Research Article

GESearch: An Interactive GUI Tool for Identifying Gene Expression Signature

Ning Ye,[1,2] Hengfu Yin,[3,4] Jingjing Liu,[1,5] Xiaogang Dai,[1,5] and Tongming Yin[1,5]

[1]*The Southern Modern Forestry Collaborative Innovation Center, Nanjing Forestry University, Nanjing 210037, China*
[2]*College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China*
[3]*Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Fuyang, Zhejiang 311400, China*
[4]*Key Laboratory of Forest genetics and breeding, Chinese Academy of Forestry, Fuyang, Zhejiang 311400, China*
[5]*College of Forest Resources and Environment, Nanjing Forestry University, Nanjing 210037, China*

Correspondence should be addressed to Tongming Yin; tmyin@njfu.edu.cn

The huge amount of gene expression data generated by microarray and next-generation sequencing technologies present challenges to exploit their biological meanings. When searching for the coexpression genes, the data mining process is largely affected by selection of algorithms. Thus, it is highly desirable to provide multiple options of algorithms in the user-friendly analytical toolkit to explore the gene expression signatures. For this purpose, we developed GESearch, an interactive graphical user interface (GUI) toolkit, which is written in MATLAB and supports a variety of gene expression data files. This analytical toolkit provides four models, including the mean, the regression, the delegate, and the ensemble models, to identify the coexpression genes, and enables the users to filter data and to select gene expression patterns by browsing the display window or by importing knowledge-based genes. Subsequently, the utility of this analytical toolkit is demonstrated by analyzing two sets of real-life microarray datasets from cell-cycle experiments. Overall, we have developed an interactive GUI toolkit that allows for choosing multiple algorithms for analyzing the gene expression signatures.

1. Introduction

High-throughput gene expression technologies, such as microarray or RNA-seq, can rapidly generate expression profiles of a large number of transcripts at a time, which extract a snapshot of global expression at a certain cellular state of samples [1, 2]. Such technologies are powerful for exploring the genome-wide expression signatures of transcripts. However, the efficient and effective methods in the context of analyzing large expression datasets remain challenging.

To facilitate the data processing and mining, the original dataset is commonly needed to be transformed into a reduced-dimension matrix [3, 4]. Then, the unsupervised clustering algorithms, mainly including hierarchical clustering, *K*-means clustering, and self-organizing neural network (SOM), are widely employed to interpret the expression data [5–7]. Since formats of expression data vary greatly, clustering algorithms generally require to reformat the original data, which might cause loss of useful information [8, 9]. During clustering analysis, it is hard to determine the number of categories. A small number of categories tend to merge the unrelated groups into one category. On the contrary, a large number of categories lead to separating the related members into different categories [8, 9]. Therefore, clustering algorithms might be helpful in understanding the global profiles of gene expression but might not be suitable for identifying coexpression genes with defined expression signatures.

Identifying the gene expression signatures (or molecular signatures) is critical in different biological studies, such as studies on cancer [10, 11], on cell growth and differentiation [12], and on disease diagnose [13, 14]. The group of genes with similar expression profiles in response to internal or external factors could further be used to model biologically relevant networks, which were essential for better understanding the

underlying molecular mechanisms [15, 16]. For instance, Pujana et al. constructed a network relevant to breast cancer by integrating the gene coexpression signatures with functional genomics data and by using prior known information of tumor suppressors [17]. More recently, Aijo et al. proposed methods based on nonparametric Gaussian process to characterize the time-course RNA-seq data and to determine temporally correlated genes during human T helper 17 cell differentiation [18]. More and more attentions have been paid to the extracting gene expression signatures, and there are several public databases, such as gene expression omnibus (GEO) [19, 20], molecular signatures database (MSigDB) [21, 22], and gene signatures database (GeneSigDB) [23], that provide useful platforms for analyzing the gene networks [19, 21–23]. There are also several public servers, such as Cell-Montage, CRCView, and FARO, that allow novel and content-based search for identifying gene expression signatures [24–26]. The efficient approaches for candidates identification and for functional analysis will allow developing new applicable biomarkers, as well as facilitating the understanding of biological processes. To date, several computing packages have been developed for identifying gene expression signatures based on similarity searching. For instance, Fujibuchi et al. defined the periodically expressed genes in different cell-cycle phases by using the sine and cosine vectors [27]; and Xiang et al. designed a method that queried data repositories based on gene expression patterns rather than textual annotations on gene expression omnibus (GEO) [28]. However, it is still challenging to identify meaningful groups of coexpressed genes in many biological scenarios. Different clustering algorithms might produce outputs of genes with distinctly different functions. So it is of great importance to provide a user-friendly platform allowing selecting and comparing of different algorithms and models for biologists. Also the capability of inputting priori knowledge is helpful for functional analysis of specific biological pathways.

In this study, we report an interactive GUI-based package written in Matlab, GEsearch, which can be easily used by click-and-pick. This package enables us to filter the input data, to import the prior knowledge for specific group of genes or expression profiles, and to choose different searching models. Another feature of this package is that the output of coexpression genes can be grouped based on expression abundance and be transferred or visualized for further analysis. We subsequently tested the feasibility of this package by using two independent gene expression datasets from cell-cycle experiments. The results showed that this package is not only efficient to find periodically expressed genes in different cell-cycle phases but also highly capable of predicting the downstream coexpressed genes regulated by a cell-cycle specific transcription factor. Taken together, this package is a useful analytical toolkit for dealing with large-scale gene expression datasets in functional genomic studies.

2. Methods and Implementation

The GEsearch package was implemented in MATLAB. Users without access to the MATLAB need to use the MATLAB Runtime Compiler (MRC) for deploying the package. The GEsearch allows users to filter the input data and to select appropriate algorithms. Then, users can select the number of displaying genes within a window to browse the candidates (Figure 1(a)). Also, users can import expression data with prior knowledge to identify the coexpressed genes (Figure 1(a)). The package and user manual can be found on the webpage (http://115.29.234.170/software/).

2.1. Data Filtering and Preprocessing.

Some gene expression profiles might not meet for the variation requirement (e.g., all-zero expression) and thus could not provide meaningful "signature" information. Such data would be eliminated in the analyzing process. To filter the input data, genes with small variations can be removed by selecting an appropriate threshold as described by the following formula (Figure 1(b)): $y_{i\sigma} = \sqrt{(1/N)\sum_{j=1}^{N}(y_{ij} - \mu_i)^2}$, where y_{ij} stands for the value of row i, column j in the gene expression dataset, μ_i represents the mean value of row i, N is the number of columns, and $y_{i\sigma}$ is the standard deviation for data in row i. The range of values in row i is determined by formula: $\Delta_i = \max(y_i) - \min(y_i)$, where $\max(y_i)$ is the maximum value of data in row i and $\min(y_i)$ is the minimal value of data in row i. If $\Delta_i < T$, the record will be removed in data analysis. In the established toolkit, a scrolling bar provides the threshold information for data filtering (Figure 1(b)), and T threshold for filtering is defined by the value from slider selection (Figure 1(b)).

2.2. Selections of Models

2.2.1. The Mean Model.

One has $\overline{y} = (\sum_{i=1}^{k} y_i)/k$, where k stands for the number of selected genes, y_i is the expression level of gene i, and $\overline{y}$ represents the mean value of selected genes. The mean model is suitable for analyzing datasets with more uncertain prior knowledge, which calculates the average gene expression level. With this model, the similarity is strengthened by suppressing the Gaussian noise.
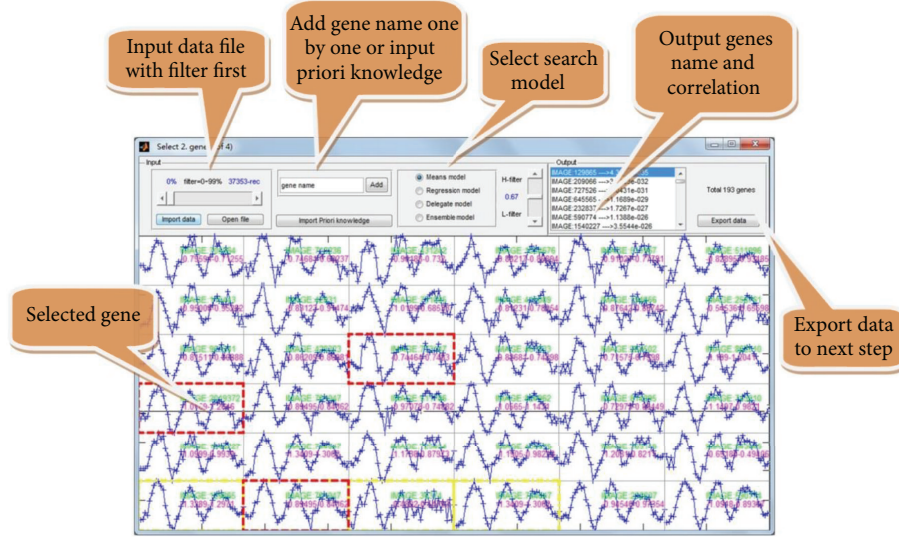
2.2.2. The Regression Model.

The regression model developed a scale-independent algorithm to fix the flaw of strong noise in the mean model. A nonlinear regression model is derived from the linear regression model through regression transformation listed as follows:

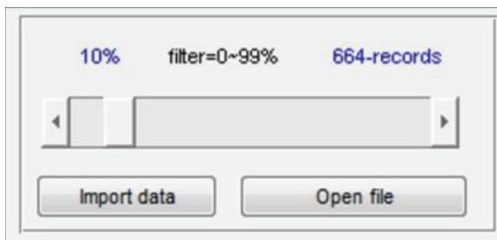$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon_i,$$
$$\varepsilon_i \in N(0, \sigma). \tag{1}$$
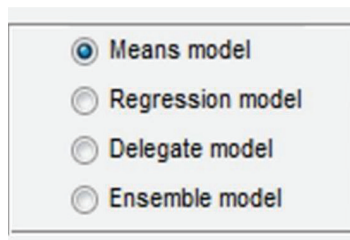
Construct A matrix:

$$A = X'X = \begin{bmatrix} N & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^p \\ & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{p+1} \\ & & & \cdots & \vdots \\ & & & & \sum x_i^{2p} \end{bmatrix},$$
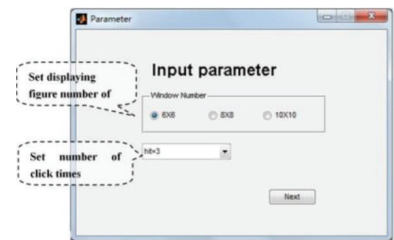
FIGURE 1: The input interface of GESearch. (a) The main window of GESearch, popping-up windows describe the utilities of software. Detailed manual is available at http://115.29.234.170/software/. (b) Panel for data filtering, (c) panel for selecting the searching models, and (d) panel for setting up parameters in GESearch package.

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^P \\ 1 & x_2 & x_2^2 & \cdots & x_2^P \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^P \end{bmatrix}. \tag{2}$$

Calculate the coefficients of regression equation:

$$b = A^{-1}B = \left(X^T X\right)^{-1} X'Y. \tag{3}$$

The regression model is suitable for datasets with gradient changes among sampling points, for example, time-series datasets.

2.2.3. The Delegate Model. The delegate model selects a "delegate" based on a group of records. The delegate model can enhance the scale-independence ability. In many cases, the gene expression level varies greatly. This model is suitable for analyzing expression datasets with dramatic variation ranges.

Therefore $y = \max(\max(y_i) - \min(y_i))$, where $\max(y_i)$ and $\min(y_i)$ stand for maximal and minimal values of selected genes, respectively, and the one with maximal range value was chosen as delegate.

2.2.4. The Ensemble Model. The ensemble model mixes the above three models to give the user a balanced solution for both accurate matching and general searching. The best optimization of this algorithm is to deliver a best-effort matching of all possible results. The ensemble model is developed for the flexible identification of candidates. For each record, this model extracts multiple signatures for similarity searching, which can provide more information for the final results. Comparing to the methods using single signature, ensemble model can be more efficient and accurate.

Therefore $y = UR(y_i)\ i = 1 \cdots k$, where k stands for the number of selected genes, U stands for the total searched genes after removing redundancy, and $R(y_i)$ stands for the results by using gene i.

2.3. Multiple Rounds of Click-and-Pick Search. Users can start selecting genes of interest by randomly browsing the displayed genes. On the interface, options for the numbers of genes are provided. According to the size of the screen and the data, pop-up window allows user to pick an option from 6×6, 8×8, and 10×10 displaying genes (Figure 1(d)).
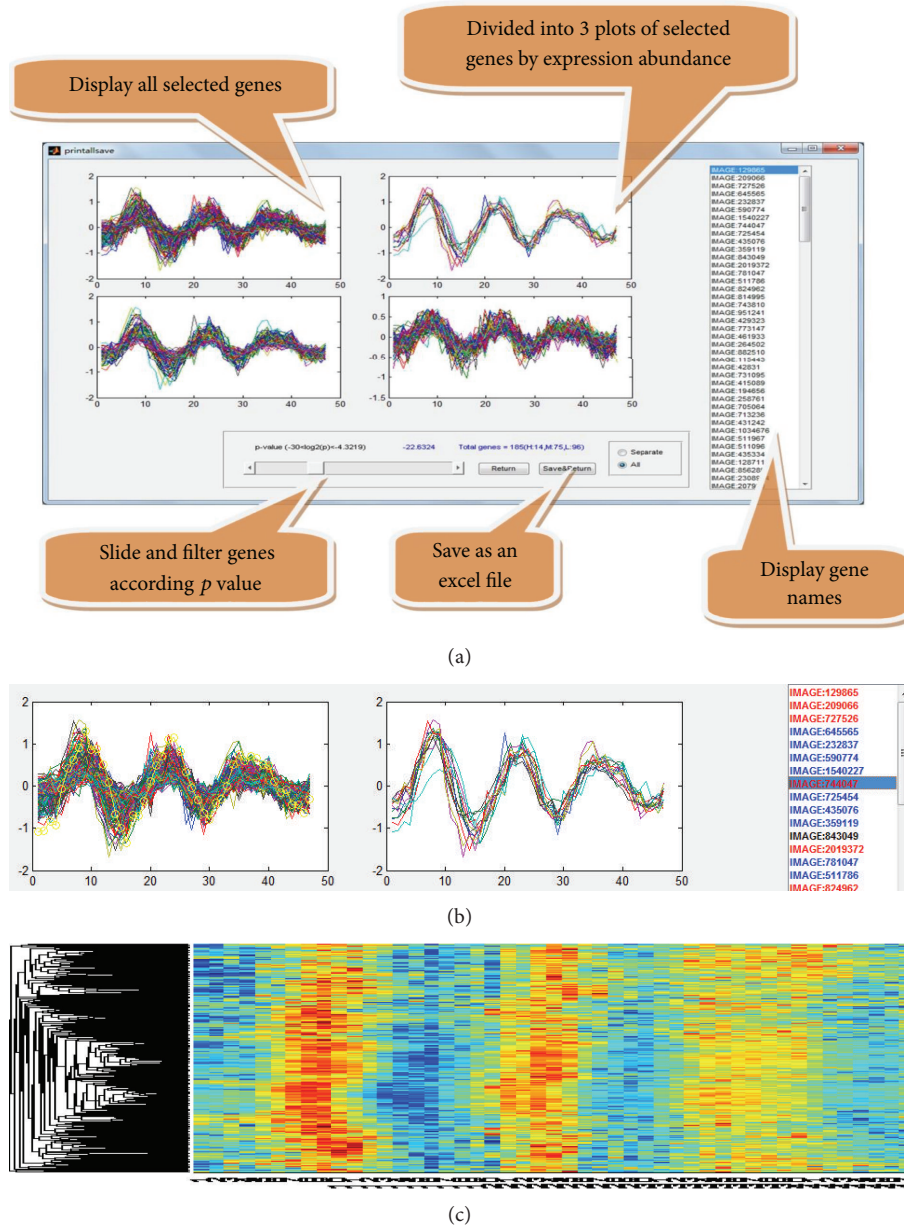
(a)

(b)

(c)

FIGURE 2: The main output interfaces of GESearch. (a) Popping-up windows display the utilities of output. (b) In "separate" model, results are split. (c) Heat map displays the final results.

At the same time, users are allowed to select the number of candidates for identifying coexpression genes (Figure 1(d)). If a subset of candidates is already identified, a pop-up window for importing priori knowledge will initialize the input of the designated gene group.

2.4. Characterization and Exportation of the Output Genes.
There are four options of models, including the mean, the regression, the delegate, and the ensemble models, that can be selected for identifying the coexpressed genes (Figure 1(c)). After model selection, users can define the threshold of output data by scrolling the correlation p value. In the output panel, numbers of identified genes with correlation p values are shown (Figure 2(a)). After optimizing the output parameters, the results can be exported into a text file by selecting the "export data" button (Figure 2(a)). In the output window, the final results can be visualized into three separated panels according to expression abundance (Figure 2(a)). By selecting the "separate" button, the output data will be deposited in three files according to expression abundance. To better display the outputs, red color is used to highlight the highly expressed genes, and blue color is used to highlight the low expressed genes in the displaying window (Figure 2(b)). Furthermore, each output of genes can be easily visualized by the heat map (Figure 2(c)).
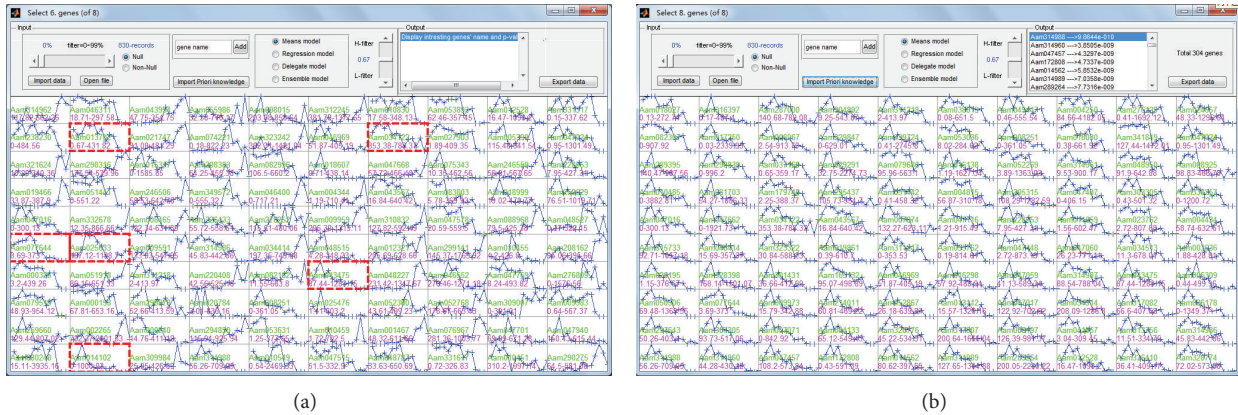
Figure 3: An example of random selecting gene profiles of interest. (a) Display the input data before searching. (b) Resulting panel after selection and search.

3. The Utility of GEsearch

For initial exploration of gene expression dataset, a random selecting and searching approach can be efficient to identify genes of interest. For example, we developed an in-house RNA-seq dataset containing over 90,000 transcripts (derived from *de novo* assembly), with 8 time points. After randomly picked profiles of interest, GESearch found a group of 304 coexpressed genes with highly similar expression patterns (Figure 3). This approach is efficient and straightforward for small or moderate scale datasets in which gene expression signatures can be easily detected and visualized. To examine the feasibility of GESearch, we retrieved and analyzed two large gene expression datasets from human and yeast cell-cycle studies.

3.1. An Example of Using Priori Knowledge. In this example, we selected the human Hela cell-cycle dataset to test the feasibilities of this package [27]. Previous studies have shown that there were specified marker genes representing phases of cell cycle, so we chose a subset of the dataset containing 118 time points and used 20 cell-cycle marker genes representing G1/S, S, G2, G2/M, and M/G1 as priori knowledge to search for the coexpressed periodic genes (http://genome-www.stanford.edu/Human-CellCycle/Hela/data/). This dataset also contained 42920 transcript IDs in which the periodic patterns were not easily seen. In this case, the use of prior known genes was essential for the identification of coexpression genes with defined expression signatures. By using the mean model and default filtering parameter (p value threshold < 0.67), the searching process identified five groups of genes which had very similar periodic expression patterns (Figure 4(a)). The final results were visualized by a heat map chart (Figure 4(a)). To further evaluate the accuracy of the output results, the mean values of each group were plotted together and the expected progression of cell-cycle phases was evident (Figure 4(b)). As a result, this example addressed a complex dataset by using priori knowledge, and the resulting groups of genes were ready for further functional analysis.

3.2. Identification of Downstream Targets of a Transcription Factor. Another useful application of coexpression analysis is to identify potential downstream targets for transcription factors, such as in yeast. The yeast cell-cycle data are from website (http://genome-www.stanford.edu/cellcycle/data/rawdata/) [29], which contains 6187 genes and 18 samples. We select yeast cell-cycle transcription factors MBF (YNL309W) as priori knowledge. Analyzed with GESearch, 74 genes were detected to have the similar expression pattern (results are not shown), among which YGR109C, YHR153C, YPL256C, and YGR221C had been proved to be the targets of MBF, and participating the regulation of cell cycle [30]. Contrast to this study, it was reported that there were 41 downstream targets of YNL309W (http://www.yeastract.com/), and four of them, including YGR109C, YOR372C, YGR221C, and YPL256C, were cell-cycle target genes (Figure 5(a)) [30, 31]. Compare these two studies; it was notable that YOR372C was not detected by GESearch. A detail check showed that the expression pattern of YOR372C shifted and peaked about 3-4 hours earlier than that of MBF (Figure 5(b)). This observation showed that YOR372C did not coexpress with MBF and explained why it was not detected by GESearch. It was also noteworthy that our search engine detected a new cell-cycle gene (YHRL153C) in the coexpression group of MBF (Figure 5(a)), which encoded the G1 cycling activating G1 to S phase transition, and its expression level was regulated by MBF complex [31, 32]. This example confirmed that the GESearch package was efficient in finding coexpression genes as well as in predicting the target genes of transcription factor.

4. Conclusions and Discussion

The design of this package follows the logic that gene expression profiles contain biologically relevant signatures; gene expression signatures are predictable and complement for understanding biological processes. And these notions have been extensively investigated by studies on cancer. For example, gene expression analysis of microarray data has played pivotal roles in breast cancer classification, prognostication, and prediction [33, 34]. It was noteworthy that
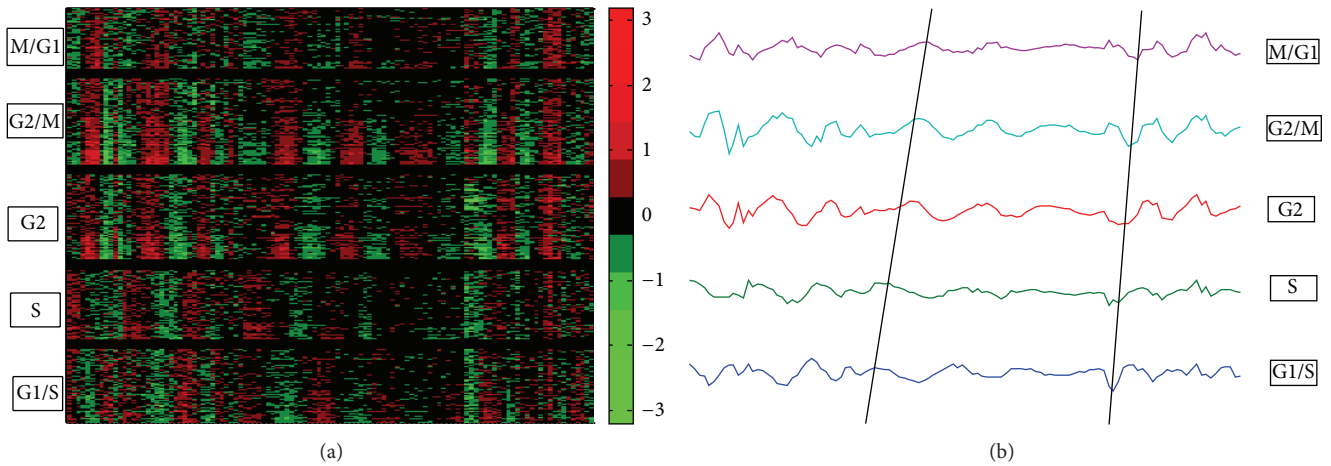
(a)   (b)

FIGURE 4: Results of using priori knowledge of cell-cycle marks in human Hela cell data. (a) Heat map displays the searching results. (b) Plotting the mean values of each group, in which black lines indicate the trends of peaks.

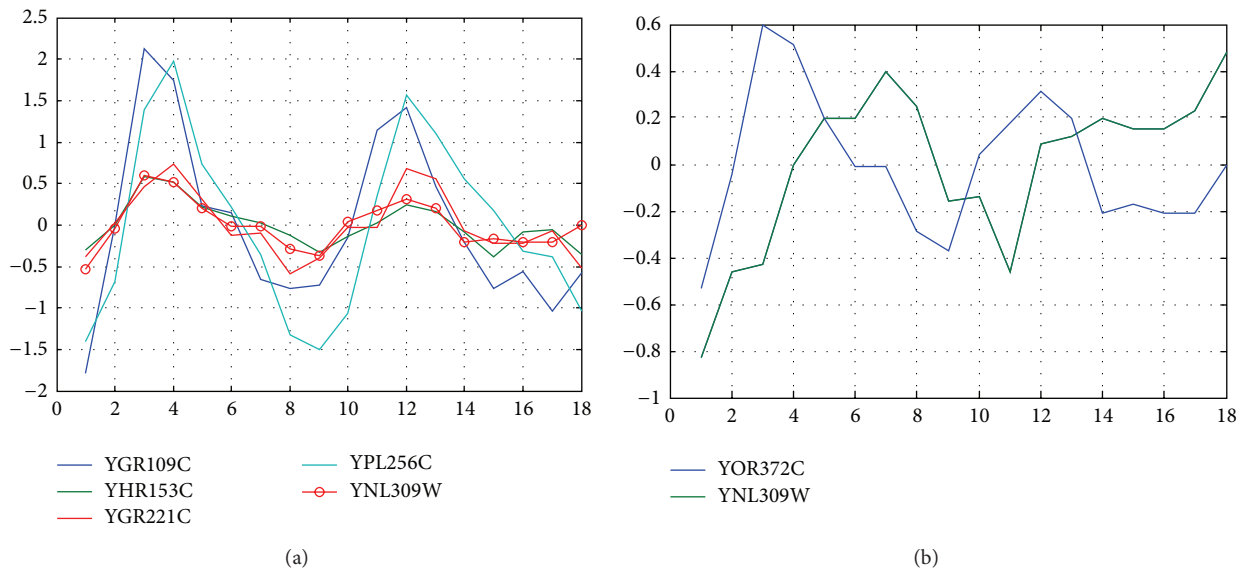(a)   (b)

FIGURE 5: Coexpression analysis of yeast cell-cycle transcription factor YNL309W. (a) GESearch detected four cell-cycle target genes, including the known targets of YGR109C, YPL256C, and YGR221C, and the newly detected YHR153C. (b) The target gene of YOR372C was in shifted expression profiles, which was not detected by GESearch, but was identified by the time delay analysis.

gene expression signatures based predictions were severely affected by the measuring methods and selected datasets [34]. To obtain reliable and repeatable results from gene signatures extraction, multiple choices and comparisons of algorithms are essential.

GESearch provided a comprehensive meaning for finding coexpression genes set of interest, and its multiple choices of searching models allowed rapid identification, regardless of dependent variables. Moreover, this analytical platform is user friendly. As shown previously, the output datasets from cell-cycle experiments were ready for visualization (Figure 2). With the availability of functional information, output datasets could be transformed to be adopted for the functional enrichment tools [35]. GESearch was not only limited to coexpression analysis [36] but also useful for identifying the downstream candidates. Although the time delay analysis was successful for target prediction, the resolution of sampling affected the outcomes of expression signatures due to the specific interaction of transcription factors [37]. It needs further investigations on the defined co- and shift-expressed genes to find the true downstream targets of a transcription factor.

In conclusion, GESearch provides an interactive platform for exploiting the large-scale gene expression datasets. Its choice of multiple models is more users friendly for identifying the coexpressed genes, and its options for data output and visualization provide an efficient way to integrate data, which can help the further analyses.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Ning Ye and Hengfu Yin contributed equally to this work.

Acknowledgments

References

[1] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, no. 1, pp. 55–65, 2006.

[2] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[3] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[4] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data," *BMC Bioinformatics*, vol. 14, article 91, 2013.

[5] S. Datta and S. Datta, "Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes," *BMC Bioinformatics*, vol. 7, article 397, 2006.

[6] J. Sivriver, N. Habib, and N. Friedman, "An integrative clustering and modeling algorithm for dynamical gene expression data," *Bioinformatics*, vol. 27, no. 13, pp. i392–i400, 2011.

[7] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, "An improved algorithm for clustering gene expression data," *Bioinformatics*, vol. 23, no. 21, pp. 2859–2865, 2007.

[8] F.-X. Wu, "Genetic weighted k-means algorithm for clustering large-scale gene expression data," *BMC Bioinformatics*, vol. 9, supplement 6, article S12, 2008.

[9] R. Sharan and R. Shamir, "CLICK: a clustering algorithm with applications to gene expression analysis," *Proceedings/International Conference on Intelligent Systems for Molecular Biology; ISMB International Conference on Intelligent Systems for Molecular Biology*, vol. 8, pp. 307–316, 2000.

[10] J. E. Gershenwald and M. Bar-Eli, "Gene expression profiling of human cutaneous melanoma: are we there yet?" *Cancer Biology & Therapy*, vol. 3, no. 1, pp. 121–123, 2004.

[11] K. M. Carr, M. Bittner, and J. M. Trent, "Gene-expression profiling in human cutaneous melanoma," *Oncogene*, vol. 22, no. 20, pp. 3076–3080, 2003.

[12] R. Nair, A. V. Ngangan, M. L. Kemp, and T. C. McDevitt, "Gene expression signatures of extracellular matrix and growth factors during embryonic stem cell differentiation," *PLoS ONE*, vol. 7, no. 10, Article ID e42580, 2012.

[13] D. Chaussabel, W. Allman, A. Mejias et al., "Analysis of significance patterns identifies ubiquitous and disease-specific gene-expression signatures in patient peripheral blood leukocytes," *Annals of the New York Academy of Sciences*, vol. 1062, pp. 146–154, 2005.

[14] A. Salam and J. A. McGrath, "Diagnosis by numbers: defining skin disease pathogenesis through collated gene signatures," *The Journal of Investigative Dermatology*, vol. 135, no. 1, pp. 17–19, 2015.

[15] M. Sonachalam, J. Shen, H. Huang, and X. Wu, "Systems biology approach to identify gene network signatures for colorectal cancer," *Frontiers in Genetics*, vol. 3, article 80, 2012.

[16] U. C. Kalluri, H. Yin, X. Yang, and B. H. Davison, "Systems and synthetic biology approaches to alter plant cell walls and reduce biomass recalcitrance," *Plant Biotechnology Journal*, vol. 12, no. 9, pp. 1207–1216, 2014.

[17] M. A. Pujana, J.-D. J. Han, L. M. Starita et al., "Network modeling links breast cancer susceptibility and centrosome dysfunction," *Nature Genetics*, vol. 39, no. 11, pp. 1338–1349, 2007.

[18] T. Aijo, V. Butty, Z. Chen et al., "Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation," *Bioinformatics*, vol. 30, no. 12, pp. i113–i120, 2014.

[19] D. Sean and P. S. Meltzer, "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.

[20] T. Barrett and R. Edgar, "Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*," *Methods in Molecular Biology*, vol. 338, pp. 175–190, 2006.

[21] A. Liberzon, "A description of the molecular signatures database (MSigDB) web site," in *Stem Cell Transcriptional Networks*, vol. 1150 of *Methods in Molecular Biology*, pp. 153–160, Springer, New York, NY, USA, 2014.

[22] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[23] A. C. Culhane, M. S. Schröder, R. Sultana et al., "GeneSigDB: a manually curated database and resource for analysis of gene expression signatures," *Nucleic Acids Research*, vol. 40, no. 1, pp. D1060–D1066, 2012.

[24] M. P. Manijak and H. B. Nielsen, "FARO server: meta-analysis of gene expression by matching gene expression signatures to a compendium of public gene expression data," *BMC Research Notes*, vol. 4, article 181, 2011.

[25] M. L. Whitfield, G. Sherlock, A. J. Saldanha et al., "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1977–2000, 2002.

[26] J. M. Engreitz, A. A. Morgan, J. T. Dudley et al., "Content-based microarray search using differential expression profiles," *BMC Bioinformatics*, vol. 11, article 603, 2010.

[27] W. Fujibuchi, L. Kiseleva, T. Taniguchi, H. Harada, and P. Horton, "CellMontage: similar expression profile search server," *Bioinformatics*, vol. 23, no. 22, pp. 3103–3104, 2007.

[28] Z. Xiang, Z. S. Qin, and Y. He, "CRCView: a web server for analyzing and visualizing microarray gene expression data using model-based clustering," *Bioinformatics*, vol. 23, no. 14, pp. 1843–1845, 2007.

[29] P. T. Spellman, G. Sherlock, M. Q. Zhang et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[30] J. A. Ubersax, E. L. Woodbury, P. N. Quang et al., "Targets of the cyclin-dependent kinase Cdk1," *Nature*, vol. 425, no. 6960, pp. 859–864, 2003.

[31] R. A. M. de Bruin, T. I. Kalashnikova, and C. Wittenberg, "Stb1 collaborates with other regulators to modulate the G1-specific transcriptional circuit," *Molecular and Cellular Biology*, vol. 28, no. 22, pp. 6919–6928, 2008.

[32] S. Takahata, Y. Yu, and D. J. Stillman, "The E2F functional analogue SBF recruits the Rpd3(L) HDAC, via Whi5 and Stb1, and the FACT chromatin reorganizer, to yeast G1 cyclin promoters," *The EMBO Journal*, vol. 28, no. 21, pp. 3378–3389, 2009.

[33] P.-E. Colombo, F. Milanezi, B. Weigelt, and J. S. Reis-Filho, "Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction," *Breast Cancer Research*, vol. 13, no. 3, article 212, 2011.

[34] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?" *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005.

[35] The Gene Ontology Consortium, "Gene ontology consortium: going forward," *Nucleic Acids Research*, vol. 43, no. 1, pp. D1049–D1056, 2015.

[36] S. J. Kiddle, O. P. F. Windram, S. McHattie et al., "Temporal clustering by affinity propagation reveals transcriptional modules in *Arabidopsis thaliana*," *Bioinformatics*, vol. 26, no. 3, Article ID btp673, pp. 355–362, 2009.

[37] Y. Shi, T. Mitchell, and Z. Bar-Joseph, "Inferring pairwise regulatory relationships from multiple time series datasets," *Bioinformatics*, vol. 23, no. 6, pp. 755–763, 2007.