ACS Combinatorial Science

RESEARCH ARTICLE

pubs.acs.org/acscombsci

Identification of Biologically Significant Genes from Combinatorial Microarray Data

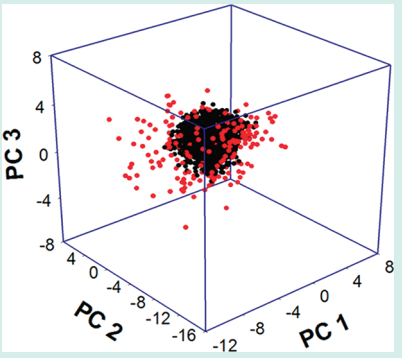
Chang Sun Kong,[†,‡] Jing Yu,[§] F. Chris Minion,[§] and Krishna Rajan*[,†,‡]

[†]Department of Materials Science and Engineering, Iowa State University, Ames, Iowa 50011, United States

[‡]Institute for Combinatorial Discovery, Iowa State University, Ames, Iowa 50011, United States

[§]Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, Iowa 50011, United States

ABSTRACT: High-throughput microarray technology has enabled the simultaneous measurement of the abundance of tens of thousands of gene-expression levels, opening up a new variety of opportunities in both basic and applied biological research. In the wealth of genomic data produced so far, the analysis of massive volume of data sets has become a challenging part of this innovative approach. In this study, a series of microarray experimental data from *Yersinia pestis* (*Y. pestis*), the etiologic agent of plague in humans, were analyzed to investigate the effect of the treatments with quorum-sensing signal molecules (autoinducer-2 and acyl-homoserine lactones) and mutation (Δ*ypeIR*, Δ*yspIR*, and Δ*luxS*) on the variation of gene-expression levels. The gene-expression data have been systematically analyzed to find potentially important genes for vaccine development by means of a coordinated use of statistical learning algorithms, that is, principal component analysis (PCA), gene shaving (GS), and self-organizing map (SOM). The results achieved from the respective methods, the lists of genes identified as differentially expressed, were combined to minimize the risk that might arise when using a single method. The commonly detected genes from multiple data mining methods, which reflect the linear/nonlinear dimensionality and similarity measure in gene-expression space, were taken into account as the most significant group. In conclusion, tens of potentially biologically significant genes were identified out of over 4000 genes of *Y. pestis*. The "active" genes discovered in this study will provide information on bacterial genetic targets important for the development of novel vaccines.



KEYWORDS: microarray, statistical learning, data mining, high throughput, *Yersinia pestis*, gene expression

INTRODUCTION

In experiments involving two-color microarrays, two sets of biological samples are prepared in parallel with different fluorescent labels (e.g., Cy3 and Cy5) and hybridized simultaneously to probes on the array, that is, one *treated* and the other *untreated* for the test and control, respectively (Figure 1). Variation of the gene-expression levels under any treatments such as mutation, temperature, chemical modification or time upon the respective samples is quantitatively measured by the relative intensity ratios of the optical signals from the probe spots emitting the two different fluorescent signals. This high-throughput microarray technology, which makes it feasible for researchers to investigate expression data from tens of thousands of genes simultaneously, is now accepted as a commonly used method in both basic and applied fields of biological research.

In the wealth of the genetic data produced so far, the analysis of massive volumes of array data sets has become the central challenge of this innovative strategy.[1−3] The essential tasks of microarray analysis involve (i) the grouping of *similar* genes in terms of the expression level (i.e., classification/clustering), (ii) the identification of significantly differentially expressed genes (i.e., outlier detection), and (iii) the investigation of the interaction among the genes as organized in biological pathways. From the biological information of some known genes, the function of unknown genes can often be inferred through the so-called functional classification of gene expression. A group of differentially expressed genes can be identified by sorting out genes by either the relative distance measure in gene expression space or by using any, rather empirical, threshold cutoff value. The interactions among genes can be statistically investigated by tracking the variation of expression levels of the potentially relevant genes.

Statistical learning algorithms are an essential tool for the identification of differentially expressed genes from microarray data. However, the results achieved from a single method might be misleading, resulting in the wrong conclusion since different methods often return rather different results.[4−6] Univariate-based statistical methods, e.g. ordinary t-statistic, need to be modified for the comparative analysis of the microarray data with a multivariate nature. When focusing on one specific statistical learning method, it is assumed that one knows the correct metric of the manifold structure on which the data exist. We relax that assumption by rationally merging multiple algorithms and take advantage of both linear and nonlinear metric of data manifold. In this manner, we acknowledge that the data streams coming

Received: July 1, 2011
Revised: July 21, 2011
Published: August 10, 2011

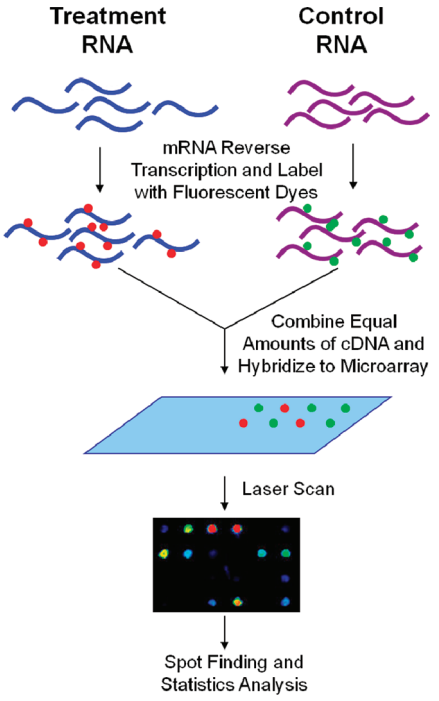


Figure 1. Schematic of the microarray process. The process begins with the construction of the array by spotting gene-unique oligonucleotides (the probes) to a chemically modified glass substrate. Targets are generated from control and experimental samples by extracting the RNA, converting to cDNA and labeling the two targets with different fluorescent dyes (e.g., Cy3 and Cy5). The two differentially labeled targets preparations are mixed and hybridized to the array. Following washing, the array is scanned using a laser scanner, the signal intensity for each spot is quantified, and the data is analyzed statistically for differentially expressed genes.

Table 1. Design of Microarray Experiments (Controlled Temperature 37°C)

signal studies (wild type)	mutant studies (OD 1.0)
control vs 3 signals (AI-2 and two AHLs)	wild type vs triple mutants (Δ*luxS*, Δ*ypeIR*, Δ*yspIR*)
control vs AI-2	wild type vs Δ*luxS* mutant
control vs two AHLs	

	QS signal molecule analysis			Mutant analysis	
	Array 1	Array 2	Array 3	Array 4	Array 5
Gene 1	$G_{1,1}$	$G_{1,2}$	$G_{1,3}$	$G_{1,4}$	$G_{1,5}$
Gene 2	$G_{2,1}$	$G_{2,2}$	$G_{2,3}$	$G_{2,4}$	$G_{2,5}$
Gene 3	$G_{3,1}$	$G_{3,2}$	$G_{3,3}$	$G_{3,4}$	$G_{3,5}$
⋮	⋮	⋮			
Gene *m*	$G_{m,1}$	$G_{m,2}$	$G_{m,3}$	$G_{m,4}$	$G_{m,5}$

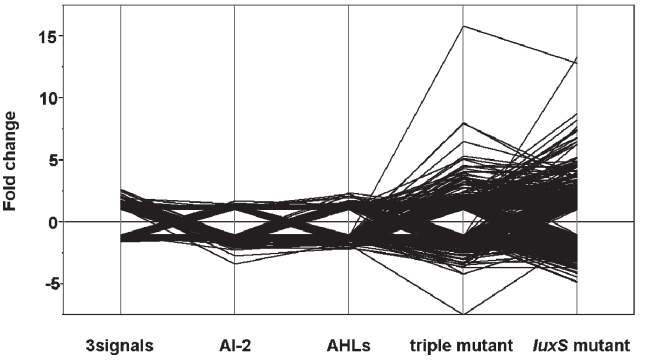


Figure 2. Schematic of microarray data table (G_{ij} denote gene-expression levels) and the corresponding parallel plot of the fold change of gene expressions across five different samples. Each straight line shows the variation of the expression level of individual genes under the different microarray conditions. The positive and negative value of fold change corresponds to up-regulated and down-regulated gene expressions, respectively. For the definition of each sample, see Table 1.

from different types of genes may not necessarily be in a single manifold structure. Jeffery et al.[5] investigated the lists of differentially expressed genes from 9 different microarray data by using 10 different statistical methods. They found out that all the dissimilar methods distinguish almost entirely different gene list, depending on the data set used. At present, the single-most efficient procedure for combining the outcomes from multiple algorithms has not been properly formulated yet. The results achieved from diverse analysis techniques would require the use of appropriate procedure for combining them. As shown in most of the published literature, the validity of a newly suggested method (or algorithm) to extract knowledge from microarray data can only be demonstrated by verifying with the training of already-known data resource, hoping that the method can be equally useful for a new unknown system. The work by Famili et al.[4] points out that the use of multiple statistical methods (i.e., rank products, significance analysis of microarray, and *t* test) could lead to better results than any of single method by equally weighting the outcome from each method. In fact, it becomes more critical aspect when new biological data, in which the available information is limited, are explored for practical application.

In this paper, we show a biologically significant (i.e., differentially expressed) gene list identified from microarray data by using a rational integration of multiple statistical learning algorithms which take into account more general similarity metric of high dimensional gene-expression space. That is, gene-expression data have been systematically analyzed by means of unsupervised multivariate data mining strategies, that is, principal component analysis (PCA), gene shaving (GS), and self-organizing map (SOM). Microarray experiments of *Yersinia pestis* (*Y. pestis*), a bubonic and pneumonic plague bacterium, were performed to investigate the effect of the treatments with quorum-sensing-signal molecules, autoinducer-2 (AI-2) and acyl-homoserine lactones (AHLs), and the mutations Δ*ypeIR*, Δ*yspIR*, and Δ*luxS* on gene-expression levels. All of these five individual treatments were implemented at the controlled temperature of 37 °C. Out of over 4000 genes, tens of distinct genes were commonly identified using different analysis methods, and those genes can be considered as the most biologically significant genes for further investigation to develop novel vaccine materials. We suggest that the highly ranked "active" genes detected from the overlapping gene lists of multivariate data mining methods need to be studied with a priority. After the fundamental algorithm of each approach is described, the level of agreement among the lists of differentially expressed genes declared by the respective methods

dx.doi.org/10.1021/co200111u |*ACS Comb. Sci.* 2011, 13, 562–571

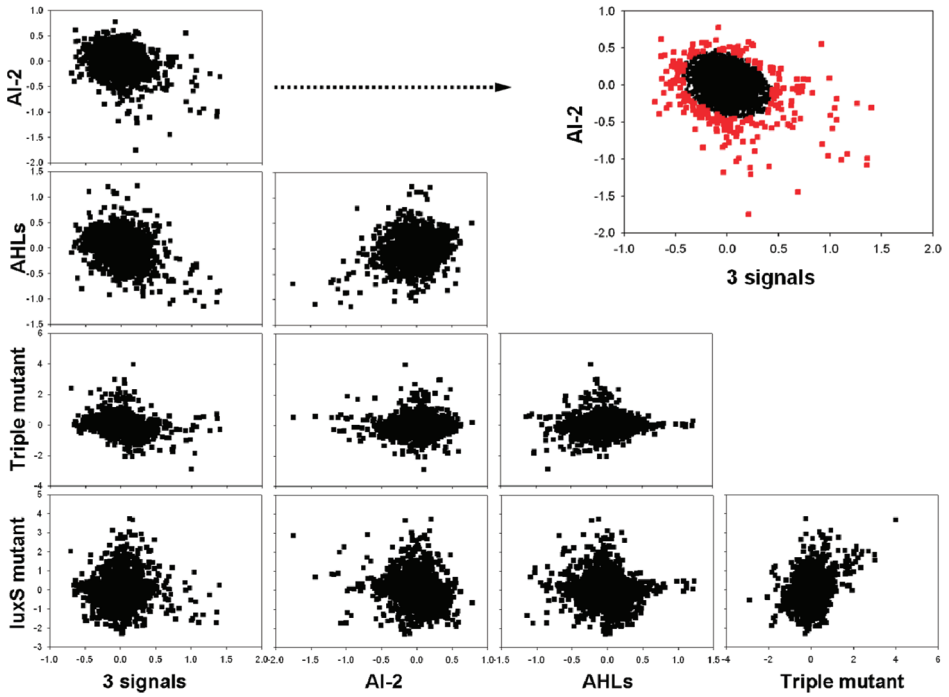


Figure 3. Pairwise comparison of the expression levels (fold changes are expressed in base 2 logarithmic scale) of 4,254 genes of *Y. pestis* characterized by the five different microarray conditions. Data points colored in red shown at the [AI-2 vs 3 signals] plot of the upper right panel are outlying genes detected by Mahalanobis distance measure (*F*-quantile = 0.99). Most of the data points are clustered together and some outlying data points are spread around the main cluster. Depending on the direction that the outliers are positioned, it can be determined that whether the genes are highly differentially expressed (up- or down-regulated) in only one or both of the microarray condition.

is taken into account along with the comparison to some known biological information.

EXPERIMENTAL PROCEDURES

Microarray Experimentation. *Bacterial Strains and Culture Conditions.* The strains *Y. pestis* CO92 ΔPgm (subsequently referred to as wild type), *Y. pestis* CO92 ΔPgm Δ*luxS*, and *Y. pestis* CO92 ΔPgm Δ*luxS* Δ*ypeIR* Δ*yspIR* were used in this study. The bacterial cells were grown in brain heart infusion broth plus 2.5 mM $CaCl_2$ at 37 °C. Cell growth was monitored on a Bausch and Lomb Spectronic 20 Spectrophotometer at wavelength 600 nm.

Experimental Design and Microarray Design. A schematic of the microarray process is shown in Figure 1. The theory for the two color array experiment is that two samples (control and experimental) can be compared simultaneously on the same array so that any spot-to-spot variation is encountered by both samples. This reduces the biological variation that is inherent to these types of experiments. In this analysis, there are five microarray comparisons, three are signals added-in studies (wild type vs added-in AI-2, wild type vs only AHLs added-in, and wild type vs added-in all three signals), and two mutant comparison studies (wild type vs ΔPgm Δ*luxS* and wild type vs ΔPgm Δ*luxS* Δ*ypeIR* Δ*yspIR*). For the three added-in signal studies, overnight wild type cultures were washed twice with PBS buffer to remove the endogenous quorum sensing signals, the cells were diluted 1:100 in fresh culture medium, and then the cells were incubated for 2 h at 37 °C. Purified signals were then added to the cultures at the following concentrations: AI-2 (500 nM final concentration), AHLs (5 μM *N*-(3-oxooctanoyl)-L-homoserine lactone and *N*-hexanoyl-DL-homoserine lactone) either as AI-2 alone, AHLs alone or all three signals. The control consisted of cells grown and treated under the same conditions without added signals. After 4 h of induction, all of the cultures were centrifuged and RNA prepared as described below.

For the mutant studies, overnight cultures of wild type, ΔPgm Δ*luxS* and ΔPgm Δ*luxS* Δ*ypeIR* Δ*yspIR* strains were diluted 1:100 in fresh culture medium and incubated at 37 °C until the cell density reached OD_{600} = 1.0, about 10 h. The cells from each culture were collected and RNA isolated as below. For each array comparison, six independent biological replicates were performed.

RNA samples from treated cultures were paired with six independent RNA samples from control cultures. For three arrays, the control RNA samples were labeled with Cy3 dye and the treatment RNA samples were labeled with Cy5 dye; the dyes were reversed for the other three arrays to account for any dye bias.

RNA Isolation, Target Generation, and Hybridization. All of the cell pellets were treated with RNAprotect Bacterial Reagent (Qiagen, Valencia, CA) and stored at −70 °C. RNA was extracted from frozen cell pellets using the RNeasy Mini kit (Qiagen). After extraction, the RNAs were treated with DNase I (Ambion, Austin, Texas) at 37 °C for 30 min to remove genomic DNA. The RNAs were purified and concentrated by Microcon Ultracel/YM 30.

Aminoallyl-labeled cDNAs were generated by reverse transcription reactions containing 10 μg of total RNA, 10 μg of random hexamer primers (Integrated DNA Technologies, Iowa City, IA), 1 × RT buffer, 10 mM dithiothreitol, 300 U Superscript III reverse transcriptase (Invitrogen, San Diego, CA), 500 μM final concentration each of dATP, dCTP, and dGTP, 100 μM final concentration dTTP, and 400 μM final concentration aminoallyl dUTP (Fermentas, Glen Burnie, MD). The reaction was incubated overnight at 46 °C. RNAs were hydrolyzed with

dx.doi.org/10.1021/co200111u |*ACS Comb. Sci.* 2011, 13, 562–571

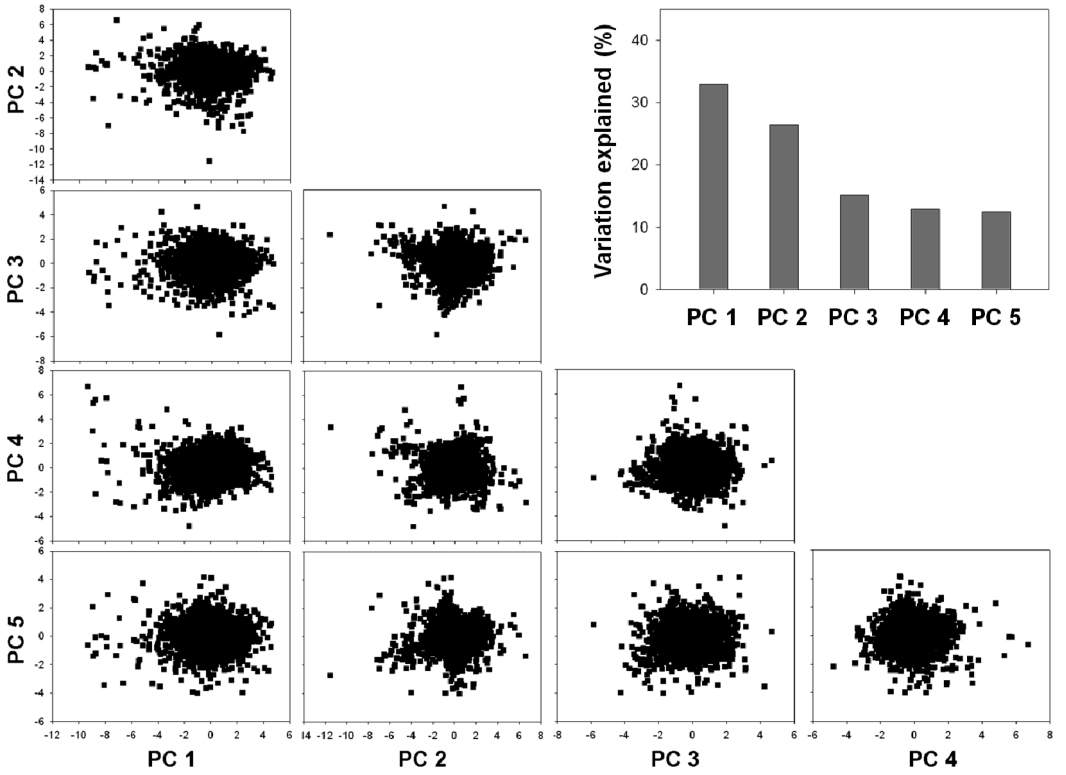


Figure 4. Pairwise comparison of the expression levels (fold changes are expressed in base 2 logarithmic scale) of 4,254 genes of *Y. pestis* projected to the principal component (PC) space. The variance of the gene-expression levels accounted for by each PC (in percentile) is shown in the bar chart on the upper right corner.

10 mM final concentration EDTA and 10 mM final concentration sodium hydroxide for 10 min at 65 °C, and the solutions were neutralized with 500 mM final concentration of HEPES buffer (pH 7.0). The cDNA targets were then purified using the UltraClean PCR Clean-Up kit (Mo Bio Laboratories, Carlsbad, CA). The following coupling of Cy3 or Cy5 dyes (GE Healthcare, Piscataway, NJ) to the purified aminoallyl-labeled cDNA was performed in a total 17.5 μL reaction volume including 10 μL of nuclease-free water, 1.5 μL of 100 mM sodium bicarbonate (pH 8.7), and 6 μL of dye. The reaction was incubated at room temperature in the dark for 3 h. The dye labeled cDNAs were then purified using the UltraClean PCR Clean-Up kit (MoBio Laboratories) and dye incorporation efficiency evaluated by ND-1000 NanoDrop spectrophotometry (NanoDrop Techologies, Wilmington, DE).

Microarray hybridization and postwashes were performed using a Lucidea Slidepro Hybridization Station (GE Healthcare, Piscataway, NJ). Corresponding equal amounts of dye labeled cDNA targets were mixed and dried by Thermo Scientific Savant DNA SpeedVac Concentrators. The mixed targets were suspended in 225 μL of long oligo hybridization solution (Corning), incubated at 95 °C for 5 min, centrifuged (10 000 × *g*, 4 min), and kept at room temperature until injection into the hybridization station. The hybridization lasted for 16 h at 42 °C and washes were performed with a series of buffers (2× saline-sodium citrate (SSC), 0.1% SDS; 1× SSC, and 0.1× SSC) by the hybridization station and dried by centrifugation at 1500 × *g* for 30 s.

Data Acquisition, Normalization, and Data Analysis. The hybridized arrays were scanned three times under varying laser power and photomultiplier tube values using a ScanArray HT scanner (Perkin-Elmer) to detect Cy3 and Cy5 signals. The images were quantified using the softWorRx Tracker analysis software package (Applied Precision, Inc., Issaquah, WA). Spot-specific mean signals were corrected for background, log transformed, and adjusted to a common median. The median of these adjusted-log-background-corrected signals across multiple scans was then computed for each spot to obtain one value for each combination of spot, array, and dye channel. A separate mixed

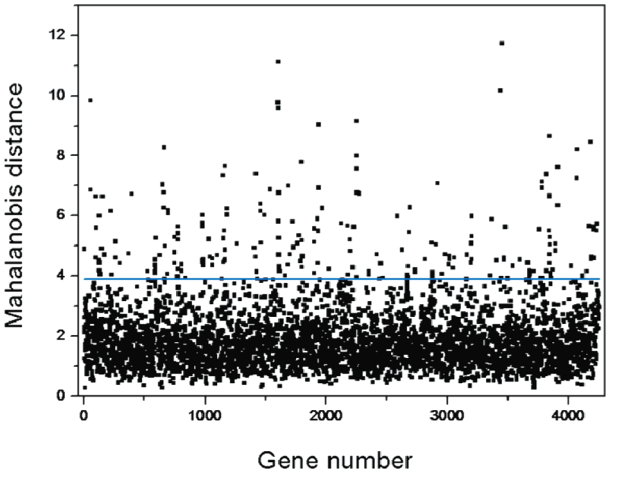


Figure 5. Outlier identification analysis plot. Data points above the blue guide line (threshold cutoff distance = 3.886869) indicate 129 outlying genes determined by the Mahalanobis-distance (*F*-quantile = 0.99). Among those identified genes, genes with FC < 1.3 or *p*-value > 0.05 in all five experiments have been removed from the list.

Table 2. Thirty Outlying Genes out of 129 Genes (*F*-Quantile = 0.99) Identified from the 4254 Gene Group by Mahalanobis Distance-Based Outlier Analysis

gene ID	description	function (class)
YPO3300	S-ribosylhomocysteinase, autoinducer-2 production protein	amino acid metabolism; cysteine and methionine metabolism
YPO1299	1-phosphofructokinase	energy metabolism; glycolysis
YPO3279	putative sigma 54 modulation protein	genetic information processing; translation
YPO1300 (*fruA*)	fructose-specific PTS system IIBC component	metabolism; environmental information processing; membrane transport; phosphotransferase system (PTS)
YPO1298	bifunctional fructose-specific PTS IIA/HPr protein	metabolism; environmental information processing; membrane transport; phosphotransferase system (PTS)
YPO1300	fructose-specific IIBC component; PTS system	transport/binding protein
YPO1993	putative dehydrogenase	unknown
YPO3711	maltoporin	transport/binding protein
YPO4080	alpha-amylase protein	degradation of polysaccharides
YPO0286	putative coproporphyrinogen III oxidase	biosynthesis of cofaacofactors, prosthetic groups and carriers
YPO3954	putative gluconate permease	transport/binding proteins
YPO1994	hypothetical protein	unknown
YPO1507	galactose-binding protein	transport/binding proteins
YPO0832	putative tagatose 6-phosphate kinase	degradation of carbon compounds
YPO3788	5-methyltetrahydropteroyltriglutamate−homocysteine methyltransferase	aspartate family biosynthesis
YPO1995	hypothetical protein	unknown
YPO3681	putative insecticidal toxin	cell processes−pathogenicity
YPO3953	putative gluconokinase	degradation of carbon compounds
YPO3643	major cold shock protein Cspa2	adaptions and atypical conditions
YPO2705	conserved hypothetical protein	unknown
YPO0276	putative LysR-family transcriptional regulator	broad regulatory functions
YPO3644	major cold shock protein Cspa1	adaptions and atypical conditions
YPO1222	outer membrane protein C, porin	cell envelop
YPO0284	orfY protein in hemin uptake locus	unknown
YPO1996	hypothetical protein	unknown
YPO1303	pH 6 antigen precursor (antigen 4) (adhesion)	surface polysaccharides, lipopolysaccharides and antigens
YPO2012	putative membrane protein	membranes, lipoproteins, and porins
YPO0003	aspartate−ammonia ligase	aspartate family biosynthesis
YPO3714	maltose-binding periplasmic protein precursor	transport/binding carbohydrates, organic acids and alcohols
malF	putative maltodextrin transport permease	transport/binding carbohydrates, organic acids and alcohols

The information on the gene function was achieved as described.[17] As the criteria of fold change (FC) and *p*-value, if a gene corresponds to either FC < 1.3 or *p*-value >0.05 under all five different experiments, then the gene was removed from the outlier list. Thus, all the genes up-regulated in at least one of the five experiments have been considered as significant ones.

linear model was constructed for each probe sequence using the normalized data.[7] The *t* tests for comparison between treatment and control for each probe were conducted. The *p*-values from these tests were converted to *q*-values using the method of Storey and Tibshirani.[8] The *q*-values were used to approximate the false discovery rate (FDR) for any given *p*-value as described by Benjamini and Hochberg.[9] Fold changes of the expression between treatment and control were estimated for each probe by taking the inverse natural log 2 of the estimated mean treatment difference. The GEO supergroup number for 5 microarray data is series GSE22850 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=xhcbrqscuokoclu&acc=GSE22850).

Multivariate Data Mining Algorithms. Suppose a microarray data set, $\boldsymbol{M}$, that consists of an $i \times j$ matrix, where i rows denote i genes (e.g., $i = 4254$ in this study) and j columns correspond to j different samples or conditions (e.g., $j = 5$ in this study). The individual gene is positioned in the j-dimensional space according to the corresponding expression level. That is, the coordinates of gene space represent the expression levels of the genes under various conditions. Each univariate microarray data set achieved from the five different conditions shown in Table 1 comprises a column of a 5-dimensional data matrix, which is referred to as a gene-expression profile. Then the algorithms detailed below are applied for the reduction of data dimensionality, the measure of the similarity of genes, and the detection of highly regulated genes. For the description of gene-expression levels, logarithmic fold change (i.e., the difference in logarithmic expressions of each sample) was used.

Principal Component Analysis (PCA). Principal component analysis (PCA) is a multivariate dimensionality reduction technique in which the dimension of original, usually correlated, variable space, $\mathbf{X} = \mathbf{X}(x_1, x_2, .., x_m) \in \mathscr{R}^m$ is transformed into a new

Table 3. Thirty Genes Identified from the 4254 Gene Group by Gene Shaving Method[a]

gene ID	description	function (class)
YPO1654	beta-D-galactosidase	degradation of carbon compounds
YPCD1.08c	hypothetical protein	NaN
YPO0158	siroheme synthase	heme and porphyrin biosynthesis
YPO3272	putative acetyltransferase	unknown
YPO0440	purine nucleoside phosphorylase	salvage of nucleosides and nucleotides
YPO3279	putative sigma 54 modulation protein	broad regulatory functions
YPO0285	conserved hypothetical protein	unknown
YPO3713	hypothetical protein	unknown
YPO3712	maltose/maltodextrin transport system ATP-binding protein	environmental information processing; membrane transport
YPO4080	periplasmic alpha-amylase precursor	carbohydrate metabolism; starch and sucrose metabolism
YPO3711	maltoporin	transport/binding carbohydrates, organic acids and alcohols
YPO1507	galactose-binding protein	environmental information processing; membrane transport; transporters
YPO3643	major cold shock protein CspA2	genetic information processing; transcription; transcription factors
YPO1299	1-phosphofructokinase	carbohydrate metabolism, fructose and mannose metabolism
YPO1298	bifunctional fructose-specific PTS IIA/HPr protein	carbohydrate metabolism; fructose and mannose metabolism; environmental information processing; membrane transport; transporters
YPO2180	bifunctional acetaldehyde-CoA/alcohol dehydrogenase	carbohydrate metabolism ; glycolysis/gluconeogenesis
YPO3954	putative gluconate permease	transport/binding carbohydrates, organic acids and alcohols
ompC	outer membrane porin protein C	environmental information processing ; signal transduction ; two-component system
YPO1994	hypothetical protein	unknown
YPO4012	two-component system response regulator	environmental information processing; signal transduction; two-component system
YPO0410	putative ABC transporter permease protein	environmental information processing; membrane transport; transporters
YPO0436	deoxyribose-phosphate aldolase	carbohydrate metabolism; pentose phosphate pathway
YPO1138	galactose-1-phosphate uridylyltransferase	carbohydrate metabolism; galactose metabolism
YPO3024	probable *N*-acetylneuraminate lyase	carbohydrate metabolism; amino sugar and nucleotide sugar metabolism
YPO1139	UDP-galactose-4-epimerase	carbohydrate metabolism; galactose metabolism; amino sugar and nucleotide sugar metabolism
YPO0407	autoinducer-2 modifying protein LsrG	unknown
YPO2012	putative membrane protein	membranes, lipoproteins and porins
YPO0409	putative periplasmic solute-binding protein	environmental information processing; membrane transport; transporters
YPO1137	galactokinase	carbohydrate metabolism; galactose metabolism; amino sugar and nucleotide sugar metabolism
YPO3300	autoinducer-2 production protein	amino acid metabolism; cysteine and methionine metabolism

[a] The information on the gene function was achieved as described.[17]

latent variable space referred to as principal components (PCs), $\mathbf{Y} = \mathbf{Y}(y_1, y_2, .., y_n) \in \mathscr{R}^n$, that is, $\mathscr{R}^m \rightarrow \mathscr{R}^n$ (where $m \geq n$), as the independent linear combination of original variables. The data matrix **X** is decomposed into two matrices **U** and **V**, which are orthogonal each other. That is,

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} \tag{1}$$

Where, **S** is the diagonal matrix of the eigenvalues. The product **US** and **V** are called the score matrix and loading matrix, respectively. The eigenvectors of the covariance matrix consist of the PCs. The first PC accounts for the maximum variance (eigenvalue) in the original data set. The second PC is orthogonal (i.e., uncorrelated) to the first one and accounts for most of the remaining variance. Thus, the *n*th PC is orthogonal to all others and has the *n*th largest variance in the set of PCs. Transforming the original to this new coordinate system in the high dimension, PCA seeks the operational benefits for the visualization and recognition of the major pattern of data structure in the reduced dimensional PC space with minimum information loss. Recently, some results have shown that PCA can be effectively used as a clustering tool for microarray data by combining with additional criteria of the threshold value for the determination of the genes differentially expressed but is not impressive by itself.[10,11]

Mahalanobis Distance-Based Outlier Detection. In the multivariate data, covariance matrix is taken into account and the differentially expressed genes that resolve a certain distance from the average value are selected. One of the classical distance metrics is Mahalanobis distance. The Mahalanobis distance between genes $\boldsymbol{G_1}$ and $\boldsymbol{G_2}$ (both are vectors) is a distance measure which reflects the similarity between genes as the proximity of them in the expression vector space. It is defined as

$$\mathbf{d}(\boldsymbol{G}_1, \boldsymbol{G}_2) = \sqrt{(\boldsymbol{G}_1 - \boldsymbol{G}_2)^T \boldsymbol{C}^{-1} (\boldsymbol{G}_1 - \boldsymbol{G}_2)} \tag{2}$$

where *C* is the variance−covariance matrix, which is defined as

$$C = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{3}$$

dx.doi.org/10.1021/co200111u |*ACS Comb. Sci.* 2011, 13, 562–571

where σ_1^2 and σ_2^2 are the variances and $\rho_{12}\sigma_1\sigma_2$ is the covariance, respectively. If the covariance matrix is a diagonal unit matrix, then Mahalanobis distance is identical to Euclidean distance. This Mahalanobis distance can be calculated in either original data space or principal component (PC) space. The values are in fact equal when all PCs are taken into account.[12]

Gene Shaving (GS). Gene shaving (GS) is a gene clustering algorithm in which the PCA technique is iteratively applied to search subgroups of genes with higher similarity than others from the original gene data set. This GS method was first proposed by Hastie et al.[13] Looking for a group of genes, S_t (where t is the number of genes of the cluster) with highest variance, that is, eigen genes, the genes with lowest correlation with the eigen genes are cut out (i.e., shaved off) through an iterative multistep elimination procedure. The size of the cluster, t (i.e., the number of eigen genes in a cluster), is determined through the gap statistics using the variances for subsets of the genes which consist of a $k \times l$ gene-expression (G_{ij}) matrix.

$$V_W = \frac{1}{l}\sum_{j=1}^{l}\left[\frac{1}{t}\sum_{i \in S_t}(G_{ij} - \bar{G}_j)^2\right] \quad (4)$$

$$V_B = \frac{1}{l}\sum_{j=1}^{l}(\bar{G}_j - \bar{G})^2 \quad (5)$$

$$V_T = \frac{1}{tl}\sum_{i \in S_t}\sum_{j=1}^{l}(G_{ij} - \bar{G})^2 \quad (6)$$

where V_W, V_B, and V_T ($= V_W + V_B$) denote the within, between, and total variances for gene clusters, respectively. The ratio of V_B/V_W (or V_B/V_T) is used as the measure to determine the size, t, of a gene cluster S_t.

Self-Organizing Map (SOM). Self-organizing map (SOM), or Kohonen map, is a nonparametric learning algorithm in which the original high-dimensional data space, $\mathbf{X} = \mathbf{X}(x_1, x_2, .., x_m) \in \mathscr{R}^m$, is mapped preserving the topology of the data structure on the prespecified low-dimensional, 2-D planar geometry representations, $\mathbf{X}' = \mathbf{X}'(x_1', x_2', .., x_m') \in \mathscr{R}^2$, namely $\mathscr{R}^m \rightarrow \mathscr{R}^2$. The prototype reference vectors are trained by input data vectors and modified according to their similarity to the input vector, that is, "self-organized". In brief, SOM can be considered as a nonlinear version of a dimensionality reduction technique such as PCA for linear data space.[14,15] However, the two algorithms are dissimilar in that SOM finds the locally based *principal* pattern, which takes into account only the neighboring data, whereas the latter seeks the *principal* directions, which considers the entire data structure. The distance between original input vector $\vec{x}$ and output parametric vector $\vec{m}_c$ is

$$||\vec{x} - \vec{m}_c|| = \min_i\{||\vec{x} - \vec{m}_i||\} \quad (7)$$

where $\vec{x}$ and $\vec{m}_c$ denote input and output vectors, respectively.

Similarity measure in the SOM is based on the Euclidian distance. In this paper, the batch learning method that is computationally effective was used for the outlier detection. The difference of this simpler version of the SOM algorithm is to use the input vectors at the same time. For the analysis, a SOM toolbox for Matlab environment was used.[16]

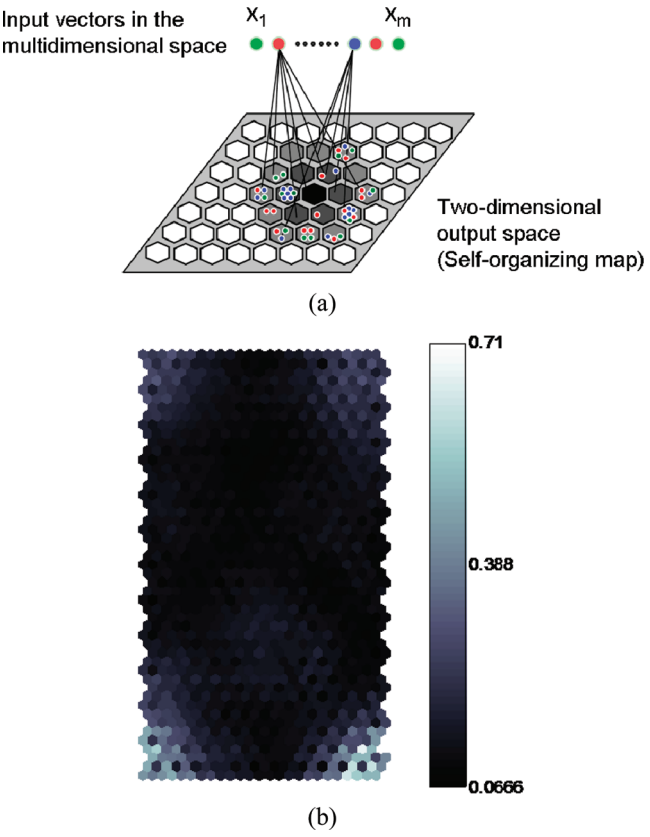


Figure 6. Self-organizing map (SOM) produced by the batch-type learning algorithm: (a) a schematic of SOM constructed in the batch-type algorithm; the data in the multidimensional original space are compressed into the reduced dimension (2-D) of predefined grid structure by topologically mapping with Euclidean distance between data points as the similarity measure, and (b) unified-distance matrix (U-matrix); note that, in the U-matrix, there exist additional grids between neighboring grid units of the SOM with the map size of (25×13) to show the distance between the grids. In the U-matrix, it is shown that two bright regions of the left and right bottom of the map that indicate the clusters of outlying genes that are far apart from other genes.

RESULTS AND DISCUSSION

The analysis of transcription-level data to identify differentially expressed genes across the individual microarray experiments has long been a statistical and computational challenge. Figure 2 shows the variation of gene-expression levels in terms of the fold change achieved across five different samples. The fold change (FC) of gene-expression levels varies apparently independently in different samples along either positive (i.e., up-regulated) or negative (i.e., down-regulated) direction. That is, a gene that is up-regulated (i.e., FC $\gtrsim 2$) in one experimental condition might be down-regulated in other conditions. Thus, the regulation behavior of genes measured under various array conditions cannot be directly analyzed with thresholding by a univariate-based approach. In fact, rather different results are often achieved using different analytical methods for the expression data with multivariate nature. In this study, the gene-expression data obtained from five microarray experiments were explored by using three multivariate data mining tools, that is, PCA—Mahalanobis distance measure, GS, and SOM. The main difference among them is originated from that of the algorithms for clustering genes in which the original input space is mapped onto

Table 4. Thirty Genes Identified from the 4254 Gene Group by Self-Organizing Map[a]

gene ID	description	function (class)
YPO0625	hypothetical protein	unknown
YPO0821	hypothetical protein	unknown
YPO0833	putative phosphosugar isomerase	degradation of carbon compounds
agaY	tagatose-bisphosphate aldolase	carbohydrate metabolism
YPO0823	putative exported protein	cell envelop
YPO0626	hypothetical protein	unknown
YPO0624	putative membrane protein	cell envelop
malG	maltose transport system permease protein	transport/binding proteins
YPO0822	putative exported protein	cell envelop
YPO3715	maltose transporter membrane protein	environmental information processing; membrane transport; transporters
YPO0623	putative aminotransferase	unknown
YPO3716	maltose transport system permease protein MalG	transport/binding proteins
YPO3712	maltose/maltodextrin transport ATP-binding protein	transport/binding proteins
YPO0844	putative aldolase	degradation of carbon compounds
malF	putative maltodextrin transport permeaase	transport/binding proteins
YPO3200	putative maltodextrin glucosidase	degradation of carbon compounds
YPO0845	ThiJ/PfpI-family thiamine biogenesis protein	biosynthesis of cofactors, prosthetic groups and carriers
YPO3714	maltose-binding periplasmic protein precursor	transport/binding proteins
malK	maltose/maltodextrin transport ATP-binding protein	transport/binding proteins
YPO0325	single-strand binding protein	DNA replication, restriction/modification, recombination and repair
YPO0931	S-adenosylmethionine synthetase	central intermediary metabolism
YPO3710	maltose operon periplasmic protein	transport/binding proteins
YPO3681	putative insecticidal toxin	pathogenicity
YPO1858	putative exported protein	cell envelop
YPO0324	excinuclease ABC subunit A	DNA replication, restriction/modification, recombination and repair
YPO3788	5-methyltetrahydropteroyltriglutamate−homocysteine methyltransferase	aspartate family biosynthesis
YPO0832	putative tagatose 6-phosphate kinase	degradation of carbon compounds
YPO3711	maltoporin	transport/binding proteins
YPO4080	alpha-amylase protein	degradation of polysaccharides
YPO3300	autoinducer-2 production protein	unknown

[a] The information on the gene function was achieved as described.[17]

the reduced-dimensional output space and the distance among genes is measured. Figure 3 shows the distribution of gene-expression data points that correspond to each gene characterized by the expression level measured from different microarray experiments. The regulation behavior of differentially expressed genes can be readily identified by the outlying position of the corresponding genes from other clustered genes in the expression space.

Principal Component Analysis. One of the inherent assumptions of the PCA approach is that most of the information of the data structure is in the first few PCs, and the information carried by the rest of the PCs is less significant. Recently, Yeung and Ruzzo showed that simple coordinate transformation of original microarray data space to the corresponding PC space may not improve the performance of the microarray analysis.[11] The use of the first few PCs (usually two or three PCs) to capture the major features of microarray data is often not effective due to the independency of the experiments. That is, the transformation of original data space into the orthogonal PC space for the purpose of dimensionality reduction results in inevitable information loss without any significant benefits in the interpretation. One reason might be that the individual sample or experiment (each column of data matrix) is already designed to be independent of other samples or experiments.

Figure 4 shows the pairwise plots of the distribution of the genes in the PC space. There is no apparent change in the data structure, compared to Figure 3. The variation of the data explained by each PC (a bar chart on the upper right panel) indicates that even the variation accounted for by the fifth PC is not negligible. Also, considering that one of the assumptions of PCA is the Gaussian distribution of the data, it is shown that the PC projection of microarray data fails to find any useful information. When investigating differentially expressed genes comparing different treatment conditions, one will observe that a gene is expressed as highly up- or down-regulated (meeting the criterion of p-value) under one treatment but not other treatments. In this case, designating the differentially expressed genes as an outcome of combining the results of multiple conditions is an important aspect of the microarray analysis. We first implemented the analysis without regarding the FC and p-value cutoff, and then after the outliers are detected, the criteria of FC > 1.3 and p-value <0.05 were employed.

Outlier Detection by Mahalanobis Distance. In both the original input space and the corresponding PC space, the distance between genes was measured by using Mahalanobis distance and the list of outlying genes identified by this distance metric (F-quantile = 0.99) was compared. The Mahalanobis

Table 5. Ranking of Differentially Expressed Genes Commonly Detected from Multiple Analysis Methods[a]

group A	group B, C, D		group E, F, G (PCA only)		group E, F, G (gene shaving only)		group E, F, G (SOM only)	
gene ID	gene ID		gene ID		gene ID		gene ID	
YPO3300	YPO3279	YPO3788	YPO1222	YPO2705	YPO1654	YPO4012	YPO0625	YPO3716
YPO3711	YPO1298	YPO3681	*fruA*	YPO0276	YPCD1.08c	YPO0410	YPO0821	YPO3712
	YPO4080	YPO3643	YPO1300	YPO3644	YPO0158	YPO0436	YPO0833	YPO0844
	YPO3954	YPO2012	YPO1993	YPO0284	YPO3272	YPO1138	*agaY*	YPO3200
	YPO1994	*malF*	YPO0286	YPO1996	YPO0440	YPO3024	YPO0823	YPO0845
	YPO1507	YPO1299	YPO1995	YPO1303	YPO0285	YPO1139	YPO0626	*malK*
	YPO0832	YPO3714	YPO3953	YPO0003	YPO3713	YPO0407	YPO0624	YPO0325
					YPO3712	YPO0409	*malG*	YPO0931
					YPO2180	YPO1137	YPO0822	YPO3710
					ompC		YPO3715	YPO1858
							YPO0623	YPO0324

[a] Group A (the highest significant gene group) is a list of genes detected by all three methods. Group B, C, D (the second gene group) is a list of genes detected by two different methods. Group E, F, G (the third gene group) is a list of genes detected by only one method.

distance of all the genes in the original space and the equivalent distance measure in the PC space (Euclidean norm) provide the identical list of outlying genes. Figure 5 represents the outlying genes determined by the cutoff line (based on the *F*-quantile of 0.99). The partial list of 129 genes is shown in Table 2. As mentioned, since the results are the same in both cases, only one table is provided.

Gene Shaving. In the GS analysis as a nonparametric clustering method of orthogonal subgroups, a gene group with the largest PC, which is referred to as eigen genes, is selected. Then a subgroup of the genes having the lowest correlation with the eigen genes is sought as the distinctive gene cluster. This filtering process is iterated with the rest of the genes. Biologically significant genes can be detected during this GS process. Although this method uses the PCA as the main clustering algorithm of determining the major pattern of data, in our study, the criterion for the detection of outlying data is applied as a way to find the differentially expressed genes. The list of active genes identified by the GS method is shown in Table 3.

Self-Organizing Map. SOM projects the input data space onto the two-dimensional grid structure. Figure 6a schematically illustrates the clustering implemented by SOM. The map grid is the fixed matrix structure on which the original input vectors are sorted out by iterative training. The size of the map was determined empirically. That is, the map size was approximately determined as $5(x)^{1/2}$, where, x denotes the number of genes, x = 4254 for the data set we used. Thus, the map size is 325; the dimensional ratio (= 25 × 13) of the map was determined as the ratio of the first and second highest eigenvalues of the covariance matrix. The U-matrix of Figure 6b shows the distances between map grid units facilitating the identification of dissimilar groups on the map. In the grid structure of the U-matrix, there are additional grids for the representation of the distance between map grids. Although each unit of grids is colored with one color, the density of data points in individual grids is different from each other. The distinct gene group verified from the classification of the SOM in Figure 6b, those included in bright-color grids of the map, is listed in Table 4.

The Genes Identified as Highly Regulated Genes from the Three Methods. The results from three different analysis methods, that is, Mahalanobis distance-based outlier analysis, GS, and SOM, were summarized in Table 5. Among them, the commonly detected genes from multiple, more than two, methods can be considered as the more important genes. There were only two genes that were identified as outlying genes from all three methods: YPO3300 (AI-2 production protein) and YPO3711 (outer membrane protein). Fourteen genes, that is, YPO3279, YPO1298, YPO4080, YPO3954, YPO1994, YPO1507, YPO0832, YPO3788, YPO3681, YPO3643, YPO2012, *malF*, YPO1299, and YPO3714 were selected as common by two different methods. Thus, a total of sixteen genes can be considered as the most significant gene group. Among the genes identified as highly regulated, YPO3300 (in the group A) and YPO1994 (in the group B,C,D) have been identified as the genes leading the transcriptional induction in human plasma.[18] It is also seen that many genes in the group A, B, C, and D are related to membrane transport (e.g., YPO1507, YPO1298, YPO2012, YPO3711, YPO3714, and *malF*). As a separate note, among the three outlier detection methods, self-organizing maps (SOM) showed the highest detection efficiency for up-regulated genes which meet the criteria of fold change (>1.3) and *p*-value (<0.05). Also, many of the coregulated genes (mainly up-regulated genes under both triple mutants and Δ*luxS* mutant conditions) were automatically identified by SOM.

CONCLUSIONS

In this paper, we investigated the gene-expression data of *Y. pestis* achieved from oligonucleotide microarrays to identify genes that are significantly differentially expressed during quorum sensing that may be useful as potential vaccine candidates. The combined use of principal component analysis, self-organizing maps, gene shaving, and outlier analysis algorithms shown in this paper have facilitated us to screen out biologically significant genes from combinatorial microarray data as the starting point worthy of further study. Under the condition of limited available biological information, we suggest that the proper outcome can best be achieved by focusing with a priority on the genes commonly indicated by the statistical learning algorithms in which both linear and nonlinear metric of gene-expression space are considered.

AUTHOR INFORMATION

Corresponding Author

*E-mail: krajan@iastate.edu. Phone: 515-294-2670. Fax: 515-294-5444.

■ ACKNOWLEDGMENT

The authors would like to acknowledge financial support from the ONR-MURI Award 429 (NN00014-06-1-1176). K.R. will also like to acknowledge support through the Wilkinson Professorship for Interdisciplinary Engineering.

■ REFERENCES

(1) Allam, A. R.; Shyambabu, M.; Srinubabu, G. Microarray analysis of differentially expressed genes between diabetes vs healthy. *J. Proteomics Bioinf.* **2008**, *S1*, S055–S084.

(2) Bassett, D. E. J.; Eisen, M. B.; Boguski, M. S. Gene expression informatics. *Nat. Genet. Suppl.* **1999**, *21*, 51–55.

(3) Loguinov, A. V.; Mian, I. S.; Vulpe, C. D. Exploratory differential gene expression analysis in microarray experiments with no or limited replication. *Genome Biol.* **2004**, *5*, R18.

(4) Famili, F.; Phan, S.; Liu, Z.; Pan, Y.; Djebbari, A.; Lenferink, A.; O'Connor, M. Discovering informative genes from gene expression data: a multi-strategy approach. In *2nd Workshop in Data Mining in Functional Genomics and Proteomics*; DMFGP'07, ECML/PKDD 2007, Warsaw, Poland, 2007.

(5) Jeffery, I. B.; Higgins, D. G.; Culhane, A. C. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinf.* **2006**, *7*, 359.

(6) Pan, W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **2002**, *18*, 546–554.

(7) Wolfinger, R. D.; Gibson, G.; Wolfinger, E. D.; Bennett, L.; Hamadeh, H.; Bushel, P.; Afshari, C.; Paules, R. S. Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* **2001**, *8*, 625–637.

(8) Storey, J. D.; Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9440–9445.

(9) Benjamini, Y.; Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **2000**, *25*, 60–83.

(10) Roden, J. C.; King, B. W.; Trout, D.; Mortazavi, A.; Wold, B. J.; Hart, C. E. Mining gene expression data by interpreting principal components. *BMC Bioinf.* **2006**, *7*, 194.

(11) Yeung, K. Y.; Ruzzo, W. L. Principal component analysis for clustering gene expression data. *Bioinformatics* **2001**, *17*, 763–774.

(12) De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L. The mahalanobis distance. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18.

(13) Hastie, T.; Tibshirani, R.; Eisen, M. B.; Alizadeh, A.; Levy, R.; Staudt, L.; Chan, W. C.; Botstein, D.; Brown, P. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* **2000**, *1*, 1–21.

(14) Kohonen, T. *Self-Organizing Maps*, 3rd ed.; Springer: New York, 2001.

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

(16) Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. SOM toolbox for Matlab5. http://www.cis.hut.fi/projects/somtoolbox/ (accessed Feb 4, 2007).

(17) GenomeNet Database Resources. http://www.genome.jp/. Kyoto University Bioinformatics Center, 1995.

(18) Chauvaux, S.; Rosso, M.-L.; Frangeul, L.; Lacroix, C.; Labarre, L.; Schiavo, A.; Marceau, M.; Dillies, M.-A.; Foulon, J.; Coppée, J.-Y.; Médigue, C.; Simonet, M.; Carniel, E. Transcriptome analysis of *Yersinia pestis* in human plasma: an approach for discovering bacterial genes involved in septicaemic plague. *Microbiology* **2007**, *153*, 3112–3123.