[Frontiers in Bioscience 14, 4058-4070, January 1, 2009]

A strategy for meta-analysis of short time series microarray datasets

Ruping Sun[1], Xuping Fu[1], Fenghua Guo[2], Zhaorong Ma[1], Chris Goulbourne[3], Mei Jiang[1], Yao Li[1], Yi Xie[1], Yumin Mao[1]

[1]State Key Laboratory of Genetic Engineering, Institute of Genetics, School of Life Science, Fudan University, Shanghai 200433 PR, China, [2]Shanghai BioStar Genechip Institute, Shanghai 200092, P.R. China, and [3]Department of Biological and Biomedical Sciences, University of Durham, Durham, UK

TABLE OF CONTENTS

1. ABSTRACT

Many time series microarray experiments have relatively short (less than ten) time points and lack in repeats, weakening the confidence of results. Combining the microarray data from different groups may improve the statistical power of detecting differentially expressed genes. However, few efforts have been taken to combine or compare the time-course array datasets generated by independent groups. Here we demonstrated a suitable strategy for meta-analysis of short time series microarray datasets and implemented this strategy on four published heat shock microarray datasets of *Saccharomyces Cerevisiae*. We first assessed the significance of each gene in each datasets based on area calculation and the null distribution of the areas. Then the similarity of significance values across datasets was assessed with meta-analysis methods, yielding a set of transient heat shock stress sensitive genes. Following correlation calculation helped us to combine the transformed data at the same time points of each gene. Further bioinformatic investigation showed the significance of our strategy, and also indicated some interesting features of regulatory systems in *S. cerevisiae* during transient heat stress.

2. INTRODUCTION

To get a comprehensive view of the transcriptome in different organisms at different stages, numerous microarray experiments have been carried out during the past few years. Accordingly, more and more related microarray datasets are publicly available. With the accumulation of these datasets submitted by independent groups, a corresponding step in analyzing the expression data is to combine the results of these studies, which has been called meta-analysis generally. Although combining array data from different groups or platforms remains a challenge, it is still feasible and necessary for avoiding artifacts of individual studies (1). Until now, several reported meta-analysis of microarray data have focused on cancer research to find commonly dysregulated genes in a particular cancer (2-6) or other diseases (7, 8), or to identify common transcriptional profiles of diverse cancer microarray datasets (9). From these studies, a number of amazing robust statistical methods have been developed, which can effectively combine and compare the related expression datasets generated by different groups and can even integrate datasets from different array platforms, oligonucleotide arrays and cDNA arrays (2).

Table 1. Datasets included in this meta-analysis

Data	Reference Number	Chip type	Strain	Reference Sample	Samples (Time Points)
1	18	cDNA	Wild Type	Mixed samples	25°C [1] (0) 37°C [1] (5/15/30/45/60)
2	18	cDNA	Wild Type	Mixed samples	25°C [1] (0) 37°C [1] (5/10/15/20/30/40/50/60)
3	19	cDNA	Wild Type	25°C [1]	37°C [1] (5/10/15/20/30/40/60/80)
4	20	cDNA	Wild Type	30°C [1]	25°C [1] (0) 37°C [1] (0/5/15/30/60)
5	21	oligo	Wild Type	25°C [1]	37°C [1] (15/30/45/60/120)

[1] The temperatures□25°C and 37°C represent the cell culture conditions of samples and reference pools in heat shock experiments. Since the reference pools in individual studies were not consistent, zero transformation was performed (see Materials and Methods).

Microarray datasets can be divided into two classes: static and time series data. Static expression experiments, in which a snapshot of gene expression levels is taken, are often used in detecting expression levels of tumour cells from different cancer types. In time series microarray experiments, a temporal process is measured (for example, response to environmental conditions or the cell cycle). Time series array datasets provide a precise view of the instantaneous expression level at a particular time point and also exhibit the gene expression changes along time. So it has become a useful approach for exploring biological processes. A number of statistical methods have been produced to identify the differentially expressed genes in time-course microarray data (10-13). However, many time series microarray experiments contain relatively short time points (less than ten) and lack in repeats, weakening the confidence of results. In individual short time series microarray experiments, there would be some error brought by technical and biological sources of variability that may lead to wrong cognizance of the hidden temporal genetic response. Combining the short time series microarray datasets from independent groups may improve the statistical power of detecting differentially expressed genes. However, few efforts have been taken to combine and compare the short time-course expression datasets generated by independent groups. Combination of time-course microarray datasets is hindered by biological and experimental inconsistencies such as differences in sampling rates, variations in the timing of biological processes, and the lack of repeats (14). Whereas, if similar time-course expression experiments are performed under the same condition, combination of these time-course array datasets may be feasible and lead to a more robust result.

It has been pointed out that for short time series microarray experiments, two error based methods (such as area calculation) are better than the cubic spline fitting based methods which are appropriate for relatively long experiments (15). In this study, we designed an area calculation and permutation based approach for meta-analysis of short time series microarray studies, and illustrated its application on four publicly available time-course array datasets pertaining to transient heat shock response in wild *S. cerevisiae* (from 25°C to 37°C). Firstly, we assessed the significance of each gene in each dataset based on area calculation and the null distribution of the areas. Then the similarity of significance values across datasets was assessed with meta-analysis methods, helping us to identify a set of commonly transient heat stress sensitive genes (CTHS genes). Subsequently, after we transformed the data to a comparable form, correlation coefficients were calculated to select the genes with high time point correlation among different datasets. Lastly, we combined the time-course data of these handpicked CTHS genes and clustered these genes according to their time-dependent characteristics.

Transient heat shock microarray datasets of *S. cerevisiae* have been chosen as a case study for this meta-analysis. Heat shock is a kind of environmental stress that alters gene expression in prokaryotic and eukaryotic cells. The response to heat shock is characterized by a rapid induction of a conserved group of heat shock proteins (HSPs) (16). In *S. cerevisiae*, this response involves two regulatory systems: the heat shock transcription factor (Hsf1) and the Msn2 and Msn4 (Msn2/4) transcription factors. However, the contribution of each system independently is just beginning to emerge (17). Using available literatures and databases, we validated the results obtained by our strategy and found some interesting genetic response tendencies and regulatory manners in *S. cerevisiae* during transient heat shock.

3. MATERIALS AND METHODS

3.1. Data collection and preprocessing

Five normalized time-course array datasets generated by four independent groups (18-21) were downloaded from public websites including GEO (Gene Expression Omnibus) (22) and SMD (Stanford Microarray Database) (23). Missing data were allowed. Four of the experiments used cDNA array and one used oligonucleotide array (Table 1). All of them contained short time points (less than ten) and lacked in repeats. The heat shock process was from 25°C to 37°C. Although cells grown at 25°C were collected for the zero time point reference in most datasets, original references used in the five array experiments were not consistent. For comparison, data were mathematically zero transformed by dividing the expression ratios of each gene at each time point by the corresponding ratios measured for the unshocked cells (25°C, 0 time point). To see if the timings of biological process in the five datasets were similar, we roughly judged the number of differentially expressed genes at each time point in each array experiment by a log ratio threshold. Dataset 5 was then filtered out.

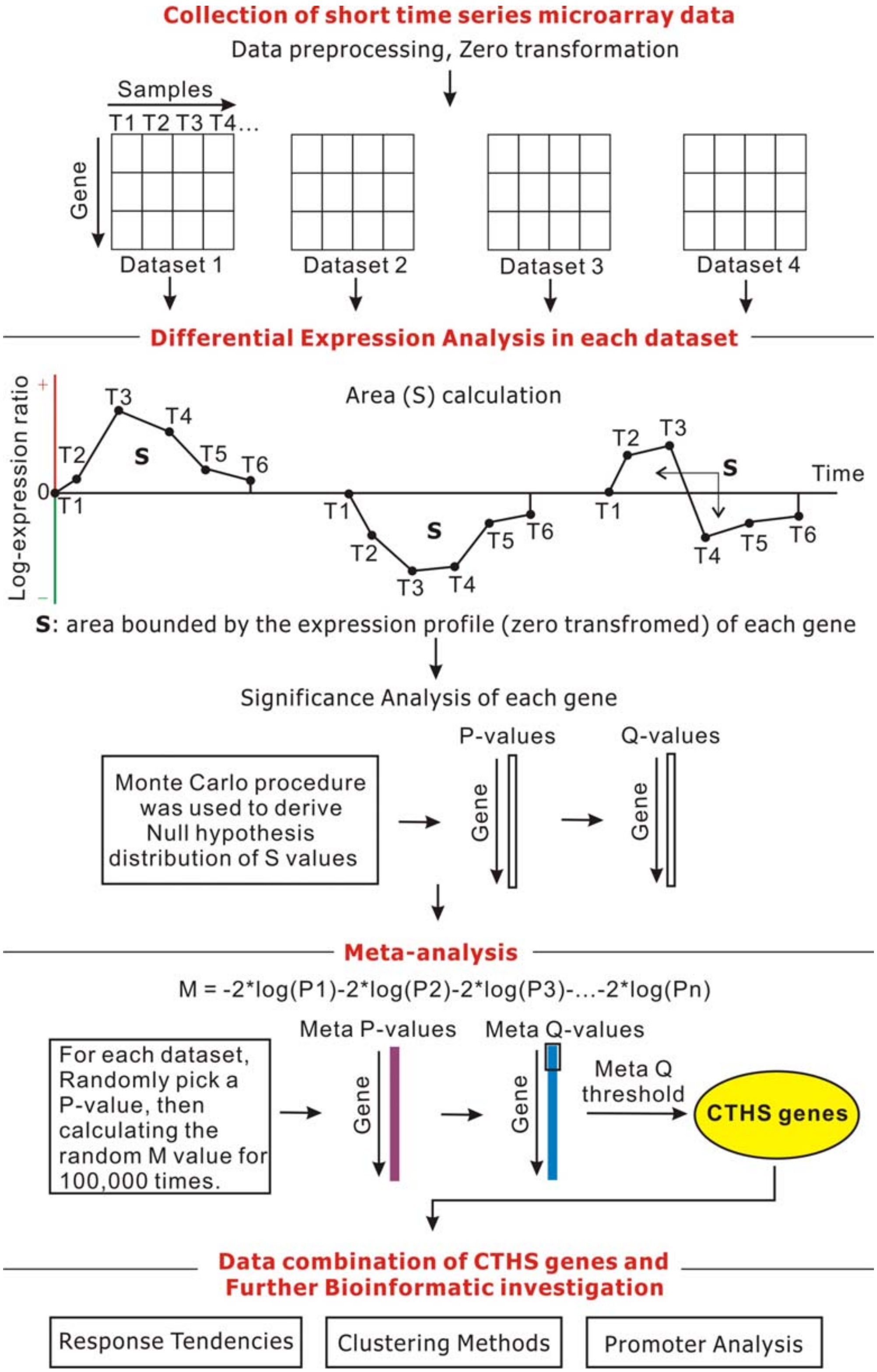
Collection of short time series microarray data
Data preprocessing, Zero transformation
Samples
T1 T2 T3 T4...
Gene
Dataset 1
Dataset 2
Dataset 3
Dataset 4
Differential Expression Analysis in each dataset
Log-expression ratio
+
0
−
Area (S) calculation
T1
T2
T3
T4
T5
T6
S
Time
S: area bounded by the expression profile (zero transfromed) of each gene
Significance Analysis of each gene
P-values
Q-values
Monte Carlo procedure was used to derive Null hypothesis distribution of S values
Gene
Meta-analysis
M = -2*log(P1)-2*log(P2)-2*log(P3)-...-2*log(Pn)
Meta P-values
Meta Q-values
For each dataset, Randomly pick a P-value, then calculating the random M value for 100,000 times.
Meta Q threshold
CTHS genes
Data combination of CTHS genes and Further Bioinformatic investigation
Response Tendencies
Clustering Methods
Promoter Analysis

Figure 1. Four major steps in our meta-analysis strategy for short time series microarray datasets. Step 1: Data collection and preprocessing, including data filter and zero transformation. Step 2: Differential expression analysis in each dataset, including area calculation and a Monte Carlo procedure. Step 3: Meta method for yielding commonly transient heat-shock sensitive genes (CTHS genes). Step 4: Data merging and further bioinformatic investigation, including clustering analysis and promoter analysis.

3.2. Differential expression analysis in each dataset

Our analysis was performed with custom software written in Perl. Considering the time-ordered aspect and short time points of the data, in each dataset, we calculated the area (S value) bounded by the expression profiles for each gene in 37°C and 25°C conditions (X axis after zero transformed, Figure 1). Let's call X (T_n) the log-expression ratios in 37°C condition, available for a generic gene X at time sample T_n (n = 1, ..., k, with k number of time samples). Then S value was calculated for each gene as the sum of the contributions of partial areas from consecutive samples (equation 1). Each contribution S_n was calculated from the deviation of expression in 37°C and 25°C conditions (X axis here) between neighbouring time points. To decide the induced or repressed genes during the heat shock process, we also adopted S' value computed by subtracting the area below X axis from the area above X axis for each gene. Genes with positive S' values were decided to be induced and genes with negative S' values were considered to be repressed

$$S = \sum_{n=1}^{k-1} S_n \quad \text{(equation 1)}$$

The gene X is considered differentially expressed if the area S of this gene was greater than the threshold, which was in correspondence to a significant level based on the null hypothesis distribution of the areas. Since the four datasets were in data-poor condition, i.e. a sufficient number of replicates was not available, a Monte Carlo procedure was used to derive the null distribution of the S values. A more detailed explanation of this procedure was previously described (15). Briefly, the null distribution of the expression level at each time sample was derived from the true expression values obtained from available replicates (at least two replicates for each time sample would be necessary), then R profiles of length k were sampled (here R = 10^4) from the null distribution of the expression level at each time sample and the S values of R profiles were calculated. Subsequently, different distribution models (Gamma, Log-normal and Weibull) were used to fit the entire set of S values of R profiles and the best model was chosen on goodness of fit. Once the null hypothesis of S values was obtained, gene-specific *P* values were assigned according to the true S values of the genes. To calculate the gene-specific false discovery rate (called Q value here), genes were sorted by *P*, and then the ratio of the expected number of occurrences at or better than each *P* to the actual number of occurrences was calculated (equation 2, N denotes total number of genes, I denotes number of genes at or better than *P*)

$$Q-value = \frac{P*N}{I} \quad \text{(equation 2)}$$

3.3. Meta-analysis and identification of commonly transient heat stress sensitive genes

The meta-method was modified from Rhodes *et al.* (24). For each possible combination of four datasets, we performed a meta-analysis to test the null hypothesis that significant results from individual studies do not correspond to the same genes. For each gene, a *P* meta statistic (M) was computed using the *P*s from individual datasets (equation 3, n denotes dataset number).

$$M = -2\log(P_1) - 2\log(P_2) - ... - 2\log(P_n)$$
(equation 3)

Then meta *P* value were calculated by a comparison to 100,000 meta values generated by randomly selecting a *P* from each dataset contributing to the respective meta-analysis. The meta statistic *P* value equaled the fraction of random summary statistics that were greater than or equal to the actual. For each meta-analysis, we sorted genes by meta *P* value, calculated the meta Q value of a gene as the ratio of the expected number of occurrences at or better than the *P* of the gene to the actual number of occurrences (same equation 2). We assimilated results from all of the meta-analyses by selecting the minimum meta Q value for each gene. Finally, all the genes were ranked by their minimum meta Q values and the commonly transient heat stress sensitive genes (the CTHS genes) were identified according to the meta Q value threshold (minimum meta Q value < 0.1).

3.4. Selective combination of time series data for further bioinformatic analysis

Since the four time series datasets were not sampled uniformly, a standard linear interpolation was performed to make the time points come into line. The points were joined by straight line segments and each segment can be interpolated independently, in consistence with the S value calculation. We selected five time points: 5, 15, 30, 45, 60 min to parallel for further combination. All the four datasets contained 5, 15, 30 and 60 min time points and we just needed to interpolate the 45 min for dataset 2, 3, 4. Interpolation was performed using equation 4. X_i, X_2 and X_1 are the expression values at the interpolating time point and two source time points, respectively. t_i, t_2 and t_1 are the corresponding time points. After interpolation, the four datasets were transformed to a comparable form. This process was carried out using equation 5. X_t and X_i mean transformed and untransformed values at the same time point in each array dataset, mu and sigma are the mean and SD of distribution of X_i. Therefore all the values in each array dataset were turned into variables which were normally distributed with mean amounted to 0 and SD to 1.

$$X_i = \frac{X_2 - X_1}{t_2 - t_1} t_i + \frac{X_1 t_2 - X_2 t_1}{t_2 - t_1} \quad \text{(equation 4)}$$

$$X_t = \frac{X_i - \mu}{\sigma} \quad \text{(equation 5)}$$

For better combination, the correlation coefficients were calculated between two random array datasets for each CTHS gene. Equation 6 shows that the correlation coefficient r_{xy} is computed using transformed values X_i, Y_i of each gene and their means at all time points between two given array datasets. For each CTHS gene, the array datasets whose r_{xy} was greater than or equal to 0.6

Table 2. The number of promoter elements in each cluster

Label	**Elements**	**Total**	**Early**	**Middle**	**Late**
	Single	38	4	24	10[1]
H1	Perfect HSE	17	2	9	6
H2	Gap HSE	17	1	14[1]	2
H3	Step HSE	4	1	1	2
	Mixed	46	16[2]	26	4
H12	(Perfect + gap) HSE	5	3	2	0
H123	(Perfect + gap + step) HSE	1	1	0	0
SH1	STRE + perfect HSE	13	3	9	1
SH2	STRE + gap HSE	18	5	12	1
SH3	STRE + step HSE	5	1	2	2
SH12	STRE + (perfect + gap) HSE	3	2	1	0
SH23	STRE + (gap+ step) HSE	1	1	0	0

[1] Weakly correlated (not significant) with the corresponding phase ($P < 0.09$) and [2] Significantly correlated with the corresponding phase ($P < 0.05$, X^2 test).

were chosen to combine the transformed values at each time point. Combined values were the mean of transformed values of correlative array datasets.

$$r_{xy} = \frac{\sum_{i=1}^{5}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{5}(X_i - \overline{X})^2 \sum_{i=1}^{5}(Y_i - \overline{Y})^2}} \qquad \text{(equation 6)}$$

3.5. Clustering method and promoter analysis

First, we analyzed the functional distribution of the CTHS genes according to CYGD database. Then we imported the combined time course data to Genesis software (version 1.7.0) developed by Alexander Sturn. Genesis is a platform independent Java package of tools to simultaneously visualize and analyze a whole set of gene expression experiments (25). By using the expression view function of this software, we obtained the overall genetic response tendency of heat stress sensitive genes. Hierarchical clustering analysis was used to the CTHS genes from a case functional category since it has been reported that hierarchical clustering method could produce the same result as which were produced by the complex hidden Markov model based clustering algorithm when dealing with short time series microarray data (14).

For further promoter analysis, we analyzed 1000 base pairs upstream of the start codon of the loci verified using a single pattern identification program (programmed using Perl). Sequences were retrieved from SGD database. We concentrated on three reported types of HSEs and STRE. HSEs consist of three inverted repeats of the nGAAn unit in perfect, gap and step arrangements. The perfect HSE consists of contiguous units, either GAAnnTTCnnGAA or TTCnnGAAnnTTC. The gap HSE consists of 5 any base pair gap units, TTCnnGAAnnnnnnnGAA and its complement TTCnnnnnnnTTCnnGAA. The perfect and gap type also allow a single mismatch, such as GAR or YTC. R is any purine base and Y is any pyrimidine base. The step type has five base pair inserts between three direct repeat units, TTCnnnnnnnTTCnnnnnnnTTC or GAAnnnnnnnGAAnnnnnnnGAA. STRE is a core promoter sequence CCCCT bound by Msn2/4. We labelled perfect HSEs as H1, gap HSEs as H2, step HSEs as H3 and STRE as S for convenience. If the promoter region of a locus contained mixed elements, we composed these labels such as SH1, H23 etc (Table 2). Hierarchical clustering analysis was performed on the CTHS genes which contained such elements in their promoter regions.

4. RESULTS

A flow diagram showing the strategy of our meta-analysis is seen in Figure 1. There are four major steps in our strategy: 1) Data collection and preprocessing (including data filter and zero transformation); 2) Differential expression analysis in each dataset (based on area calculation and the null distribution of the areas); 3) Meta method for yielding commonly transient heat shock sensitive genes (CTHS genes); 4) Data merging and further bioinformatic investigation. Details of each step are described in Materials and Methods. The results found by using this strategy are presented below.

4.1. Collection of short time series microarray datasets

We downloaded five time-course microarray datasets generated by four independent groups (18-21) (Table 1). These datasets pertained to transient heat shock response in wild type *S. cerevisiae* (from 25°C to 37°C). All of them contained short time points (less than ten) and lacked in repeats. Since reference samples used in different array experiments were not consistent, the datasets were mathematically zero transformed by dividing the expression ratios of each gene at each time point by the corresponding ratios measured for the unshocked cells (25°C, 0 time). A critical prerequisite for combination of time series datasets is that the timings of biological process in these datasets are similar. We then roughly determined the number of differentially expressed genes at each time point in each array experiment by a log ratio threshold (log ratio > 1 or < -1). As shown in Figure 2, dataset 1-4 showed a similar expression tendency that the number of differentially expressed genes reached the peak at 15 min but dataset 5 did not, suggesting the timing of biological process in dataset 5 was not as same as that in other datasets. Hence, we selected dataset 1-4 for further meta-analysis.

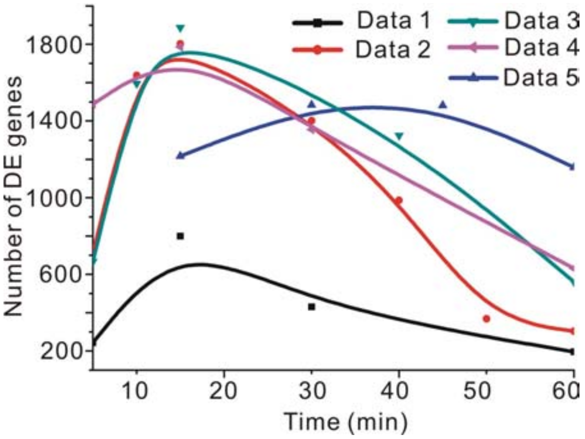

Figure 2. A rough investigation of the expression changes in each datasets. DE genes means differentially expressed genes identified using expression level thresholds (log ratio value >1 or <-1) at each time point in individual microarray datasets. Solid curves represent a computer described best fit to the data obtained.

4.2. Differential expression analysis in each dataset

We evaluated the results of the individual datasets before performing the interstudy analysis. For each gene in each study, we calculated the area (named as S value) bounded by the expression profile and X axis (Figure 1). The S value reflects the intensity of the gene expression change during transient heat stress. To decide the induced or repressed genes during heat shock process, we also computed the S' value by subtracting the area below X axis from the area above X axis. Because the four datasets lacked in repeats, a Monte Carlo procedure was used to derive the null distribution of areas (see Materials and Methods). *P* values (*P*s) were assigned to each gene according to the Gamma distribution which was the best fit for the distribution of S values. For multiple tests, we then adjusted the *P* values by calculating the estimated lowest gene-specific false discovery rates (Q values), which have been suggested as a measure of significance analogous to *P*s but adapted to multiple inference scenarios. Q values were calculated by ranking genes in each dataset by their *P*s and then calculating the ratio of expected random occurrences at or better than a *P* of a gene to the actual number of occurrences. Figure 3A depicts the Q value plot of the four datasets analyzed (the X axis is the rank index sorted by *P*s). All the four datasets, especially dataset 3 and 4, had hundreds of genes with Q value much less than one, suggesting that many genes were differentially expressed during transient heat shock.

4.3. Meta-analysis and identification of commonly transient heat stress sensitive genes (CTHS genes)

We then implemented a meta-analysis model modified from Rhodes *et al*. (24) to assess the similarity of the results between studies to identify reliable sets of commonly transient heat stress sensitive genes (named as CTHS genes). *In-silico* interstudy validation and significance analysis were carried out for all the genes by integrating the methods of meta statistics (M value) and false discovery rates (see Materials and Methods). Briefly, to test the null hypothesis that the significant results in the four individual datasets do not correspond to the same genes, we performed a meta-analysis for each possible combination and then assimilated the results. In each possible combination, we calculated the meta M statistics for each gene, and evaluated the significance (meta *P* value) of M value based on a distribution of randomly generated meta M statistics. Finally, to estimate the false discovery rates of meta *P* value, a meta Q value was assigned to each gene in each combination. If a gene significantly responded to transient heat stress, the meta-analysis M statistic of a gene would also be significant (represented by a low meta Q value). On the contrary, if a gene was significant in only one dataset, the M value would not be significant. The meta Q value plot showed that all of the possible combinations of datasets yielded sets of significantly similar genes (Figure 3B). As expected, increasing the number of datasets in the combination increases the significance and number of commonly differentially expressed genes, for example, the combination of all four datasets (Data1234, Figure 3B) yielded the largest number of significantly similar genes. Such a result suggests that our meta-analysis of short time series microarray dataset improves the statistical power of detecting differentially expressed genes.

Results from the various combination analyses were assimilated by selecting the lowest meta Q value for each gene and then sorting genes based on the meta Q values. At a meta Q value < 0.1, 972 commonly sensitive loci were identified during heat shock. By importing these loci to SGD database (26), 826 loci were verified. Approximately 14% of the genes in the yeast genome were found to be involved in the response to heat stress, showing a complex process during heat stress. This finding of a large number of genes enabled us to give a functional interpretation to the genetic response to heat stress. The subcellular localization of these proteins was shown in Figure 4 according to CYGD database (27). The largest two groups of identified proteins originated from the cytoplasm and nucleus. It is noteworthy that 13.85% proteins were originated from mitochondria and 5.79% from ER. Among the CTHS genes, 625 (528 verified) were induced and 347 (298 verified) were repressed during heat stress. The induced and repressed genes were functionally categorized according to CYGD database (Figure 5). The total gene number was greater than 972, because many genes were found in many categories. The number of induced genes was larger than that of the repressed genes in most functional categories, especially in “Energy”, “Protein activity of regulation” and “Cell rescue, defence and virulence”. However, the number of repressed genes in category of “Protein synthesis” and “Transcription” was larger than that of the induced genes, in agreement with a previous expression profiling study of heat shock stress in wild *S. cerevisiae* (28).

4.4. Combination of time series data finds overall genetic response tendency

To further investigate the expression response tendency of the CTHS genes, we managed to combine the four time series datasets. First, we should make the time

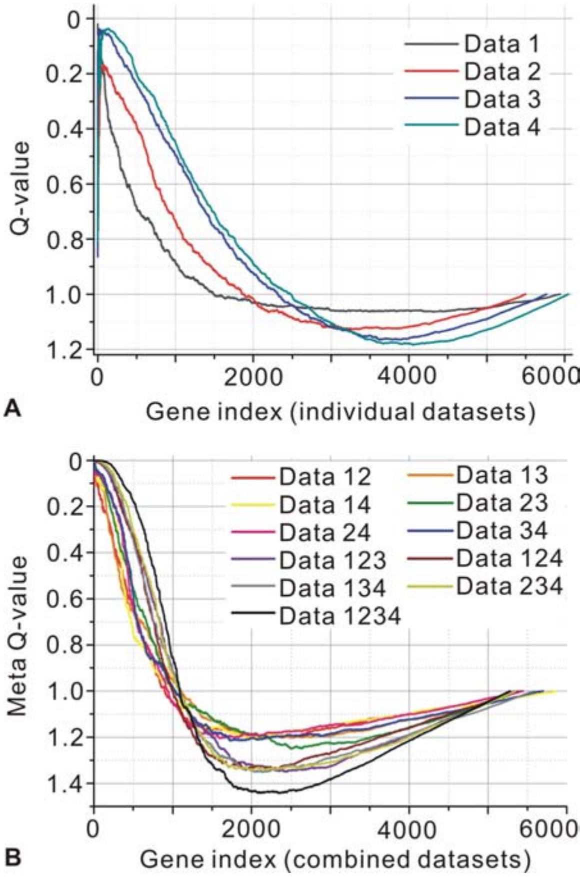

Figure 3. Q-value and meta Q-value plots. (A) Evaluating individual time series datasets by estimated lowest false discovery rates (Q-value). The X axis represents the gene index as ranked by *P*-values generated from the null hypothesis distribution of the S values, and the Y axis represents Q-value. (B) Comparison of meta Q-value plots for analyses of different combinations of the four datasets. The X axis represents the gene index as ranked by meta *P*-values generated from random permutation test, and the Y axis represents meta Q-value.

points come into line because the four series were not sampled uniformly (Table 1). Since the datasets contained relatively short time points and lacked in repeats, it is not suitable for adopting the well developed method such as the continuous representation which requires relatively long time samples (12). To avoid too much interpolation, we picked some common time points (5, 15, 30, 60 min) in four datasets and linearly interpolated the 45 min time point for dataset 2-4. Subsequently, the four datasets were transformed to a comparable form and normalized values of each CTHS gene were attained at each time point (5, 15, 30, 45 and 60 min). Finally, we combined the normalized values of each CTHS gene at the same time points based on a correlation coefficient threshold (see Materials and Methods).

To get an overall view of the expression changes of the CTHS genes along time, the combined values of the CTHS genes were imported into Genesis software (25). As shown in Figure 6A and B, most genes were induced or repressed to the maximal extent at 15 min and showed a similar expression level at the start point 5 min and the ending point 60 min. Analogous tendencies of CTHS genes from individual functional categories were seen in Figure 6C and D. These results suggest that 15-30 min is the most active period for genetic response to transient heat shock in *S. cerevisiae* and the overall genetic response activity decreased later, in consistence with a previous study of transient heat shock in yeast (29). This finding suggests the robustness and adaptability of the genetic system to transient heat stress in *S. cerevisiae*.

4.5. Clustering analysis of a case category identifies detailed response trends

It has been pointed out that the interpretation of clustering results could be problematic when a large number of genes were present in the dataset (17). As the heat shock response process was complicated and the CTHS genes were from different functional categories, we focused on 118 induced CTHS genes categorized in "cell rescue, defence and virulence" and clustered the genes as a case study (Figure 7). In this category, we found many well known molecular chaperones such as *SSE2*, *SSA4*, *SSC1*, *LHS1*, *SSE1* from *HSP70* family, *HSC82* and *HSP82* from *HSP90* family, *HSP78*, *HSP104* from *HSP100* family. The 118 induced CTHS genes were divided into 3 clusters: the early, middle and late phases. In the early phase, almost all genes were induced to their maximal extent within 5 min after exposed to 37^{o}C. Genes in the middle phase were approximately induced to their maximum from 15 min to 45 min and were also divided into 3 clusters: middle-1, middle-2 and middle-3 phases according to their values at 45 min. Genes in the late phase were induced to the maximum during 45 min and 60 min. These results suggest that the clustering algorithm successfully identifies cooperatively regulated genes according to the combined values. A majority of the 118 genes including most HSPs were classified to the middle phase, whereas *SIS1*, *SSA4*, *HSP42*, *HSP78*, *HSP104* were classified to the early phase and only *SCJ1*, *LHS1* were classified to the late phase. These HSPs may be regulated in a different manner as compared to other HSPs during transient heat stress. We also found that the transcription factor Msn2 reached peak in the early phase and Msn4 peaked in a relatively early period (the middle-1 phase). In contrast, the heat shock transcription factor Hsf1 was weakly induced from 15 min to 45 min in the middle-3 phase. Msn2/4 are non-essential transcription factors that recognize stress response elements (STREs) found in the promoters of most HSPs. Whereas, Hsf1 is an essential protein, *S. cerevisiae* utilizes it to activate the expression of a wide variety of genes even in the absence of heat shock. Under heat stress, the induction of Hsf1 may not be as great as Msn2/4. Previous studies have pointed out that Msn2/4 and Hsf1 regulatory systems control the expression of genes classified as different functional groups (30-32): Msn2/4 controls the genes from chaperons, carbon metabolism and oxidative stress, while Hsf1 controls the genes from chaperons, energy generation and cell wall maintenance. The difference in the inducing tendency between Msn2/4 and Hsf1 indicates the different

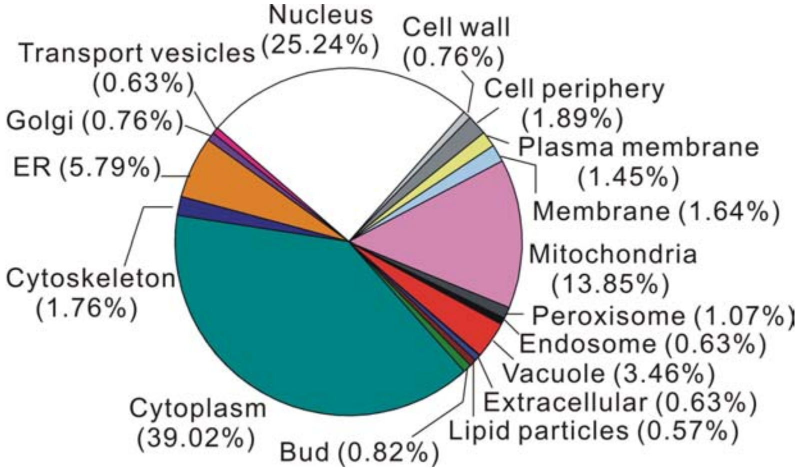

Figure 4. Subcellular localization distribution of the 972 identified ORFs as defined by the CYGD database. The percentage of ORFs present in each subcellular localization is given in parentheses.

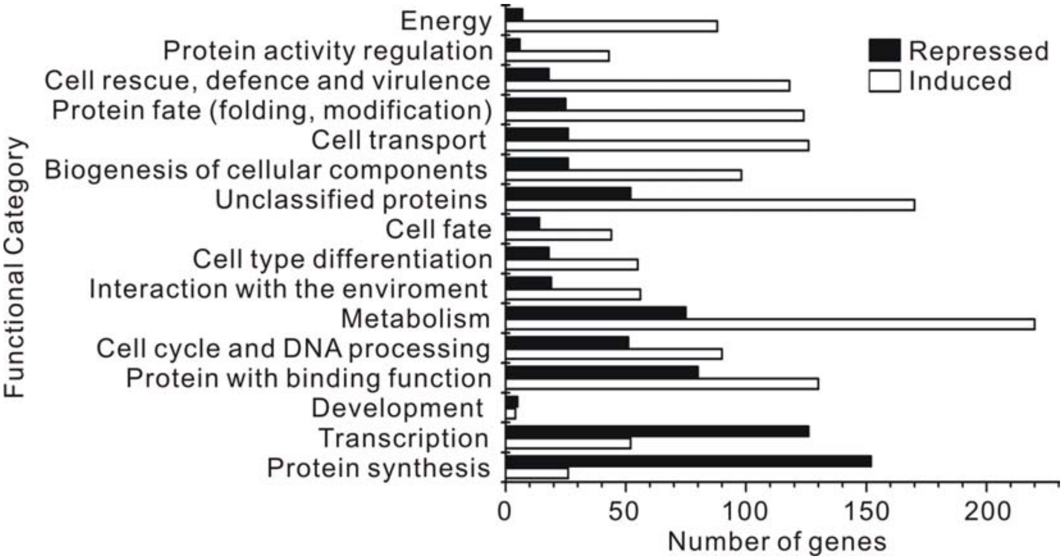

Figure 5. The overview of CTHS genes functionally categorized using the CYGD database. White and black bars represent the numbers of induced and repressed CTHS genes in each functional category, respectively.

functional roles of these two transcription factors during heat shock process.

4.6. Promoter analysis of different heat shock elements

As is described above, both Hsf1 and Msn2/4 transcriptional systems modulate the induction of specific heat shock genes. However, the contribution of Hsf1, independent of Msn2/4, is only beginning to emerge. The heat shock elements HSEs and STRE play independent roles in these two systems, but the response changes along time of the cis-regulons controlled by these elements and the contributions of these elements to transient heat shock remain unclear. To address this, we searched the promoters of the CTHS genes for three types of HSEs and STRE and then clustered the genes containing these elements. 15.8% of the induced CTHS genes contained HSEs in their promoters and 42.1% contained STRE. 52.2% of induced gene contained at least one of these elements, indicating that the Hsf1 and Msn2/4 system are crucial for heat shock response. We also found that many induced genes contained mixed elements, which were labelled as SH1, H12, SH23, etc. Hierarchical clustering method was performed to 84 induced CTHS genes containing HSEs (Figure 8). These genes contained single elements: H1, H2, H3 or mixed elements: H12, H123, SH1, SH2, SH3, SH12 and SH23. We counted the number of genes containing each kind of element in each phase (Table 2). The correlations between the genes containing given elements and the clustering phases were calculated using X^2 test. Interestingly, a distinct correlation was found between the genes containing mixed elements and the early phase ($P < 0.05$). Moreover, genes containing single elements showed

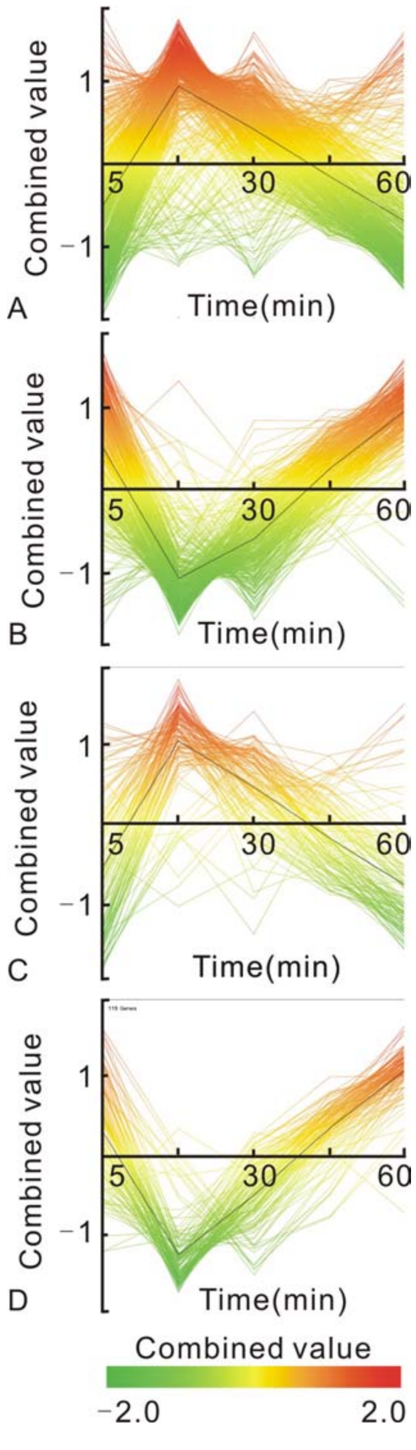

Figure 6. Expression levels (combined values) of the CTHS genes during heat shock are represented by the lines painted in gradient colour scale. Yellow denotes the average level of general combined values. Red and green denotes higher or lower levels, respectively. Mean lines are shown in black. This figure show the expression levels of (A) all the induced CTHS genes, (B) all the repressed CTHS genes, (C) induced CTHS genes from "cell rescue, defence and virulence" category and (D) repressed CTHS genes from "transcription" category.

a weak but not significant correlation with the late phase ($P < 0.09$) and most genes containing single H2 elements tended to reach peak in the middle phase ($P < 0.09$), suggesting that the genes with mixed elements prefer to be induced earlier than the genes with single elements under transient heat stress. Since most of these mixed elements contained STRE recognized by Msn2/4, it seems that Msn2/4 regulatory system reacted more quickly and strongly than Hsf1 system. These findings were consistent with the results from "Clustering analysis of a case category", indicating the validity of our combination method. Moreover, Msn2/4 system may assort with Hsf1 regulatory system in some aspects under heat shock. Several studies have reported that some genes are regulated by cooperation between two systems, such as *HSP26* and *HSP104* (33). Based on our data, this cooperation may play an important role in the early inducing of the CTHS genes in *S. cerevisiae*.

5. DISCUSSION

While tens of thousands of genes are profiled at each microarray experiment, many time series datasets are short (less than ten) and noisy. With the combined power of biologically related but distinct datasets, a suitable meta-analysis strategy for short time series array data were designed here and implemented on four independent short series pertaining to heat shock in yeast. This strategy is composed of four steps discussed below: data selection, differential expression analysis in each dataset, interstudy analysis and further bioinformatic investigation.

Firstly, an important problem in data selection is that whether the different datasets are in the same genetic response rhythm. We performed a rough investigation of the differentially expressed genes at each time point of each dataset and excluded the dataset 5 as it had a distinct expression rhythm as compared to other four datasets. This procedure was a synchronization of different datasets and would improve the analysis of temporal process. Although the four datasets were all about the same biological process, the differences in chip platforms or experimental designs would affect the results of individual studies. For example, the dataset 5 were generated by using an oligo array platform which were different from other datasets (Table 1). Secondly, in individual dataset analysis, area bounded by expression profile was used to represent the gene response intensity to heat stress. This method is suitable for short time series data since it has been demonstrated that the methods base on cubic spline fitting for time-course datasets needed long time points (12, 13) and the methods using area representation worked better than those based on curve fitting (15). In fact, the spline curve fitting methods led to an overfit when working on the short time series datasets (data not shown). Furthermore, since the four datasets lacked in repeats, the distribution model of S values could not be derived from ANOVA or other well developed model. The null distribution of S values was derived from a Monte Carlo procedure, which is applicable in data-poor condition (15). Thirdly, the meta method facilitated the validation and significance analysis of the four heat shock datasets. The combined significance (or

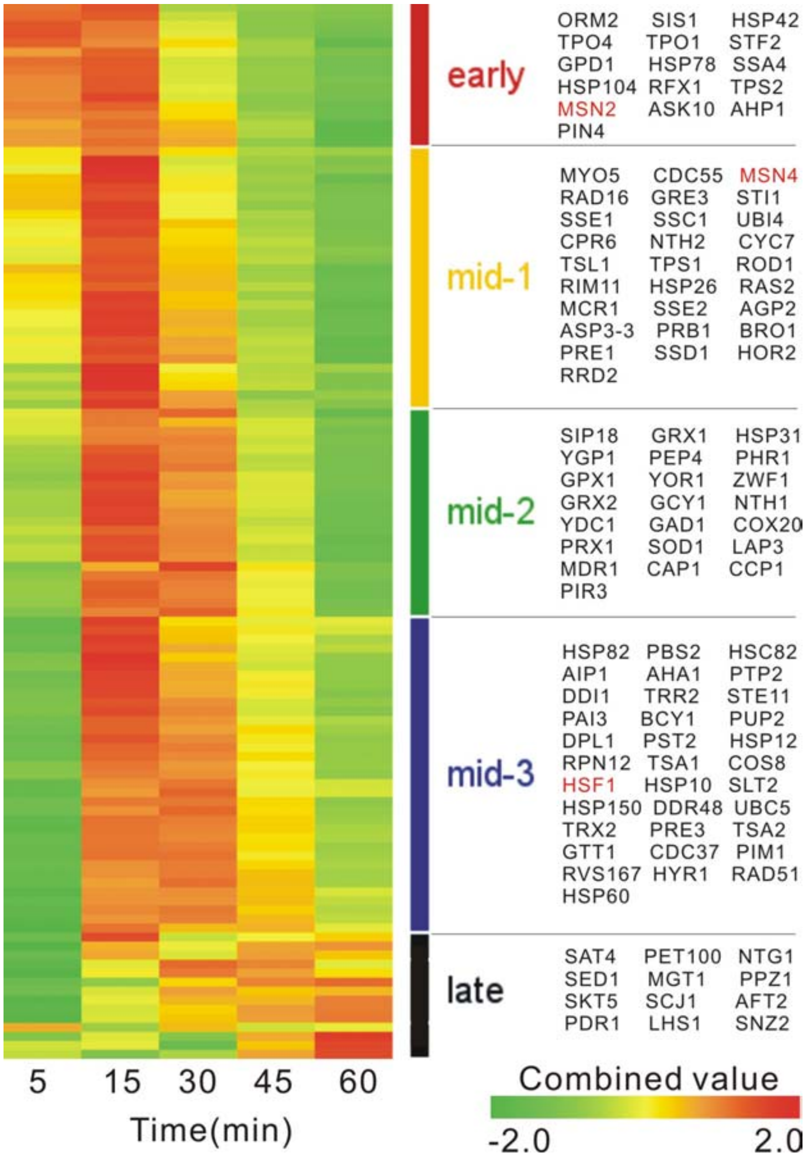

Figure 7. Clustering analysis of a case category using combined values. Hierarchical clustering analysis was performed on the induced CTHS genes classified in the category "cell rescue, defence and virulence". Annotations for clusters are shown and genes classified to each cluster are listed in the right columns. Genes printed in red are the most important transcription factors implicated in heat shock response. Color scale for combined values is shown below.

meta Q value) increased with the number of datasets combined, suggesting that the four datasets had many common differentially expressed genes. The correlation coefficient calculation alleviated the risk resulting from the combination of the array data which were not significantly correlated and led to a precise combination of correlated time course data. Finally, further bioinformatic investigation confirmed the validity of our strategy. Functional categories of CTHS genes and the genetic response tendencies found by using combined values agreed with several previous studies in yeast. Moreover, Msn2/4 transcription factors peaked more quickly and strongly than Hsf1 in clustering analysis, in consistence with the results from promoter analysis, demonstrating the validity of the data combination.

We reviewed the publications and found few suitable methods for meta-analysis of short time series datasets in poor condition. For example, Conesa *et al.* developed an approach named maSigPro for the analysis of multi-series time-course datasets (10). MaSigPro uses a two-step regression strategy to find genes with significant temporal expression alterations. Using MaSigPro to the four datasets we collected, 844 genes can be identified with significantly differential expression profiles. Then our CTHS genes were compared with these 844 MaSigPro-identified genes. Although only 387 genes were

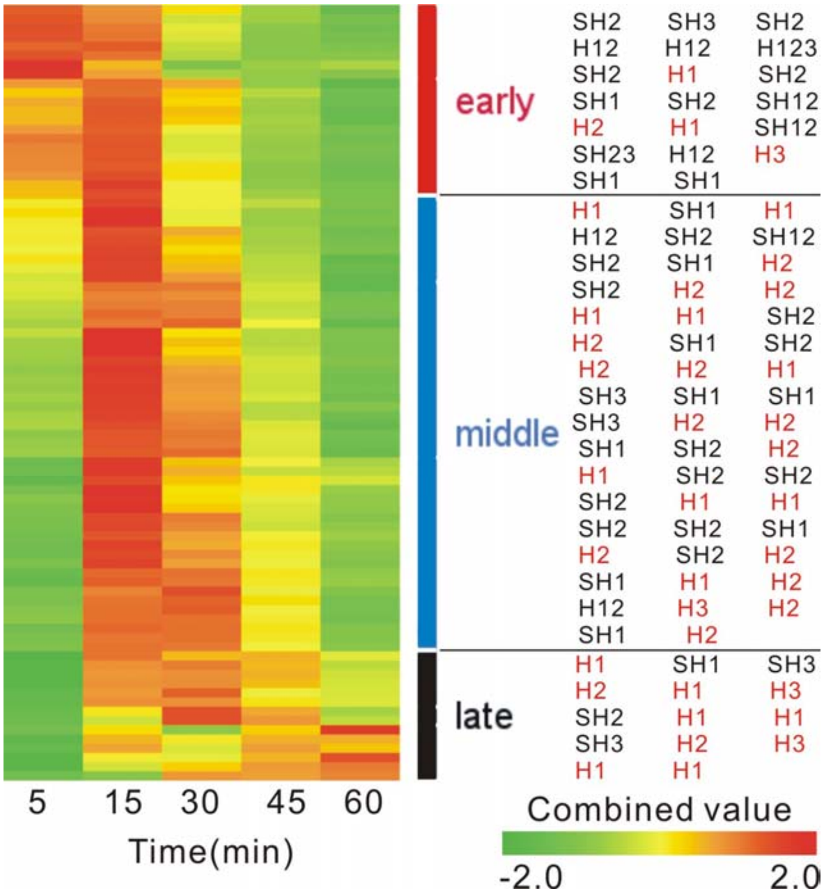

Figure 8. The results of promoter analysis are shown in the same manner as in Figure 7. The types of the promoter elements of 84 induced CTHS genes (containing HSEs) classified to each cluster are listed in the right columns. Red denotes single promoter element types and black denotes mixed promoter element types.

overlapped, most of genes implicated in the heat shock response were included in both MaSigPro-identified genes and our CTHS genes, such as *HSP70* family, *HSP90* family and some other molecular chaperones. As the most important transcription factor in heat stress in yeast, Hsf1 can be identified in both gene subsets. However, Msn2 and Msn4 transcription factors can not reveal enough significance to be identified as MaSigPro-identified gene. Luckily, these two transcription factors were regarded as CTHS genes in our analysis, suggesting that our strategy enhanced the efficiency in identifying the CTHS genes and avoided neglecting important genes in short time series datasets.

There are still some limitations in our strategy. The first limitation is the linear interpolation of missing 45 min time points in dataset 2-4. The missing values ascribed to errors occurring in the time series experimental process that lead to corruption or absence can be estimated with several well-developed approaches, such as the continuous representation method (14). However, such method would lead to an overfit in short time series datasets. So the simplest technique standard linear interpolation (34) was used to make the time points come into line. Although the linear interpolation may take a risk of incorrectly estimating non-linear expression data, the problem is not prominent in our study because most of the datasets had similar time points. For more complicated short time series datasets, an alternative imputation approach needs to be constructed. Once an appropriate interpolative technique is available, our strategy can be more generally applicable. Secondly, the selective combination method base on a correlation threshold may lead to unbalanced contributions of different platforms to the combined values. Further investigation in balancing the combined values using uncorrelated platform is under our consideration. Finally, it also should be mentioned that our study focused on the statistical models for meta-analysis of short time series datasets and further experimental investigation is also needed to validate the findings in molecular regulatory systems in transient heat shock.

6. SUMMARY

For short time-course microarray datasets generated by independent groups, the limitation of data quality (short time points and lacking repeats) and the dynamic nature of experiments pose great challenges to data analysis. Here we introduced a suitable and easy strategy in identifying sensitive genes in multi short time series array datasets and combining expression values to explore the genetic response tendencies. Additionally, the clustering and promoter analysis also provided insights for understanding the transcriptional regulatory mechanisms for several genes in transient heat shock of yeast cells. Up to now, many time series datasets in the microarray databases are short (less than ten) and lack in repeats, our study may help to shed some lights on the usage of this kind of datasets.

7. ACKNOWLEDGEMENT

We thank Mr. Alonso Fuentes for copyediting this manuscript. This work is supported by a grant 2006AA02Z324 from the National High Technology Research and Development Program of China (863 Program). This work is also supported by Shanghai Leading Academic Discipline Project, Project Number: B111.

8. REFERENCES

1. Moreau, Y., S. Aerts, B. De Moor, B. De Strooper & M. Dabrowski: Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet* 19, 570-577 (2003)

2. Wang, J., K. R. Coombes, W. E. Highsmith, M. J. Keating & L. V. Abruzzo: Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics* 20, 3166-3178 (2004)

3. Chen, X., S. Liang, W. Zheng, Z. Liao, T. Shang & W. Ma: Meta-analysis of nasopharyngeal carcinoma microarray data explores mechanism of EBV-regulated neoplastic transformation. *BMC Genomics* 9, 322 (2008)

4. Smith, D. D., P. Saetrom, O. Snove, Jr., C. Lundberg, G. E. Rivas, C. Glackin & G. P. Larson: Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. *BMC Bioinformatics* 9, 63 (2008)

5. Ghosh, D., T. R. Barette, D. Rhodes & A. M. Chinnaiyan: Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Funct Integr Genomics* 3, 180-188 (2003)

6. Grutzmann, R., H. Boriss, O. Ammerpohl, J. Luttges, H. Kalthoff, H. K. Schackert, G. Kloppel, H. D. Saeger & C. Pilarsky: Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* 24, 5079-5088 (2005)

7. Jelier, R., P. A. t Hoen, E. Sterrenburg, J. T. den Dunnen, G. J. van Ommen, J. A. Kors & B. Mons: Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease. *BMC Bioinformatics* 9, 291 (2008)

8. Kong, X., V. Mas & K. J. Archer: A non-parametric meta-analysis approach for combining independent microarray datasets: application using two microarray datasets pertaining to chronic allograft nephropathy. *BMC Genomics* 9, 98 (2008)

9. Rhodes, D. R., J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey & A. M. Chinnaiyan: Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A* 101, 9309-9314 (2004)

10. Conesa, A., M. J. Nueda, A. Ferrer & M. Talon: maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* 22, 1096-1102 (2006)

11. Park, T., S. G. Yi, S. Lee, S. Y. Lee, D. H. Yoo, J. I. Ahn & Y. S. Lee: Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics* 19, 694-703 (2003)

12. Bar-Joseph, Z., G. Gerber, I. Simon, D. K. Gifford & T. S. Jaakkola: Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc Natl Acad Sci U S A* 100, 10146-10151 (2003)

13. Storey, J. D., W. Xiao, J. T. Leek, R. G. Tompkins & R. W. Davis: Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A* 102, 12837-12842 (2005)

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

15. Di Camillo, B., G. Toffolo, S. K. Nair, L. J. Greenlund & C. Cobelli: Significance analysis of microarray transcript levels in time series experiments. *BMC Bioinformatics* 8 Suppl 1, S10 (2007)

16. Morano, K. A., P. C. Liu & D. J. Thiele: Protein chaperones and the heat shock response in Saccharomyces cerevisiae. *Curr Opin Microbiol* 1, 197-203 (1998)

17. Eastmond, D. L. & H. C. Nelson: Genome-wide analysis reveals new roles for the activation domains of the Saccharomyces cerevisiae heat shock transcription factor (Hsf1) during the transient heat shock response. *J Biol Chem* 281, 32909-32921 (2006)

18. Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein & P. O. Brown: Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11, 4241-4257 (2000)

19. Eisen, M. B., P. T. Spellman, P. O. Brown & D. Botstein: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95, 14863-14868 (1998)

20. Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller & N. Friedman: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34, 166-176 (2003)

21. Causton, H. C., B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander & R. A. Young: Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell* 12, 323-337 (2001)

22. Edgar, R., M. Domrachev & A. E. Lash: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30, 207-210 (2002)

23. Cherry, J. M., C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng & D. Botstein: SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 26, 73-79 (1998)

24. Rhodes, D. R., T. R. Barrette, M. A. Rubin, D. Ghosh & A. M. Chinnaiyan: Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 62, 4427-4433 (2002)

25. Sturn, A., J. Quackenbush & Z. Trajanoski: Genesis: cluster analysis of microarray data. *Bioinformatics* 18, 207-208 (2002)

26. Cherry, J. M., C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng & D. Botstein: SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 26, 73-79 (1998)

27. Guldener, U., M. Munsterkotter, G. Kastenmuller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. Garcia-Martinez, J. E. Perez-Ortin, H. Michael, A. Kaps, E. Talla, B. Dujon, B. Andre, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin & H. W. Mewes: CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res* 33, D364-368 (2005)

28. Matsumoto, R., K. Akama, R. Rakwal & H. Iwahashi: The stress response against denatured proteins in the deletion of cytosolic chaperones SSA1/2 is different from heat-shock response in Saccharomyces cerevisiae. *BMC Genomics* 6, 141 (2005)

29. Gasch, A. P.: The environmental stress Response: a common yeast response to diverse environmental stresses**.** In: Topics in Current Genetics vol.1 Yeast Stress Responses. Eds: Hohmann S, Mager WH, *Springer-Verlag Berlin Heidelberg*, Heidelberg, Germany 11-70 (2003)

30. Boy-Marcotte, E., G. Lagniel, M. Perrot, F. Bussereau, A. Boudsocq, M. Jacquet & J. Labarre: The heat shock response in yeast: differential regulations and contributions of the Msn2p/Msn4p and Hsf1p regulons. *Mol Microbiol* 33, 274-283 (1999)

31. Hahn, J. S., Z. Hu, D. J. Thiele & V. R. Iyer: Genome-wide analysis of the biology of stress responses through heat shock transcription factor. *Mol Cell Biol* 24, 5249-5256 (2004)

32. Yamamoto, A., Y. Mizukami & H. Sakurai: Identification of a novel class of target genes and a novel type of binding sequence of heat shock transcription factor in Saccharomyces cerevisiae. *J Biol Chem* 280, 11911-11919 (2005)

33. Amoros, M. & F. Estruch: Hsf1p and Msn2/4p cooperate in the expression of Saccharomyces cerevisiae genes HSP26 and HSP104 in a gene- and stress type-dependent manner. *Mol Microbiol* 39, 1523-1532 (2001)

34. D'Haeseleer, P., X. Wen, S. Fuhrman & R. Somogyi: Linear modeling of mRNA expression levels during CNS development and injury. *Pac Symp Biocomput* 41-52 (1999)

Abbreviations: ER: Endoplasmic Reticulum; ORF: Open Reading Frame; HSP: Heat Shock Protein; HSE: Heat Shock Element; STRE: Stress Response Element

Key words: Meta-analysis, time-series microarray dataset, S value, Permutation Test, Transient Heat Shock, Clustering Analysis, Promoter Analysis

Send correspondence to: Yumin Mao, Institute of Genetics, School of Life Science, Fudan University, Shanghai, PR China, Tel: 8621-65643573, Fax: 8621-65642502, E-mail: ymmao@fudan.edu.cn

http://www.bioscience.org/current/vol14.htm