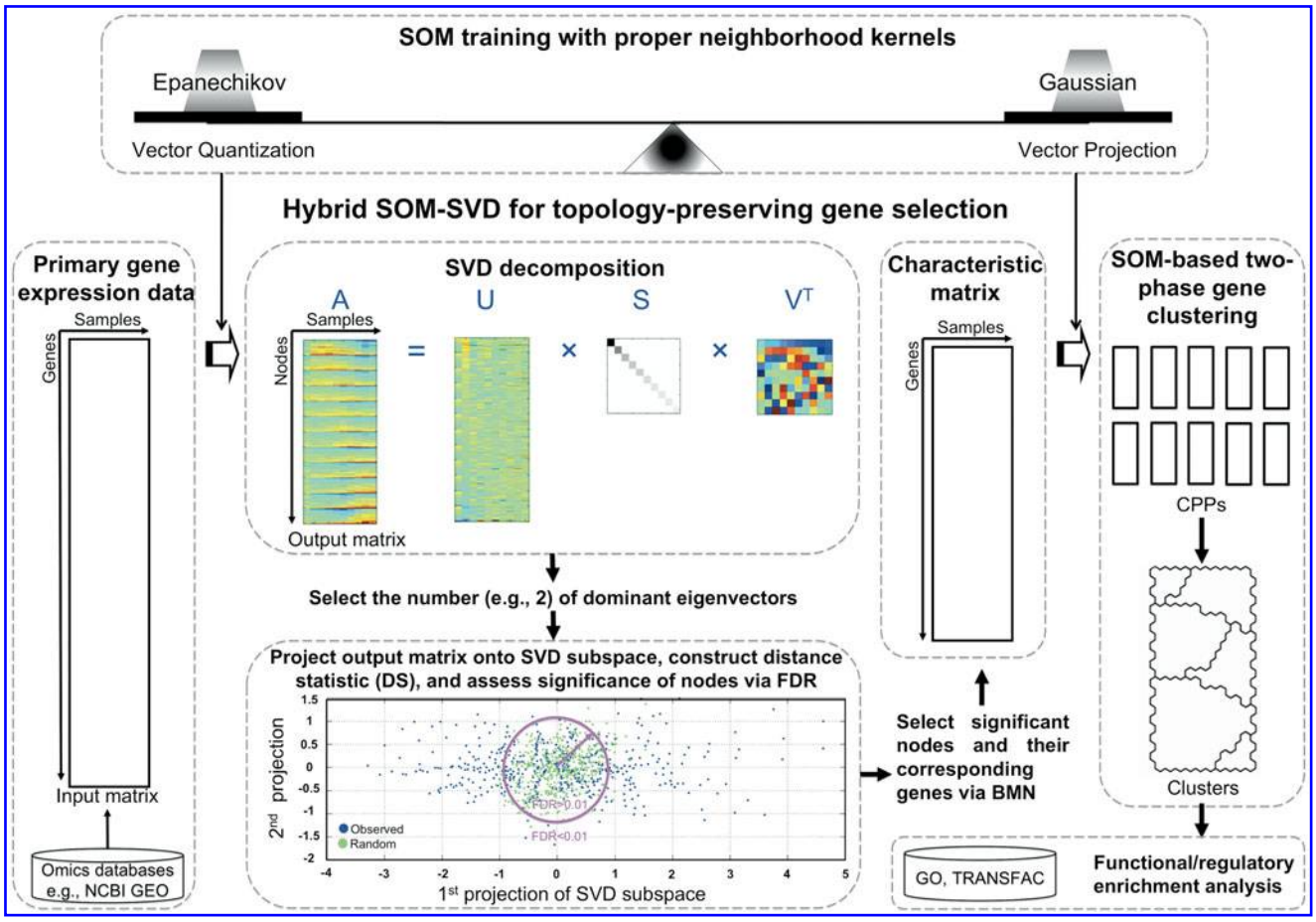

FIG. 1. Schematic flow diagram of the proposed methodology.

resultant SOM is further partitioned based on distance matrix to obtain bases/clusters without *a priori* assumption of the data structure.

Last, the potential utility of bases/clusters is to facilitate many aspects of in-depth mining of biological information, such as downstream functional/regulatory enrichment analysis using external biological annotation resources like Gene Ontology (GO) (Gene Ontology, 2008) and TRANSFAC (Matys et al., 2006).

Data mining using SOM with either an EP or Gaussian kernel function

Based on a vector space model, a gene expression matrix ($G\times N$) forms a specific data cloud in the N-dimensional hyperspace $\Re^N$, reflecting the topological structure of the data. Artificial neuronal network-based SOM (Kohonen, 2001) mimics the data structure of the input space $\Re^N$ by converting the nonlinear relationships between input vectors $\vec{x} \in \Re^N$ into simple geometric relationships on a regular two-dimensional hexagonal grid of map nodes. Each node i is represented by two kinds of vectors: the location vector on the two-dimensional map grid ($\vec{r}_i \in \Re^2$), and the codebook/prototype vector in the N-dimensional hyperspace ($\vec{m}_i \in \Re^N$). Typically, the topology of SOM refers to the lattice structure on the two-dimensional map grid, and is heuristically determined based on the input training data. Given a gene expression matrix ($G\times N$) as the input, the number M of nodes (i.e., the product of the side lengths of the map grid) is initially calculated based on the heuristic formula $M = 5\times\sqrt{G}$. Then, the ratio between the side lengths is set to the square root of the two biggest eigenvalues of the training data. Therefore, the actual side lengths are then finely set so that their product is as close to M as possible. Through VQ and VP in an interactive manner, SOM is trained as follows.

First, the winner map node c is chosen, whose prototype vector is closest to the input vector $\vec{x}$:

$$||\vec{x} - \vec{m}_c|| = \min_i\{||\vec{x} - \vec{m}_i||\}, i = 1, \ldots, M, \quad (1)$$

where $\vec{m}_i$ stands for a prototype vector of node i, $\vec{m}_c$ for the prototype of the winner node c, $\vec{x}$ for an input vector of the gene, and $||\bullet||$ for the Euclidean norm.

Next, the winner node c and its topological neighbors are stretched toward the input vector $\vec{x}$ in the input space $\Re^N$ by

$$\vec{m}_i(t+1) = \vec{m}_i(t) + \alpha(t)h_{ci}(t)[\vec{x}(t) - \vec{m}_i(t)], \quad (2)$$

where $\alpha(t)$ is the learning rate at training time t, and $h_{ci}(t)$ is a smoother neighborhood kernel function centered on the winner node c. The neighborhood kernel $h_{ci}(t)$ dictates the topology of the map grid, and thus defines the extent of regularization that VP exerts on VQ. In terms of the EP or Gaussian function, it can be written respectively as

$$h_{ci}(t) = \max\left\{0, 1 - \frac{||\vec{r}_c - \vec{r}_i||^2}{\sigma^2(t)}\right\} \text{or} \quad (3)$$

$$h_{ci}(t) = \exp\left(-\frac{||\vec{r}_c - \vec{r}_i||^2}{2\sigma^2(t)}\right), \quad (4)$$

where the positive integer $\sigma(t)$ defines the width of the kernel at training time t, and $\vec{r}_c$ and $\vec{r}_i$ are respectively the location vectors of map node c and i on the map grid. They both share the property that the value equals one when $||\vec{r}_c - \vec{r}_i|| = 0$, and goes to zero as $||\vec{r}_c - \vec{r}_i|| \to \infty$. The EP neighborhood kernel places relatively more emphasis on local topological relationships than the Gaussian function and prefers VQ to VP, whereas the Gaussian neighborhood kernel better preserves global topology relationships and thus prefers VP to VQ. Accordingly, we propose the use of EP function for tasks such as topology-preserving gene selection and use of Gaussian function with the aim of global gene clustering.

Hybrid SOM–SVD for topology-preserving gene selection

Decomposition by SVD. The output matrix A (M nodes$\times N$ samples) created by SOM with EP function (Eq. 3) is then decomposed by SVD into

$$A = USV^T, \quad (5)$$

where U is an $M\times N$ matrix whose columns are the left singular vectors (eigensamples), V^T is an $N\times N$ matrix whose rows are the right singular vectors (eigenvectors), and S is an $N\times N$ diagonal matrix of singular values, whose on-diagonal entries (eigenexpressions) are in descending order (Alter et al., 2000).

Selection of the dominant eigenvectors. Relative eigenexpression is defined to indicate the relative significance of the kth eigenvector in terms of the fraction of the overall variation captured:

$$RV_k = s_k^2 \Big/ \sum_j^R s_j^2, \quad (6)$$

where s_k is the kth singular value and R is the rank of the matrix A. SVD is performed on a row- and column-wise permutation of the output matrix A to get the relative eigenexpression spectrum of randomized variables. The similar complete randomization as the reference has been initialized in SVD analysis (Holter et al., 2000). Repeat such randomizations multiple times (e.g., 100) and select those eigenvectors as dominant eigenvectors, the observed relative eigenexpression of which is beyond the maximum random relative eigenexpression at the probability of at least 99% (corresponding to $p < 0.01$).

SVD subspace projection and distance statistic construction. The matrix A, thus, can be represented by the chosen L eigenvectors in the subspaces $\Re^L$ obtained by SVD. Projection of each prototype vector onto the SVD subspaces are performed to obtain the different projection values, forming the projection vector $\vec{q} \in \Re^L$. Let the coordinate-wise zero point be the origin (denoted as $\vec{o}$). For each node, different projection values are integrated into distance statistic (DS),

$$DS = ||\vec{q} - \vec{o}||^2, \quad (7)$$

the larger value of which likely indicates significance of the node. The intuition behind the distance statistic construction is to incorporate the prototype vector of each node in the SVD subspace into a unified yet quantized value.

Assessment of significant nodes. Based on distance statistic, the statistical significance of a node (and its prototype vector) can be assessed by the method of FDR (Benjamini and Hochberg, 1995) to account for multiple hypothesis tests. The steps for selecting significant nodes, and subsequently genes, are described as follows. First, compute the node-specific projection vector $\vec{q}$ and construct DS from $\vec{q}$ to $\vec{o}$ (Eq. 7), and order the distances: $DS_{r1} \le DS_{r2} \le \cdots \le DS_{rM}$. Second, obtain $b = 1, \ldots, B$ (i.e., 1,000) reference datasets, forming a $M\times N$ matrix A^b, which is generated by randomly permuting the output matrix A in both row and column directions. Analogously, compute the projection values of the reference prototype vectors on the chosen L eigenvectors to obtain projection vector $\vec{q}^b$, and then construct DS^b from $\vec{q}^b$ to $\vec{o}$ (Eq. 7), and order the distances: $DS^b_{r1} \le DS^b_{r2} \le \cdots \le DS^b_{rM}$. Third, assess the statistical significance of each node in terms of the FDR. For the rith node as ordered, compute the number of nodes called significant ($rM - ri + 1$), and then calculate the median number of nodes falsely called significant by calculating the median number of nodes among each of the B sets of reference data, whose DS^b_{rj} satisfy: $DS^b_{rj} \ge DS_{ri}, j = 1, \ldots, M$. Thus, the FDR for the rith ordered node can be quantized as the median number of falsely called nodes divided by the number of nodes called significant. Finally, identify significant nodes to control the FDR, and subsequently select their corresponding genes. After these steps, the characteristic matrix is extracted in a topology-preserving manner.

SOM-based two-phase gene clustering

The distance matrix-based clustering of SOM (Vesanto and Sulkava, 2002) is utilized to characterize the genes selected by hybrid SOM–SVD. In the first phase, the input data are trained by SOM with the Gaussian kernel (Eq. 4) to better preserve the topology of the data. In the second phase, the trained map is divided into a set of bases/clusters using a region growing procedure, which starts with local minima of the distance matrix as seeds (Eq. 8). The set of seed nodes i can be found by

$$f(\vec{m}_i, N_i) \le f(\vec{m}_j, N_i), \forall j \in N_i, \quad (8)$$

where $\vec{m}$ stands for the prototype vector, N_i and N_j for the sets of neighboring map nodes i and j, and the function $f(\vec{m}_i, N_i) = median\{||\vec{m}_i - \vec{m}_k|| \mid k \in N_i\}$ for the median distance between the map node i and its neighboring map nodes N_i. Then, let each local minimum be one base and find each unassigned neighboring node with smallest distance to each base. The step is repeated until all the remaining nodes are finally designed to the corresponding base.

Functional/regulatory enrichment analysis

External annotated biological databases such as Gene Ontology (GO) and TRANSFAC can be used for interpretation of

bases/clusters in terms of functional/regulatory relevance. The search for GO functional enrichments in each base/cluster was conducted with the MAPPFinder program (a component of GenMAPP version 2.0) (Doniger et al., 2003). The Westfall-Young adjusted p-value after multiple hypothesis testing was calculated as a statistical measure of significance for each GO functional category according to the published protocols. Conserved targets of transcription factors, as represented by positional weight matrix (PWM) in TRANSFAC, were curated (Xie et al., 2005) and utilized for hypergeometric distribution-based regulatory enrichment analysis in each base/cluster. Hypergeometric p-values were first calculated, and then, the Benjamini-Hochberg (BH)-derived step-up procedure of FDR (Benjamini and Hochberg, 1995) was applied to account for multiple hypothesis testing. Hypergeometric distribution-based BH-derived FDR was used to assess the significance of the PWM regulatory enrichments.

Results

Identification of many more cell cycle genes with a characteristic period

Unlike many other algorithms (Lee and Verleysen, 2007), SOM has properties of both VQ and VP, carrying out a topologically preserving projection of the prototype vectors from a high-dimensional input space onto a low-dimensional grid (Kohonen, 2001). The dimensionality reduction by this nonlinear projection algorithm may allow the retaining of information inherent to the original data. To test this assumption, we first selected a cell-cycle dataset from HeLa cells as a model system. This dataset contains expression values of 29,621 genes (represented by 43,198 cDNA elements) across 48 samples (representing 47 time points) (Whitfield et al., 2002). Through normalization from original publication and a standard clean-up procedure, with a threshold missing value of less than 20%, 36,549 cDNA elements were selected and subsequently used for SOM training with their original expression values. SOM training was conducted with 962 (37×26) nodes using the EP kernel function. As shown in Figure 2A, numerical values of the output matrix were first illustrated by CPPs, demonstrating visual advantages for the detection of genome-scale transcriptional features for each of the samples during the time course (Du et al., 2006; Fang et al., 2010; Wang et al., 2009; Zheng et al., 2005). To further examine the tightness/accuracy of local gene clustering, nodes are compressed with genes that have highly similar expression patterns (as exampled in Fig. 2B), implicating that the output matrix is well representative of the original data matrix in terms of information inherent. Because features and artifacts are separated to different nodes, this output matrix may function as an intermediate format for the selection of biologically meaningful genes.

To establish an automated procedure for pattern recognition and subsequent gene selection, we applied SVD complemented with an FDR-based method to the output matrix. As shown in the Figure 3A, the decomposition of the output matrix by SVD results in the formation of three matrices U, S, and V^T, in which the temporal pattern of any node can be expressed as a linear combination of eigenvectors, analogously described in previous publications (Alter et al., 2000; Holter et al., 2000). To set up a threshold with statistical inference for selection of those observed dominant eigenvectors, the output matrix was randomly permuted in both row/node and column/sample directions, and then similarly decomposed by SVD to generate a set of randomized eigenvectors for multiple times. By comparing the relative contribution to the overall variation of each observed eigenvector with that of randomly generated eigenvectors, the first seven dominant eigenvectors ($p < 0.01$) were selected to form a seven-dimensional SVD subspace (see Methods). Contributions of eigenvectors (observed) and randomized eigenvectors from a randomization were illustrated on the left panel in Figure 3B. To address impacts of the permutations used on the estimation of dominant eigenvectors (Supplemental Doc. 1), we also generated randomized matrix only by row-wise permutation or only by column-wise permutation (Supplemental Fig. 1A). Compared to these partially randomized matrices, the simultaneously permutated matrix allowed us to identify all dominant eigenvectors from those minor eigenvectors (Supplementary Fig. 1B). Moreover, pairwise scatter plots projected on the SVD seven-subspace showed that distribution resulting from the output matrix consistently stayed further away from the origin of the space than that resulting from the simultaneously permutated matrix (Supplemental Fig. 2). Based on this observation, we took advantage of spatial structure and constructed DS through the calculation of Euclidian distance of each data point (observed or randomized) to the origin of the subspace. The significance of each node was assessed in terms of FDR accounting for multiple hypothesis tests. Under the FDR of 0.1, 514 nodes were selected and subsequently used for analysis. As indicated in the grid map (the right panel in Fig. 3B), most of the selected nodes are organized to edge or corner areas of the map, whereas filtered nodes are mapped to center areas of the map. As exemplified in Figure 2, nodes with periodic gene expression patterns indeed locate at the corner of the map, whereas those nodes without periodic variations tend to be at the center (data not shown), thus allowing the selection in a topology-preserving fashion.

To further show that information inherent to the primary data was largely retained in the output matrix retained, we applied Fast Fourier transform (FFT) (Duhamel and Vetterli, 1990) to the selected nodes and found that 61 of them (representing 1,896 cDNA elements) revealed characteristic patterns of the cell cycle period of 15.67 h, which was the most dominant cycle period in this setting of dataset. Of these cDNA elements, 1,158 well matched the cycle period. In contrast, when the same criteria were applied to previously identified cell cycle-related cDNA elements (Whitfield et al., 2002), it yielded only 758 cDNA elements matching the cycle period. As compared in Figure 3C, the overlaps between results of two analyses are represented by 536 cDNA elements, whereas the nonoverlaps are respectively represented by 622 and 222 cDNA elements. This demonstration partially validates the power of the output matrix of SOM for pattern recognition and can allow for topology-preserving gene selection when coupled with SVD.

Probing into the global structure of genes regulated during various stress responses

To further explore the power of combining SOM with SVD for topology-preserving gene selection, we chose a second

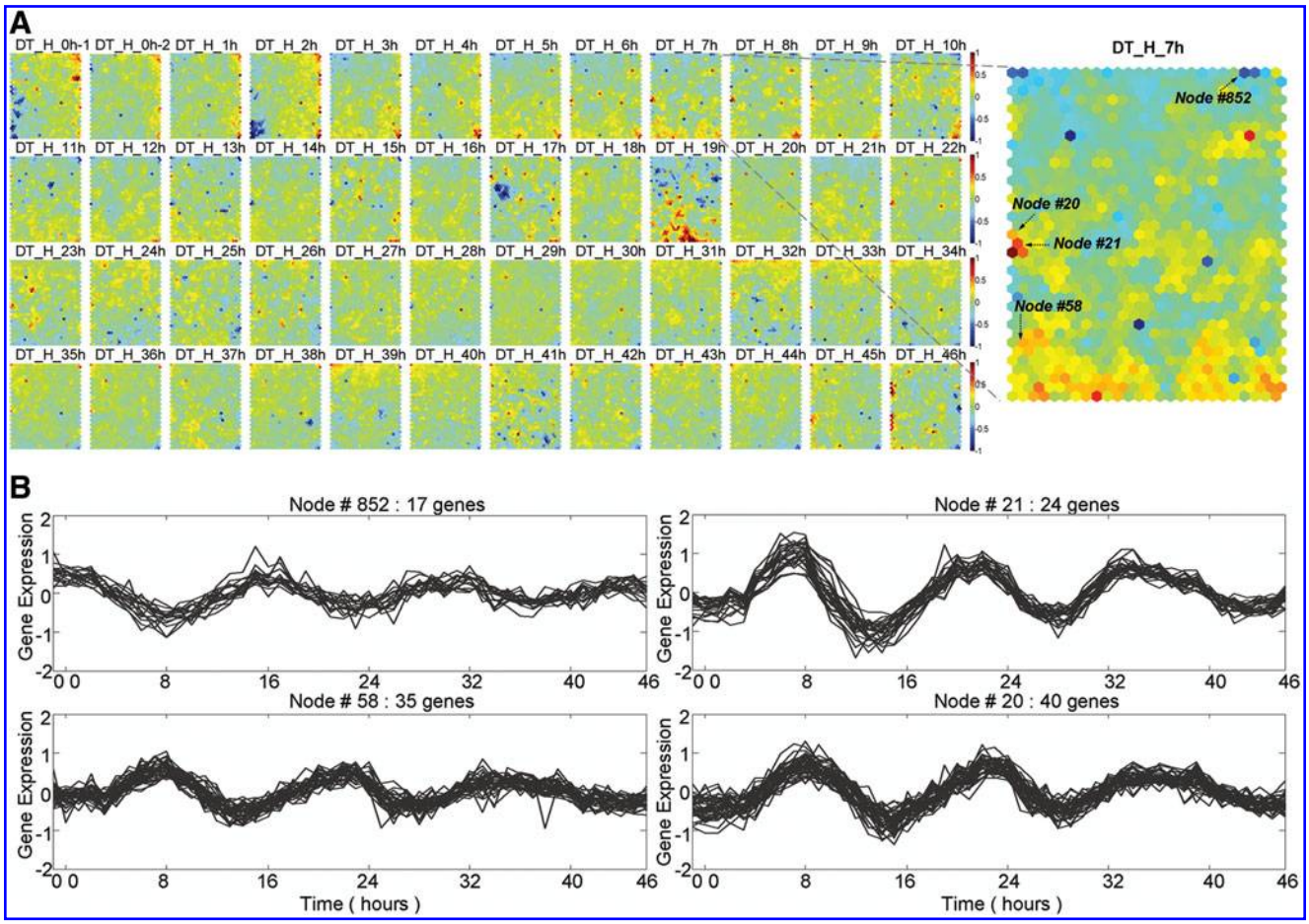

FIG. 2. Nonlinear transformation of HeLa cell proliferation dataset using SOM with EP kernel function. (**A**) Component plane presentations of the SOM outputs, depicting genome-wide transcriptional changes during the proliferation. Each presentation illustrates a sample-specific transcriptome map, in which all the upregulated (in red), downregulated (in blue), and moderately regulated (in yellow and green) genes are well delineated. Notably, the same position in all presentations contains the same group of coexpressed genes. Color coding index stands for expression values of genes, with the brighter color denoting the higher value. DT_H denotes double-thymidine synchronization of cell proliferation. On the right panel is enlarged illustration of DT_H_7h, wherein those indexed nodes below are indicated. (**B**) Illustration of gene expression patterns in nodes with typical cell cycle periods through simple line graphs, as exampled by nodes 852, 21, 58, and 20.

model dataset containing expression values of 25,802 genes (represented by over 40,000 cDNA elements) across 76 samples (i.e., three human cell lines of three different stress treatment series excluding zero time points) (Murray et al., 2004). Similarly, the expression matrix was first trained by SOM with 972 (36×27) nodes using the EP kernel function (Supplementary Fig. 3). Also, SVD was applied to decompose the output matrix of the SOM, followed by the eigenvector selection procedure (Supplementary Fig. 4A). Following an FDR procedure based on eight selected dominant eigenvectors ($p < 0.01$; Supplementary Fig. 4B), a stringent cutoff (FDR < 0.01) allowed the selection of 491 nodes, representing 17,524 cDNA elements.

To characterize these cDNA elements in a more comprehensive manner, we introduced an SOM-based two-phase gene clustering approach. Direct benefits of this approach include the reduction of the complexity of the clustering task, and the ability to get more reliable estimates of clusters in a topology-preserving manner. As shown in Figure 4A, the output matrix of the SOM is displayed by CPPs, showing transcriptome responses of each cell type to different stress conditions. Figure 4B illustrates the second phase of the gene clustering, in which nodes are well organized into bases/clusters according to their neighborhood relationships. A total of 46 bases were identified, without *a priori* assumption of data structure. Comparatively, direct application on primary data distorted the topology of global clustering, and using different clustering approaches could not obtain topology-preserving clusters (see Supplementary Doc. 2, and Supplementary Figs. 5 and 6).

When genes in each of the 46 bases are illustrated through color-coded displays, highly similar patterns of gene expression are observed (Fig. 5). This provides evidence that highly accurate gene clustering at the genome-wide scale is obtainable through our integrated approach. Importantly, this highly accurate gene clustering may facilitate many aspects of in-depth mining of functional and regulatory features. Searches for GO functional annotations enriched in bases were

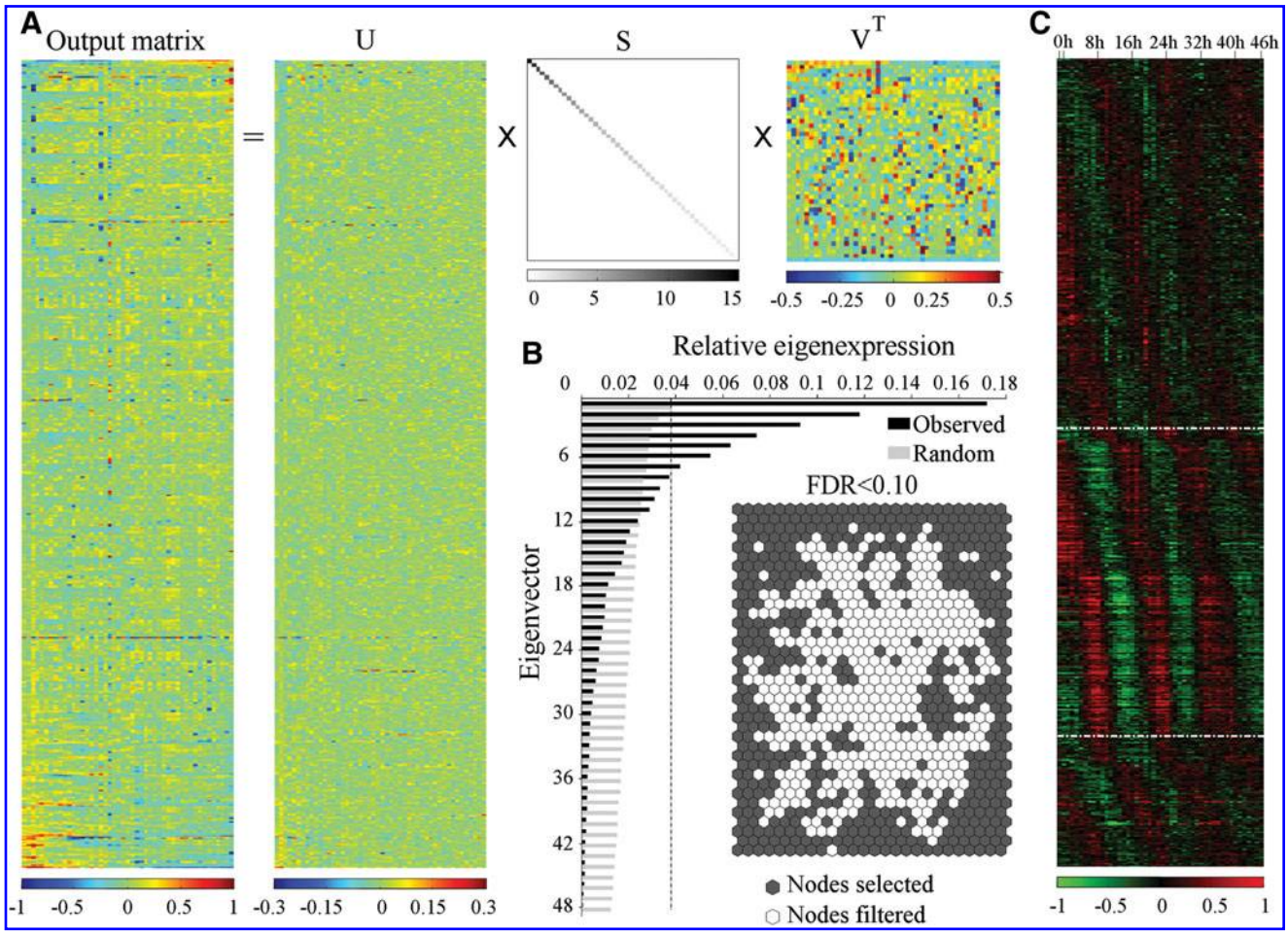

FIG. 3. Linear decomposition by SVD, feature selection and cell cycle period recognition by Fourier transformation. (**A**) Decomposition of the output matrix by SVD. The matrix is decomposed into U, S, and V^T. Values of eigensamples (columns of U), eigenexpressions (on-diagonal entries of S), and eigenvectors (rows of V^T) are coded either by color spectrum or gray magnitude. (**B**) Topology-preserving selection through the integration of FDR. Bar displays on the left respectively illustrate contributions of eigenvectors (observed) and randomized eigenvectors from a randomization. The grid map on the right demonstrates nodes selected (in heavy gray) or omitted (in white) under the indicated FDR. (**C**) Fourier transformation for the recognition of cell cycle genes with a typical period of 15.67 h. The cDNA elements unique to this study are indicated in the upper part, whereas those unique to the previous selection are shown in the lower part. The overlaps between the results of two analyses are in the middle. Expression patterns within each region are ordered by phase information from $-\pi$ to π.

performed with MAPPFINDER program (Doniger et al., 2003). Also, we utilized transcription factor (TF) putative conserved targets (Xie et al., 2005) to perform PWM regulatory enrichment analysis to infer common regulatory features associated with these bases.

Our automated approaches produce more biologically meaningful results than those published previously. The representative expression patterns during various stress responses in cultured human cells can be identified in a concise yet unbiased manner. As illustrated in Figure 5, the bases/clusters obtained provide a meaningful representation of biologically relevant expression patterns inherent in microarray data, as highlighted by stress condition-specific bases (Fig. 5A), stress-specific response bases (Fig. 5B), general stress response bases (Fig. 5C), fibroblast-dependent bases (Fig. 5D), bistress response bases (Fig. 5F), and cell cycle-relevant bases (Fig. 5G). Moreover, these well-organized gene bases can facilitate in-depth mining of functional/regulatory relevance. For instance, genes in base 20 and 16 (Fig. 5C) reveal more general responses to various stress treatments. Genes in base 20 (upregulated) are significantly related to cell homeostasis and ER transporting networks, suggesting the occurrence of cellular reorganization upon various stress treatments. Regulatory enrichment analysis shows that genes in base 20 are largely regulated by multiple stress-responsive transcription factors such as XBP1, ATF3, ATF4, and ATF6. On the contrary, genes in base 16 (downregulated) are functionally associated with DNA metabolism, RNA processing and ribosome biosynthesis, and are mostly regulated by survival-related transcription factors (e.g., NRF1, YY1, and MYC), implicating the suppression of cell growth related activities upon various stress treatments. Notably, genes in bases 1 and 2 of Figure 5G are prominently downregulated in late stages of ER stress and oxidative agent-treated fibroblasts. They are functionally involved in various aspects of cell cycle and regulated by cell cycle-specific transcription factors (i.e., NFY, E2F). Further

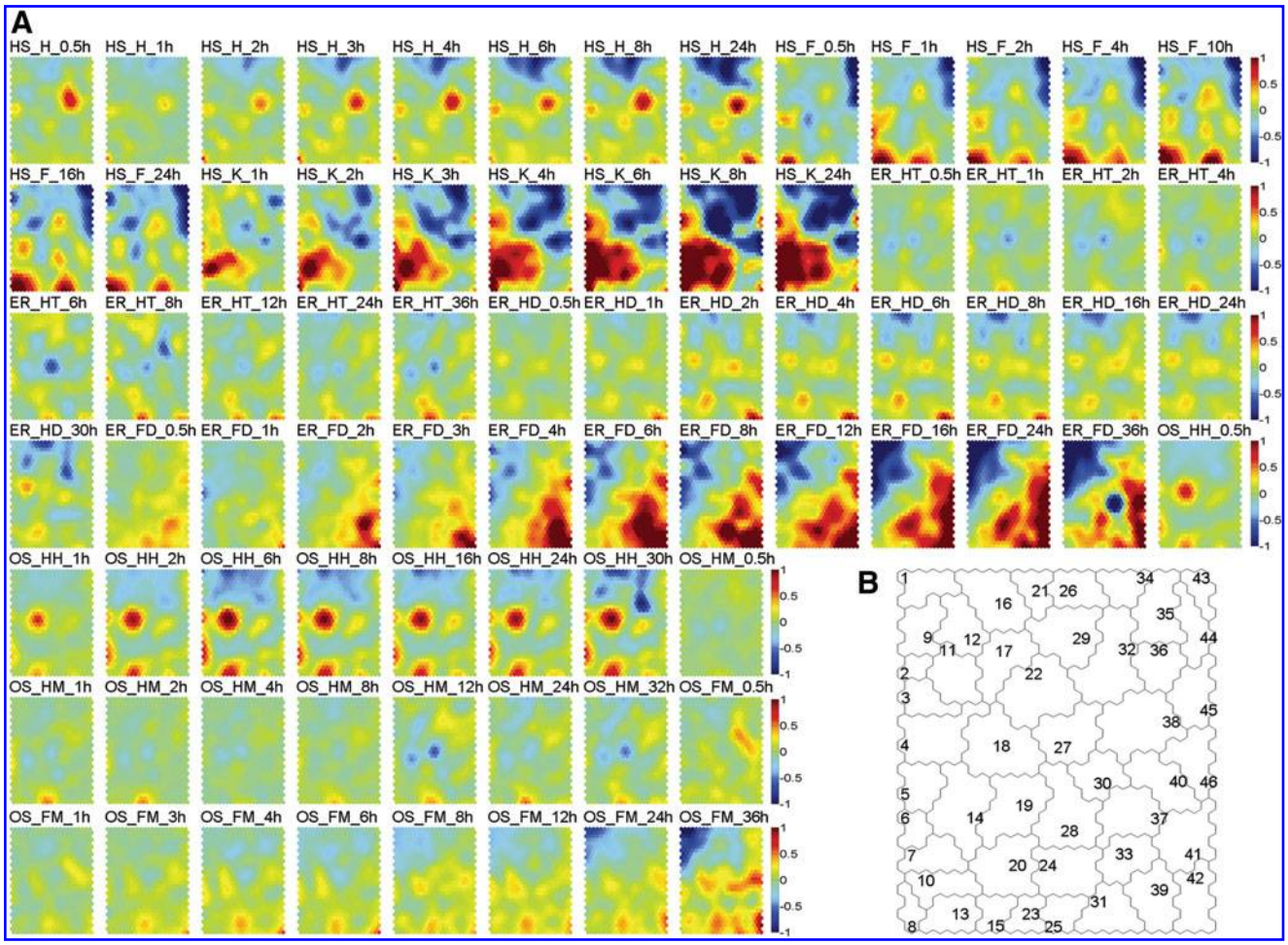

FIG. 4. SOM based two-phase gene clustering of various stress response dataset. (**A**) Component plane presentations of the SOM outputs. (**B**) Ideogram illustration of 46 gene bases on a SOM grid map. The index of each base is marked in the seed node as indicated. Abbreviations of stress responses of various cell type to different conditions are listed as follows, that is, stress responses (heat shock, HS; ER stress, ER; oxidative stress, OS) of various cell type (fibroblasts, F; HeLa, H; K-562, K) to different conditions (HS: cells shifted from 37 to 42°C; ER: cells induced by tunicamycin, T, or DDT, D; OS: cells induced by hydrogen peroxide, H, or menadione, M). For example, OS_FM_36h denotes oxidative stress (OS) in fibroblasts (F) induced by menadione (M) at 36 h.

examination by cell cycle genes identified in the previous case (Fig. 3C), about a half of the genes clustered in cell cycle bases belong to the category periodically regulated by cell cycle (15.67 h).

Revealing the impacts of host variability on responses to pathogen infection

To test the broad utilities of our automated approach, we chose a third dataset generated by Affymetrix GeneChip HG-U133A (Tailleux et al., 2008). The dataset was relevant to time course transcriptome profiles between pathogen (i.e., *Mycobacterium tuberculosis*)-infected human dendritic cells (D) and macrophages (M) in nine independent healthy donors/hosts. Compared to the corresponding reference transcriptome at the 0 time point of infection, there remained expression values of 22,283 probesets across 54 samples, i.e., nine donors (1–9)×two cell types (D and M)×three infection time points (4, 18, and 48 h). The gene expression matrix was then subjected to the hybrid SOM–SVD for gene selection, resulting in 5,280 probesets. Those selected probesets were then subjected to gene clustering and visualized by CPP-SOM (Fig. 6A), and followed by sample classification with the unsupervised hierarchical algorithm (Fig. 6B), yielding much more meaningful results than those published previously (Tailleux et al., 2008).

Similar to the previously published results, dendritic cells and macrophages were consistently classified into two groups, independent of the infection time points and donors. Within each group, samples at the 4-h infection time point were clustered together, independently of the donors analyzed, whereas samples at the other two time points were not separated. Unlike the previous results, macrophages from the same donor infected at the 18- and 48-h time points were exclusively clustered together, but the same biases were not observed with dendritic cells. This result indicates that donor-to-donor variability for macrophages at the 18- and 48-h time points is greater than that for dendritic cells. The preferential impact of donors on macrophages at the 18- and 48-h time

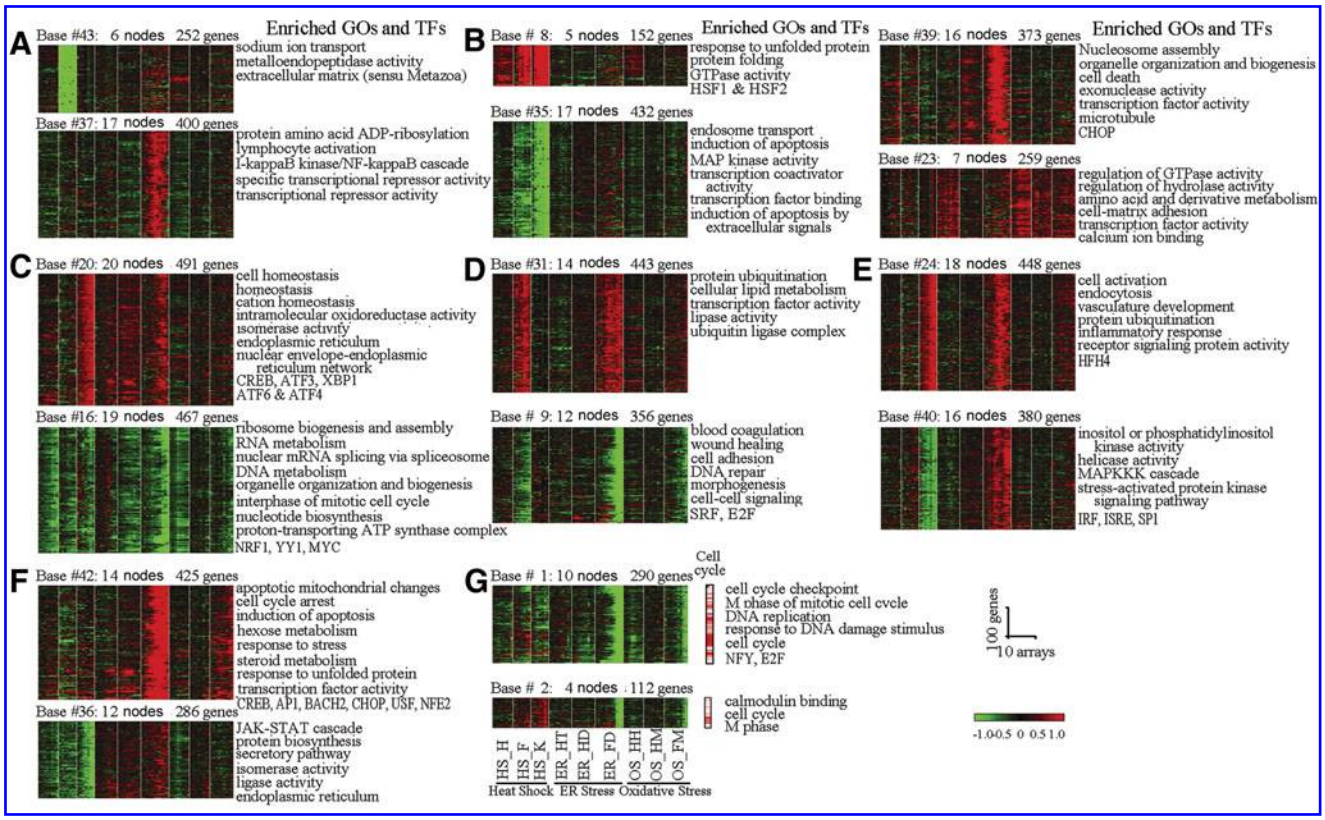

FIG. 5. Illustration of gene expression patterns, corresponding GO and transcription factor enrichments in exampled bases. (**A**) Stress condition-specific bases. (**B**) Stress-specific response bases. (**C**) General stress response bases. (**D**) Cell line-dependent bases. (**E, F**) Complex response bases. (**G**) Cell cycle bases. The cDNA elements periodically regulated by cell cycle in the category of G are marked in red within the bar on the right.

points was not detected in the original publication. Further results revealed by our approaches can be found in Supplementary Figs. 7–9.

Discussion

Our proposed approach TPSC (including hybrid SOM–SVD for gene selection, two-phase gene clustering in a topology-preserving manner) (Fig. 1) shows several distinct advantages. First, hybrid SOM–SVD takes spatial properties of data into consideration and permits topology-preserving selection of genes that show statistically significant changes in expression, omitting conventional arbitrary gene selection procedures. In particular, data preprocessing through EP kernel function based SOM yields an output matrix, allowing the separation of features, artifacts, and nonsignificant variables into different nodes. Linear decomposition of the output matrix by SVD illustrates the relative contributions of these feature- and artifact-containing nodes. Further integration of an FDR-based procedure permits the entire gene selection process to be defined on the basis of statistical inference. As demonstrated in the processing of human cell cycle data (Figs. 2 and 3), Fourier transformation of the genes selected through our automated process is much more effective than that of the genes selected through other methods.

Second, the two-phase gene clustering approach allows the clustering of all the selected genes into well-organized bases/clusters in a topology-preserving fashion, producing much accurate and complete gene clustering without *a priori* assumption of data structure. The comprehensive view of global clustering is of great use, especially for large and complex microarray data. For example, it is straightforward to perform downstream in-depth analyses of gene clusters in terms of functional/regulatory relevance. As demonstrated in the processing of the various stress response data (Figs. 4 and 5), most gene clusters well represent characteristic expression patterns under various stress responses and each associated with specific functional/regulatory features.

Third, TPSC can be applicable to other high-dimensional biological data (e.g., clinical sample plus time course-based microarray data). As demonstrated by the processing of pathogen-infected host cell response data (Fig. 6), we have revealed the preferential impacts of donor-to-donor variability on macrophages, thus generating new sound hypotheses.

Last but not least, we emphasize the topology-preserving nature of the methodology compared to our previous work (Xiao et al., 2003) or others (Tamayo et al., 1999). As shown in Figure 4A and Supplemental Figure 5, direct application of the clustering method on primary data may distort the topology of global clustering. Hence, we first select characteristic expression patterns from primary microarray data (while filtering out those noisy data at the center of data hyperspace) through hybrid SOM–SVD in a topology-preserving manner, which are then subjected to topology-preserving gene clustering. Considering the fact that SOM can preserve both local and global topological properties of high-dimensional data

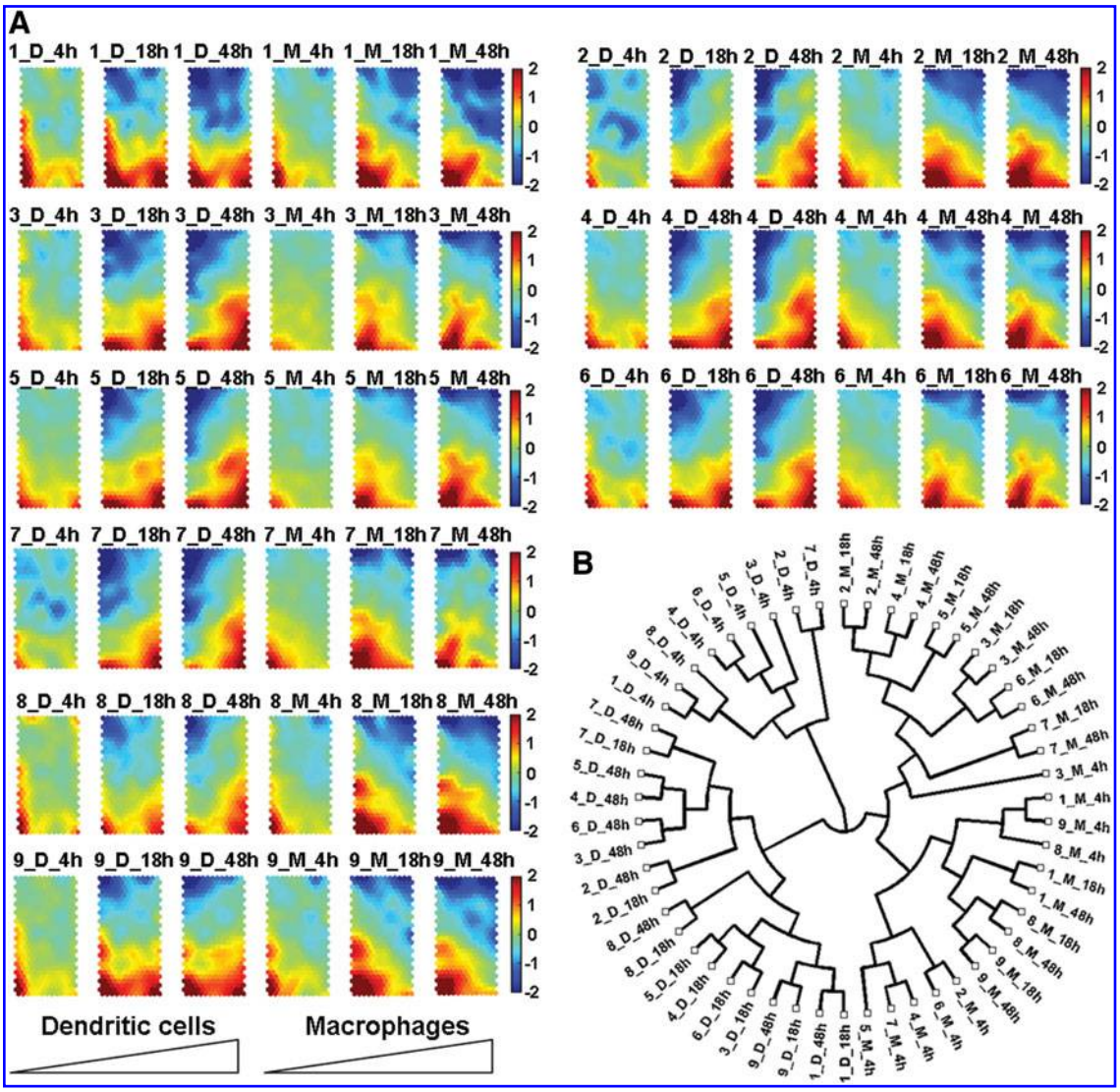

FIG. 6. The preferential impacts of host variability on macrophages rather than dendritic cells in response to pathogen infection. (**A**) Component plane presentations integrated SOM visualizations of gene expression data, following hybrid SOM–SVD to process time course (4 h, 18 h, and 48 h) transcriptome data between pathogen-infected human dendritic cells (D) and macrophages (M) in nine independent healthy donors/hosts (1–9). Intuitively, 6_M_48h denotes the infection response of macrophages in the sixth donor at 48 h. (**B**) Radial dendrogram of unsupervised hierarchical samples, which were classified using those transcriptome data processed by hybrid SOM–SVD. Of note, macrophages from the same donor exclusively stayed together at the infection of 18- and 48-h time points, but not for dendritic cells. Comparatively, the method in the original publication did not detect the preferential impact of donors on macrophages at the 18- and 48-h time points (Tailleux et al., 2008).

(Kohonen, 2001), our work will greatly enrich the other existing SOM-derived methods such as an SOM–support vector machine (SVM) approach (Wu et al., 2005). To the best of our knowledge, authors in this study used SOM to improve the supervised classification performance of the SVM. In this sense, our method can be considered to be unsupervised; it follows the data structure for topology-preserving gene selection and clustering.

Conclusions

In summary, we started this project with the awareness: how to exploit the topological structure of multidimensional microarray data might be required for in-depth mining of information relevant to a biological process of interest. Based on initial efforts (Fang et al., 2010; Wang et al., 2009), we here report an automated approach for topology-preserving gene selection and gene clustering at the genome scale. As demonstrated, our analytical approach works well in extracting characteristic features from the whole dataset, and seems to be superior in terms of genome-wide selection and clustering in a topology-preserving manner. Also, the implemented packages are provided so that potential readers (and users) can easily follow the accompanying protocols to test the proposed methodology or analyze the microarray data of interest. Promisingly, these approaches may interest a wide range of

investigators and thus expand the scope of omics applications in bioscience research.

Acknowledgments

This work was supported in part by the Knowledge Innovation Program of Chinese Academy of Sciences (KSCX1-YW-22-01), Ministry of Science and Technology of China Grants (2009CB825607 and 2011CB910202), National Natural Science Foundation Grants (30730033 and 90919059), Shanghai Postdoctoral Scientific Program (09R21414900), China Postdoctoral Science Foundation (20090450573), and European Community Grants of FP7 (TB-VIR network, 200973).

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

Supplementary Data

Supplemental Data includes Supplemental Docs. 1–3 (and Supplemental Figs. 1–9). The MATLAB package implementing the proposed methodology, together with step-by-step instructions, is also available at http://www.cs.bris.ac.uk/~hfang/TPSC.

References

Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 97, 10101–10106.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2009). NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37, D885–D890.

Basener, W.F. (2006). *Topology and Its Applications*. (Wiley-Interscience, Hoboken, NJ).

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 57, 289–300.

Bi, Y.F., Liu, R.X., Ye, L., Fang, H., Li, X.Y., Wang, W.Q., et al. (2009). Gene expression profiles of thymic neuroendocrine tumors (carcinoids) with ectopic ACTH syndrome reveal novel molecular mechanism. Endocr Relat Cancer 16, 1273–1282.

Collins, F.S., Green, E.D., Guttmacher, A.E., and Guyer, M.S. (2003). A vision for the future of genomics research. Nature 422, 835–847.

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., et al. (2007). The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic Acids Res 35, D766–D770.

Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C., and Conklin, B.R. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 4, R7.

Du, Y., Wang, K., Fang, H., Li, J., Xiao, D., Zheng, P., et al. (2006). Coordination of intrinsic, extrinsic, and endoplasmic reticulum-mediated apoptosis by imatinib mesylate combined with arsenic trioxide in chronic myeloid leukemia. Blood 107, 1582–1590.

Duhamel, P., and Vetterli, M. (1990). Fast fourier transforms: a tutorial review and a state of the art. Signal Process 19, 259–299.

Fang, H., Wang, K., and Zhang, J. (2008). Transcriptome and proteome analyses of drug interactions with natural products. Curr Drug Metab 9, 1037–1047.

Fang, H., Yang, Y., Li, C., Fu, S., Yang, Z., Jin, G., et al. (2010). Transcriptome analysis of early organogenesis in human embryos. Dev Cell 19, 174–184.

Gene Ontology Consortium. (2008). The Gene Ontology project in 2008. Nucleic Acids Res 36, D440–D444.

Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., and Fedoroff, N.V. (2000). Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci USA 97, 8409–8414.

Hood, L., Heath, J.R., Phelps, M.E., and Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. Science 306, 640–643.

Imbeaud, S., and Auffray, C. (2005). Functional annotation: extracting functional and regulatory order from microarrays. Mol Syst Biol 1, 2005.

Kohonen, T. (2001). *Self-Organizing Maps* (Springer, Berlin).

Lee, J.A., and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction* (Springer, New York).

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34, D108–D110.

Muller, F.J., Laurent, L.C., Kostka, D., Ulitsky, I., Williams, R., Lu, C., et al. (2008). Regulatory networks define phenotypic classes of human stem cell lines. Nature 455, 401–405.

Murray, J.I., Whitfield, M.L., Trinklein, N.D., Myers, R.M., Brown, P.O., and Botstein, D. (2004). Diverse and specific gene expression responses to stresses in cultured human cells. Mol Biol Cell 15, 2361–2374.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J.P. (2007). Classification of microarray data using gene networks. BMC Bioinformatics 8, 35.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., et al. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96, 2907–2912.

Tailleux, L., Waddell, S.J., Pelizzola, M., Mortellaro, A., Withers, M., Tanne, A., et al. (2008). Probing host pathogen cross-talk by transcriptional profiling of both *Mycobacterium tuberculosis* and infected human dendritic cells and macrophages. PLoS One, 3, e1403.

Vesanto, J., and Sulkava, M. (2002). Distance matrix based clustering of the self-organizing map. In *Proceedings of the 12th International Conference on Artificial Neural Networks (ICANN 2002)*. J.R. Dorronsoro, eds. Madrid, Spain, August 27–30, 2002. Lecture Notes in Computer Science, volume 2415, pages 951–956.

Wang, K., Fang, H., Xiao, D., Zhu, X., He, M., Pan, X., et al. (2009). Converting redox signaling to apoptotic activities by stress-responsive regulators HSF1 and NRF2 in fenretinide treated cancer cells. PLoS One 4, e7538.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13, 1977–2000.

Wu, W., Liu, X., Xu, M., Peng, J.R., and Setiono, R. (2005). A hybrid SOM–SVM approach for the zebrafish gene expression analysis. Genomics Proteomics Bioinformatics 3, 84–93.

Xiao, L., Wang, K., Teng, Y., and Zhang, J. (2003). Component plane presentation integrated self-organizing map for microarray data analysis. FEBS Lett 538, 117–124.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., et al. (2005). Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434, 338–345.

Zheng, P.Z., Wang, K.K., Zhang, Q.Y., Huang, Q.H., Du, Y.Z., Zhang, Q.H., et al. (2005). Systems analysis of transcriptome and proteome in retinoic acid/arsenic trioxide-induced cell differentiation/apoptosis of promyelocytic leukemia. Proc Natl Acad Sci USA 102, 7653–7658.

Address correspondence to:
Prof. Kankan Wang
State Key Laboratory of Medical Genomics
Ruijin Hospital affiliated to Shanghai Jiao Tong University School of Medicine
Shanghai 200025, P.R. China

E-mail: kankanwang@shsmu.edu.cn

OR

Prof. Ji Zhang
Key Laboratory of Stem Cell Biology
Institute of Health Sciences
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
Shanghai 200025, P.R. China

E-mail: jizhang@sibs.ac.cn