Advances and Applications in Bioinformatics and Chemistry

Dovepress

open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

A novel biclustering approach with iterative optimization to analyze gene expression data

This article was published in the following Dove Press journal:
Advances and Applications in Bioinformatics and Chemistry
6 September 2012
Number of times this article has been viewed

Sawannee Sutheeworapong[1,2]
Motonori Ota[4]
Hiroyuki Ohta[1]
Kengo Kinoshita[2,3]

[1]Department of Biological Sciences, Graduate School of Biosciences and Biotechnology, Tokyo Institute of Technology, Tokyo, Japan; [2]Graduate School of Information Sciences, [3]Institute of Development, Aging and Cancer, Tohoku University, Miyagi, Japan; [4]Graduate School of Information Sciences, Nagoya University, Nagoya, Japan

Video abstract

Point your SmartPhone at the code above. If you have a QR code reader the video abstract will appear. Or use: http://dvpr.es/MTQoCi

Correspondence: Kengo Kinoshita
Laboratory of Systems Bioinformatics, Graduate School of Information Science, Tohoku University, Aoba-ku, Sendai, 980-8579, Japan
Email kengo@ecei.tohoku.ac.jp

Objective: With the dramatic increase in microarray data, biclustering has become a promising tool for gene expression analysis. Biclustering has been proven to be superior over clustering in identifying multifunctional genes and searching for co-expressed genes under a few specific conditions; that is, a subgroup of all conditions. Biclustering based on a genetic algorithm (GA) has shown better performance than greedy algorithms, but the overlap state for biclusters must be treated more systematically.

Results: We developed a new biclustering algorithm (binary-iterative genetic algorithm [BIGA]), based on an iterative GA, by introducing a novel, ternary-digit chromosome encoding function. BIGA searches for a set of biclusters by iterative binary divisions that allow the overlap state to be explicitly considered. In addition, the average of the Pearson's correlation coefficient was employed to measure the relationship of genes within a bicluster, instead of the mean square residual, the popular classical index. As compared to the six existing algorithms, BIGA found highly correlated biclusters, with large gene coverage and reasonable gene overlap. The gene ontology (GO) enrichment showed that most of the biclusters are significant, with at least one GO term over represented.

Conclusion: BIGA is a powerful tool to analyze large amounts of gene expression data, and will facilitate the elucidation of the underlying functional mechanisms in living organisms.

Keywords: biclustering, microarray data, genetic algorithm, Pearson's correlation coefficient

Background

The complete sequencing of the genomes of many organisms has led to the launch of various omics studies. In one study, the advent of deoxyribonucleic acid (DNA) microarray technology has enabled the monitoring of the expression levels of numerous genes at a time, under many different growth conditions. This technique is now widely used in diverse types of biological research, such as identifying disease markers, reconstructing cellular signaling pathways, and inferring gene regulatory networks. DNA microarray technology has also provided numerous biological insights.[1–3] Data generated from even a few array measurements are quite complex, and the amounts of microarray data available in public databases are dramatically increasing, due to the efficiency and rapid improvement of DNA microarray technologies. As a result, the interpretation of DNA microarray data obtained under a large number of conditions has become a challenging problem.

In the analyses of a large dataset, as the first step, researchers usually search for similar patterns appearing within the data. In the case of DNA microarray data, similar patterns of gene expression data are often investigated by using cluster analyses, such as K-means

clustering[4] and hierarchical clustering.[5] Although clustering can provide considerable biological information, conventional clustering algorithms may not be suitable for some analyses of microarray data for the following two reasons. Firstly, there are many genes that encode proteins involved in several functional activities at a time, but the conventional clustering methods cannot identify these genes, because they only allow a gene to belong to one cluster at a time, instead of multiple clusters. Secondly, it is difficult to find the genes that are co-expressed under a few specific conditions but are differently expressed under other conditions because the similarity of the genes in conventional clustering is determined by the entire expression data.[6,7]

In terms of the above shortcomings, biclustering is more effective than conventional clustering, since it can cluster both genes and conditions simultaneously, and a gene (or a condition) can be involved in multiple clusters at a time.[7] The concept of biclustering was first proposed by Hartigan,[8] and Cheng and Church[9] applied it to search for the most homogeneously expressed genes over certain sets of conditions by using greedy search algorithms.[9] Most biclustering algorithms have been implemented with greedy search algorithms,[1,10,11] to reduce the calculation costs. One such bicluster, a maximum bicluster, is known as a nondeterministic polynomial time (NP)-complete problem that can possibly be solved in polynomial time using a nondeterministic Turing machine,[12] and a greedy search algorithm is required for actual applications to provide efficient approximations. Usually, one greedy search results in one bicluster, and the greedy search approach is repeatedly applied to the data, while preventing the reproduction of similar biclusters. The greedy search then tries to obtain a set of various biclusters as the final output.

Biclustering has also been implemented by using a genetic algorithm (GA) to find a practical solution to balance bicluster quality and calculation cost. A GA emulates an evolutionary processes to obtain nearly optimal solutions.[13] Initially, a set of candidate solutions is prepared; each solution being called a chromosome. The chromosomes evolve by exchanging their parts and changing some elements into a different state, and elite chromosomes are selected to survive as the parents of the next generation. This evolution and selection process is repeated over a number of generations to yield an optimal solution.[13] Bleuler et al[14] first applied GA to biclustering, whereby a binary string (representing a gene or a condition belonging to a bicluster, or not) was employed as a representation of chromosomes. To avoid any redundancy of the resulting biclusters, Bleuler et al introduced a special selection operator called environment selection. Chakraborty and Maka[15] have generated a similar GA-based biclustering, but different in terms of chromosome initialization. Initial chromosomes are prepared by K-means clustering. These methods find an optimum set of biclusters from one GA search. For such methods, it would be difficult to obtain a set of various, nonredundant biclusters, because only better chromosomes can survive by the selection process of GA, and thus the resulting biclusters tend to converge into similar results in the later generations.[14,15] Another type of GA-based biclustering, Sequential Evolutionary Biclustering (SEBI), has a distinct strategy. SEBI initially applies GA to select the optimal bicluster, and then this process is repeated so that the genes and the conditions in the biclusters already selected are less likely to be selected again. In other words, although SEBI would generate a set of diverse biclusters, it de-empathizes the overlap of biclusters, a significant feature of biclustering.[16]

In the present study, we propose BIGA as the basis of a novel biclustering approach. In BIGA, an attempt is made to progressively divide the large amounts of input data into small datasets, by iteratively using GA, such as SEBI. Instead of evaluating a set of biclusters, GA is applied to each division process. Therefore, the resulting biclusters are substantially diverse. In addition, BIGA introduces the overlap state explicitly defined in the ternary digit (or trit) encoding chromosome. In this study, the algorithm is described, the performance of BIGA is compared with those of six existing biclustering algorithms, and the biological relevance of BIGA is evaluated by using gene ontology (GO) enrichment analyses. Finally, we conclude that BIGA is a powerful and practical solution for biclustering with high-dimensional data.

Material and methods

Definition of biclusters

BIGA accepts a set of gene expression data with the matrix form $D = (G, C)$, including N rows of genes $G = \{g_1, g_2, \ldots, g_N\}$ and M columns of conditions or samples $C = \{c_1, c_2, \ldots, c_M\}$, where N and M are the total numbers of genes and conditions, respectively. All genes will be clustered into K overlapping biclusters $B = \{B_1, B_2, \ldots, B_K\}$, and each bicluster ($B_i$) corresponds to a submatrix $B_i = (X, Y)$ of D, where $X \subseteq G$ and $Y \subseteq C$. The sizes of X and Y, ie, the numbers of genes and the conditions of a bicluster, are denoted by n and m, in which $n \leq N$ and $m \leq M$, respectively.

Binary-iterative genetic algorithm

In order to decompose D into B systematically, a binary tree was introduced. Generally, a binary tree comprises nodes

and directed edges, in which each node can be extended to at most two child nodes.[17] In this work, we regarded each bicluster and each edge as a node and a parent–child relationship between a bicluster pair, respectively. We designated the method as BIGA.

BIGA consists of the following three steps. A schematic diagram of BIGA is shown in (Figure 1).

Step 1: A division of microarray data is represented by a string, a sequence of trit (0, 1, 2) with the length of n (number of genes in the parent bicluster) $+m$ (number of conditions in the parent bicluster). The trit 0, 1, and 2 means that an associated gene or condition is contained in either of two biclusters, b_{left} or b_{right}, or both, respectively. This means that one string can encode the division of one bicluster into two biclusters, while allowing overlap. An example of this encoding is shown in (Figure 1A). The "|" symbol serves as a spacer of the genes and conditions for clarity. The string is equivalent to the division illustrated by the matrix (microarray data, or a bicluster) in the middle of (Figure 1A). In the matrix, the rows and the columns correspond to the genes and the conditions, respectively. The cell of the matrix belongs to either b_{left} (blue cell), b_{right} (red), or both (violet), under the decoding rule shown in (Figure 1B). The white cells are ignored because they are not coexpressed with color cells. Consequently, the bicluster shown in the middle of (Figure 1A) represents the division into two biclusters on the right of (Figure 1A).

Step 2: To search for the best chromosome (the best trit string) representing the optimal division of a bicluster, GA is performed (rectangles in Figure 1C). In the GA procedure, a mutation and a crossover are introduced into each chromosome. Each number on a chromosome is altered to 0, 1, or 2, for the mutation; whereas two chromosomes exchange corresponding parts with each other in the crossover. Chromosomes with higher fitness scores (described in the following section) survive in the next generation, and all other chromosomes are discarded. GA was implemented via Java Genetic Algorithm Product,[18] with a mutation rate of 0.01 and a crossover rate of 0.5. Finally, the best chromosome after 100 generations of GA (the underlined string in the rectangle) is selected, based on the fitness score (see the next section). The best chromosome is then decoded into two biclusters (b_{left} and b_{right}). We decide whether to continue with further decompositions after the evaluation of the biclusters, as follows.

Step 3: Evaluation of biclusters. For each child bicluster, the numbers of genes and conditions, the average Pearson's correlation coefficient (PCC), and the parent–child

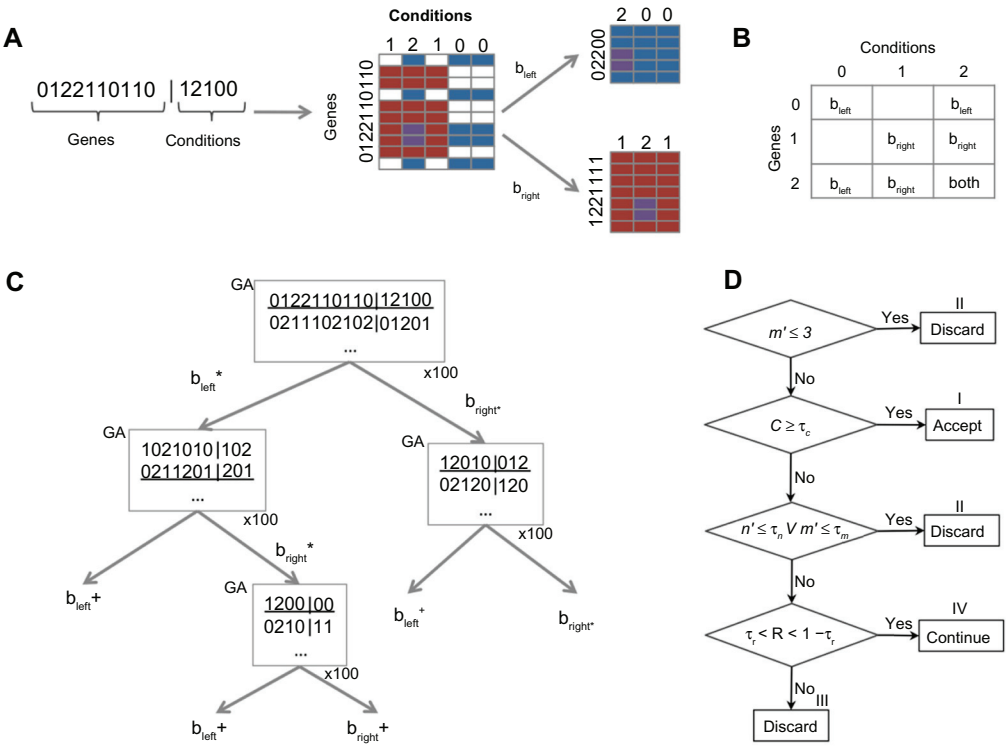

Figure 1 Schematic diagram of binary-iterative genetic algorithm. (**A**) Decomposition of a parent bicluster into two child biclusters encoded in a string (left panel). The string indicates that a parent bicluster (middle panel) is divided into two child biclusters (right panel). The red, blue, and violet cells in the biclusters belong to b_{left}, b_{right}, and both, respectively. (**B**) Decoding rule of a string. (**C**) Binary division performed by genetic algorithm (GA). The best string is underlined in the rectangle. For each GA, the generated biclusters (b_{left} and b_{right}) are evaluated to determine their states: continue the decomposition (*), quit the decomposition and accept (+), or quit the decomposition and discard (–). (**D**) Flow diagram of the bicluster evaluation.

redundancy are examined to decide whether we should quit or continue the decomposition. Subsequently, the bicluster is either accepted as an element of the final biclusters, B, or discarded. We calculate the PCC of every gene pair in a bicluster, and average them (the average PCC). The parent–child redundancy is defined as the ratio of the number of genes of the child bicluster (n') to that of the parent bicluster (n). Therefore, a small parent–child redundancy indicates that the child bicluster contains a smaller number of genes than the parent, and a large parent–child redundancy means that the number of genes in the child bicluster is almost the same as that of the parent. The average PCC and the parent–child redundancy are abbreviated as C and R, respectively. The decision process is illustrated in (Figure 1D). Briefly, the process employs four rules: (I) we quit the decomposition and accept the bicluster if C is higher than the threshold τ_c. (II) we quit the decomposition and discard the bicluster if the bicluster is "small," which is judged by the thresholds τ_n and τ_m for n' and m', respectively. (III) we also quit the decomposition and discard the bicluster if the redundancy, R, is small ($R < \tau_r$) or large ($R > 1 - \tau_r$). The latter rule was employed to reduce the calculation cost, because a child bicluster that is similar to its parent bicluster and has a low C is not considered to produce promising results. Using the forth rule: (IV) we continue the decomposition. Four thresholds, τ_n, τ_m, τ_c, and τ_r, were empirically determined as 30, 10, 0.65, and 0.15, respectively (see Table S1). The Greek symbols in (Figure 1D) indicate the rule applied in each decision. In (Figure 1C), the accepted and discarded biclusters are marked by + and – symbols. The bicluster to be decomposed is marked by a * symbol. Figure 1C indicates that four biclusters are accepted.

Fitness function

In general, large biclusters including co-expressed genes across many specific conditions are preferable. The average PCC of a bicluster was employed to evaluate the gene co-expression. Furthermore, the relative area A of the bicluster, defined by $(n'/n)^{\alpha}(m'/m)^{\beta}$, using the gene and condition numbers of the parent and child biclusters was used to evaluate the size of a bicluster. Two parameters were introduced for gene-weight (α) and condition-weight (β), to control the balance between the number of genes and that of the conditions ($0 < \alpha, \beta < 1$) in a relative area, A. The fitness function of a chromosome was defined as follows (Equation 1):

$$f(c) = A(b_{left})C(b_{left}) + A(b_{right})C(b_{right}), \qquad (1)$$

where c, b_i (i = left or right), $A(b)$, and $C(b)$ denote a chromosome, one of the child biclusters, the relative area of child bicluster b, and the average PCC of child bicluster b, respectively.

The balance between α and β was important in order to select biologically meaningful biclusters when using $f(c)$. Since a high average PCC for a large number of genes was obtained rather easily when only a small number of conditions were considered, a certain number of conditions should be required for each bicluster, to ensure the biological significance. The variation of α and β was empirically estimated, and finally 0.3 and 0.5 were chosen, respectively (see the results in Table S1).

Assessment procedure

Six existing methods were compared to evaluate the performance of BIGA: Cheng and Church algorithm,[9] Statistical-Algorithmic Method for Bicluster Analysis (SAMBA),[19,20] order-preserving submatrix (OPSM),[1] iterative signature algorithm (ISA),[11] binary inclusion-maximal biclustering algorithm (BIMAX),[21] and SEBI.[16] SEBI is selected as a representative of the GA-based biclustering approaches,[15,16] because SEBI adopts an outstanding system to reduce the redundancy of biclusters and performs iterative evolutionary searches like BIGA. The five other methods are based on greedy searches. Data provided by Gasch et al[22] was used for the analyses of *Saccharomyces cerevisiae*. The analyses contained 2993 genes and 173 stress conditions, as a result the data size was large and abundant annotations were available. Prelic et al[21] used this dataset to evaluate algorithms, and the resultant sets of biclusters for the five greedy-search algorithms are publicly available. These bicluster sets were obtained for comparison with our results. Neither the results of SEBI for the data nor SEBI itself is publicly available. The framework of SEBI was re-implemented in a second experiment.[16] Note that there might be some minor differences between SEBI and the re-implemented SEBI. Henceforth, we denote mySEBI as our implementation.

The sets of biclusters were evaluated in terms of the following four points. Since PCC is a widely used parameter to assess the similarity of expression patterns, the distribution of the average PCC of all biclusters was examined. One may consider the mean square residual (MSR) of biclusters[9] to be useful as an indicator of the coherence of biclusters, but PCC is better than MSR in terms of finding the functional relevance of genes,[23–26] in much biological data, for example, the involvement of the same pathway or the participation in the same protein complex.[27,28] The existing methods do not

necessarily optimize the correlation of biclusters, and some biclusters derived from other algorithms can contain biclusters showing strong anti-correlation (ie, genes expressed inversely). The absolute value of PCC was used to estimate such biclusters for comparisons.

Coverage and overlap are also important measures to evaluate the biclustering, as higher coverage and lower overlap are preferable for further biological analyses. Previous studies[29] used "cell coverage," by calculating the percentages of area (genes × conditions) covered by the biclusters, and "cell overlap" by measuring the intersection areas of the biclusters. In this study, "gene coverage" and "gene overlap," were adopted because higher cell coverage can be achieved even by a high coverage of conditions and a low coverage of genes, and this result is not biologically significant. In addition, cell overlap ignores the overlap of genes shared in any two biclusters, if the conditions in the biclusters are completely different. Gene coverage is defined as the ratio of genes that are assigned to any biclusters to all genes, and gene overlap is the ratio of total genes overlapping on multiple biclusters to the genes assigned to any biclusters (Equation 2):

$$Gene\ overlap = \frac{\sum_{i=1}^{k} X_i - \left|\bigcup_{i=1}^{k} X_i\right|}{\left|\bigcup_{i=1}^{k} X_i\right|} \quad (2)$$

Gene coverage can evaluate the ability of an algorithm to decide the cluster for each gene, and gene overlap can measure the ability of an algorithm to specify the clusters for genes that are not necessarily involved in multiple biological processes.

The biological significance of the results by measuring the GO enrichment was also evaluated. More precisely, FuncAssociate (2.0; Roth Laboratories, Harvard University, Boston, MA), a tool for finding overrepresented GO terms in a set of genes was utilised. Using this tool, we performed Fisher's exact test to determine the probability of the appearance of genes associated with a GO term in each bicluster.[30] FuncAssociate calculates an adjusted *P*-value (*P*adj) from the simulations, instead of the corrections of multiple tests. *P*adj is the probability of obtaining at least one false positive for any desired cutoff. We considered a biologically significant bicluster as one that is relevant to at least one GO term with a statistically significant appearance (namely, *P*adj less than significance level). The number of such biclusters, relative to the total number of biclusters (the GO enrichment), was used to estimate each algorithm. A previous study by Prelic et al[21] evaluated the biological relevance of existing algorithms, using the GO enrichment.

Results and discussion

Biclusters for the *Saccharomyces cerevisiae* microarray data

With the selected parameters and thresholds, BIGA found 164 biclusters from the *S. cerevisiae* microarray data. The average numbers of genes and conditions in the biclusters are 92.25 and 23.65, respectively (Table 1). The detailed statistics of each bicluster are provided in Table S2. The properties of the biclusters obtained by other methods are also summarized in Table 1.

Performance evaluation

The distribution of the average PCCs of the biclusters obtained by each biclustering algorithm is shown in the boxplot (Figure 2A). The thick line around the middle of the box indicates the median of the average PCCs. The top and bottom of the box indicate the upper and the lower quartiles, respectively. The circles show the outliers (more than 1.5 times the upper quartile or less than 1.5 times the lower quartile from the median). The whiskers mean the range of data between the maximum and the minimum values, other than the outliers. According to the plots, OPSM performs the best with a very small deviation in the average PCCs. Apart from OPSM, BIGA can outperform the other methods when compared by the median of the average PCC. One may consider that the fitness function of BIGA takes the average PCC into account (Equation 1), and thus it is obvious that the average PCC of BIGA is good. However, note that the results are not necessarily satisfactory if the optimization procedure does not work well, or the balance between the average PCC and the area of the bicluster in (Equation 1) is inappropriate. Next, using the the Wilcoxon signed-rank test the study examined whether the distribution of the average PCCs of BIGA is significantly better than those of the other algorithms.[31] The results showed that BIGA detects significantly more co-expressed genes in biclusters than the other methods, except for OPSM (the highest *P*-value is only 5.4×10^{-6} against SAMBA). To clarify the performance, the expression profiles of the four best biclusters with higher average PCCs are demonstrated in Figure S1. Note: the reason for the highest performance of OPSM was related to the gene coverage and these analyses will be discussed later.

The gene coverage and the gene overlap are shown in (Figure 2B and 2C), respectively. As a result, BIGA achieved the fourth-highest gene coverage among the seven

Table 1 Comparing quantitative metrics among biclustering algorithms

Properties	CC	SAMBA	ISA	OPSM	BIMAX	mySEBI	BIGA
Number of biclusters	100	100	66	12	101	100	164
Average gene number	82.01	911.52	76.27	95.58	24.03	74.98	92.25
Average condition number	19.85	25.15	8.71	12.50	3.00	80.5	23.65

Abbreviations: BIGA, binary-iterative genetic algorithm; BIMAX, binary inclusion-maximal biclustering algorithm; CC, Cheng and Church algorithm; ISA, iterative signature algorithm; OPSM, order-preserving submatrix; mySEBI, the Sequential Evolutionary Biclustering method used in this work; SAMBA, Statistical-Algorithmic Method for Bicluster Analysis.

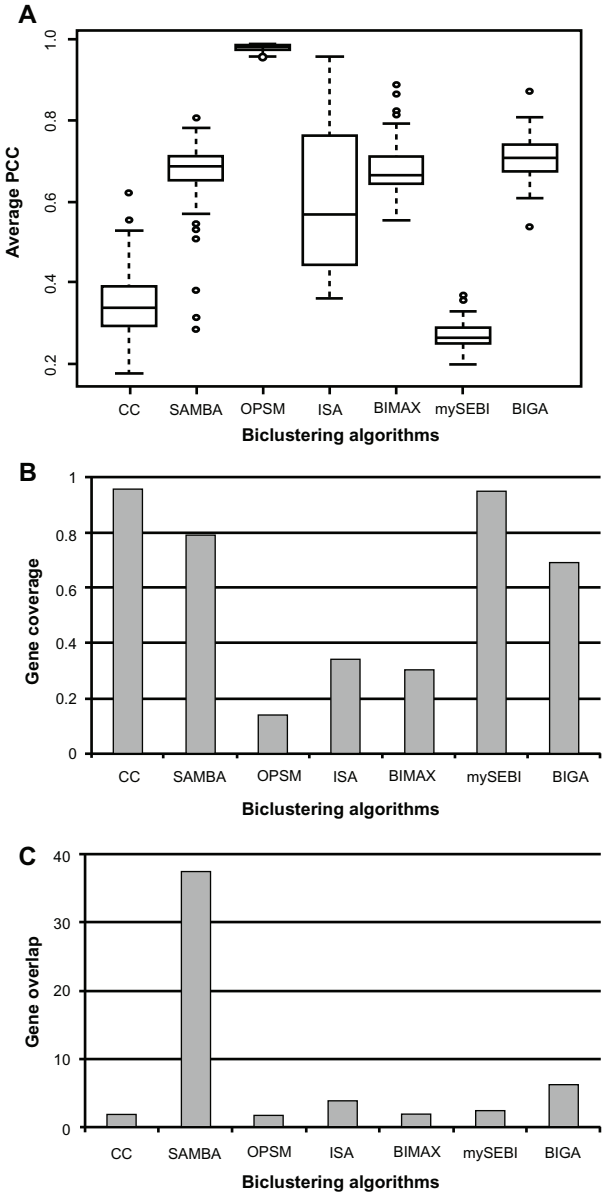

Figure 2 (**A**) Distribution of the average Pearson correlation coefficients for each biclustering algorithm, represented by a boxplot. (**B**) Histogram of gene coverage for each biclustering algorithm. The y-axis represents the coverage ratio between the union of genes appearing on biclusters and all analyzed genes. Higher coverage shows higher performance. (**C**) Histogram of gene overlap for each biclustering algorithm. The y-axis shows the gene overlap defined by (Equation 2). Lower overlap shows higher performance.

Abbreviations: CC, Cheng and Church algorithm; SAMBA, Statistical-Algorithmic Method for Bicluster Analysis; OPSM, order-preserving submatrix; ISA, iterative signature algorithm; BIMAX, binary inclusion-maximal biclustering algorithm; mySEBI, the Sequential Evolutionary Biclustering method used in this work; BIGA, binary-iterative genetic algorithm.

algorithms (Figure 2B). SAMBA could classify almost 100% of the genes into biclusters, but each bicluster contained more than 900 genes (Table 1) with extremely high overlap (Figure 2C), which will make the succeeding experimental or bioinformatics analyses difficult. mySEBI could produce a set of biclusters that would include 95% of all genes with a small amount of overlap. CC showed the best gene coverage (highest) and overlap (lowest). The results indicate that the techniques to reduce redundancy of biclusters in SEBI and CC are efficient for gaining high coverage and low overlap. However, the average PCCs of the biclusters by both algorithms were very low (Figure 2A). OPSM produced biclusters with the highest correlation (Figure 2A), but failed to achieve higher gene coverage due to the small number of clusters (Table 1). The average PCCs of OPSM and BIGA are high, because both methods adopt gene co-expression in the target function. By contrast, CC and SEBI adopt MSR instead of PCC. Although MSR can sometimes identify coherent biclusters, it is not necessarily efficient to achieve higher correlations of genes.

BIGA yielded the second-largest gene overlap, with 6.29 (Figure 2C), which may imply that the biclusters of BIGA are mutually similar. The pairwise overlap (*PO*) of two biclusters defined by $X_i \cap X_j / X_i \cup X_j$, where X_i and X_j are genes in biclusters B_i and B_j, respectively, was measured to examine the similarity of the biclusters more directly, and plotted in Figure 3A. The median of the *PO*s for BIGA was not very large, as compared with those of the other methods, indicating that the biclusters determined by BIGA are not necessarily similar. Moreover, the variety of biclusters using the single-linkage clustering method, where the distance between two biclusters defined by 1.0–*PO* was investigated. At each cut-off distance, the number of clusters was counted and normalized by the total number of biclusters, which we call the fraction of independent biclusters. When the cut-off distance is sufficiently small, no biclusters are merged and *FIB* is 1.0. This state indicates that the biclusters are independent and diverse. On the other hand, when the cut-off distance is sufficiently large, most of the biclusters may be merged together, and *FIB* will

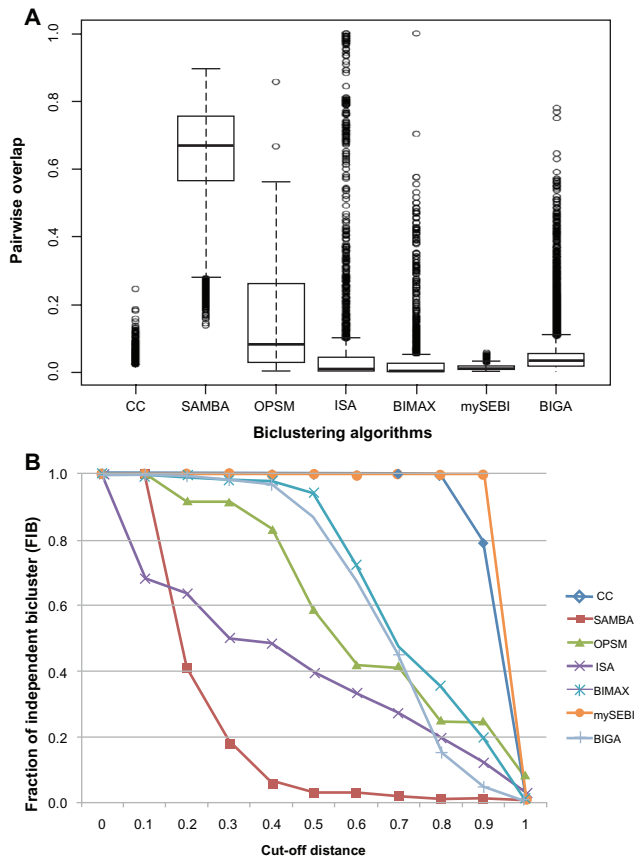

Figure 3 (**A**) Distribution of pairwise overlap (*PO*) of biclusters, shown in boxplots for each algorithm. Thick lines, boxes, whiskers, and circles indicate the same things as in (Figure 2A). (**B**) The fraction of independent biclusters (*FIB*) over the cut-off distance.

Abbreviations: CC, Cheng and Church algorithm; SAMBA, Statistical-Algorithmic Method for Bicluster Analysis (SAMBA); OPSM, order-preserving submatrix; ISA, iterative signature algorithm; BIMAX, binary inclusion-maximal biclustering algorithm; mySEBI, the Sequential Evolutionary Biclustering method used in this work; BIGA, binary-iterative genetic algorithm.

converge to 0.0. This state means that all of the biclusters are judged as being similar to each other. We consider a higher *FIB* to be an indicator illustrating the variety of the resultant biclusters. According to the plot (Figure 3B), the *FIB*s of SAMBA and ISA are obviously low in almost the whole cut-off distance range, showing that their biclusters are rather similar. The *FIB*s of OPSM show that its ability to detect diverse biclusters is moderate. CC, mySEBI, BIMAX, and BIGA provided a wider variety of biclusters than the other algorithms, when the cut-off distance was less than 0.5. In summary, the average bicluster determined by BIGA contains many genes that are shared with other biclusters (Figure 2C): however, when focusing on each pair of biclusters, a small number of genes are shared (Figure 3A). Consequently, the biclusters determined by BIGA seem to be independent (Figure 3B), and cover most of the genes efficiently (Figure 2B).

Evaluation of biological relevance by gene ontology enrichment analyses

In the study by Prelic et al[21] on the evaluation of existing methods using GO enrichment, OPSM showed the best performance (100% of the biclusters were significant at the 0.05 significance level). However, it only produced twelve biclusters (Table 1), and thus the gene coverage was the lowest (Figure 2B). Less than half of the biclusters produced by CC were judged to be significant,[21] probably because CC cannot detect biclusters with a higher average PCC (Figure 2A). The percentages of significant biclusters from mySEBI are 93%, 81%, 69%, and 42% for the 0.05, 0.01, 0.005, and 0.001, respectively. By contrast, 94.5% of the biclusters produced by BIGA were judged to be significant at the 0.05 significance level. This value was changed to 88.4%, 86.0%, and 79.3% for the 0.01, 0.005, and 0.001 significance levels, respectively. The performance of BIGA is almost the same as those of BIMAX and ISA in GO enrichment,[21] but BIGA outperforms them in the gene coverage (Figure 2B).

There was a functional relationship between the resultant biclusters by BIGA, based on the enriched GO terms at the 0.001 significance level. Among the 122 GO-enriched terms, ribosome-related terms (ribosome GO:0005840, ribosomal subunit GO:0033279, etc) are abundant in many biclusters (50 biclusters). This observation was consistent with the fact that 60% of transcription was devoted to ribosomal ribonucleic acid (RNA),[32] because genes with higher expression levels tend to be clustered. Apart from the ribosome-related terms, primary metabolic (GO:0044238), translation (GO:0006412), protein-related (GO:0044267, GO:0019538), macromolecule-related (GO:0009059, GO:0034645, GO:0044260, GO:0043170), and biopolymer-related (GO:0043283, GO:0034960, GO:0043284, GO:0034961) processes also frequently appeared in several biclusters. This indicated that the genes involved in these terms are primary or essential in many biological processes. Five GO terms that are most enriched at the 0.001 significance level for each bicluster five specific GO terms among them are shown in Table S2.

Furthermore, the novel aspects of the biclusters identified by BIGA were examined. For each bicluster defined by BIGA, the *PO* against all biclusters identified by the other five methods was measured and the maximum *PO* was derived (Table S2). The highest value of the maximum *POs* was at most 0.12, indicating that the biclusters defined by BIGA are quite different from those determined by the other methods. To explore the relationships of the genes that were detected

only by BIGA, on the study examined the biclusters of BIGA that were not similar to any of the other biclusters; that is, the biclusters with maximum pair-wise similarity scores < 0.05. In bicluster 109 (the maximum $PO = 0.039$ with bicluster 29 of CC), 16 out of 86 genes are involved in a cellular nitrogen metabolic process (GO:0034641), eg, *SAS3* (YBL052C), *TEF2* (YBR118W), and *SWD3* (YBR175W), are co-expressed under twelve conditions. In bicluster 118 (0.037 with bicluster 56 of CC), 26 out of 66 genes, eg, *RRN6* (YBL014C), *ORC2* (YBR060C), and *PAF1* (YBR279W), are involved in an RNA metabolic process (GO:0016070). In bicluster 160 (0.037, bicluster 24 of ISA), 33 out of 74 genes, such as *HEK2* (YBL032W), *ROX3* (YBL093C), and *SIF2* (YBR103W), are related to a nucleic acid metabolic process (GO:0090304). These results demonstrate that BIGA is useful to reveal the functional relevance underlying the biclusters. Furthermore, some genes belonged to the same bicluster, even though they lacked known co-functional evidence (see the biclusters in Table S2 without significant GO terms). These genes represent promising experimental targets that bridge biological processes exhibiting co-expression under specific conditions.

Conclusion

The development of biclustering algorithms has allowed biologists to start unraveling the underlying functional mechanisms in living organisms. We propose BIGA as an alternative biclustering technique, since it was designed to address the conventional problems of the pre-existing methods. Biclustering is obviously advantageous in accounting for the overlap state among clusters, but the suitable amount of overlap is still ambiguous and different algorithms often produce solutions with various degrees of overlap. We tried to develop a novel chromosome-encoding mode that explicitly defines the overlap between biclusters. BIGA revealed that the most frequently appearing genes express their functions in fundamental and essential biological processes, such as translation. A microarray often consists of relatively few conditions, with respect to a large number of genes. The weighting of genes and conditions diminishes the bias between the number of genes and conditions, which helps to eliminate unreliable results, such as biclusters with very few conditions. We also applied an alternative index, the average PCC, which impacts the biological meaning, rather than the MSR, to measure the goodness of a bicluster. The analysis of GO enrichment demonstrated that most of our biclusters were significant, with one or more enriched GO terms. When evaluated with the five pre-existing algorithms, BIGA performed well in most of the properties with good balance, although it did not show the best performance for all criteria. A pair-wise comparison of our biclusters with those obtained by the other algorithms revealed the novel aspects of the biclusters that are distinct from those of the other methods. Since biological systems are quite complicated, resulting in high-dimensional data, it is quite difficult to answer all biological questions with a single approach. For new discoveries, we recommend the application of several approaches, including BIGA.

Acknowledgments

We would like to thank the Human Genome Center for providing computational resources to analyze all of the data, as well as for a scholarship from the Ministry of Education, Culture, Sports, Science and Technology to Sawannee Sutheeworapong. We would like to acknowledge Prof Kenta Nakai for providing good facilities to Sawannee Sutheeworapong in the early stage of this work. We also thank Dr Takeshi Obayashi for useful discussions in the early stage of this work.

Authors' contributions

SS, KK, and MO contributed to the overall research and the manuscript preparation. KK, MO, and HO were responsible for the project direction and financial support.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Ben-Dor A, Chor B, Karp R, Yakhini Z. Discovering local structure in gene expression data: the order-preserving submatrix problem. *J Comput Biol*. 2003;10:373–384.
2. Ma X, Salunga R, Tuggle T, et al. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci U S A*. 2003;100: 5974–5979.
3. Yamane D, Zahoor MA, Mohamed YM, et al. Microarray analysis reveals distinct signaling pathways transcriptionally activated by infection with bovine viral diarrhea virus in different cell types. *Virus Res*. 2009;142(1–2):188–199.
4. Wang RS, Wang Y, Zhang XS, Chen L. Inferring transcriptional regulatory networks from high-throughput data. *Bioinformatics*. 2007;23(22):3056–3064.
5. Hartigan JA, Wong MA. A k-means clustering algorithm. *Appl Stat*. 1979;28:100–108.
6. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bull*. 1958;38:1409–1438.
7. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM. Trans Computat Biol Bioinform*. 2004;1(1):24–45.
8. Hartigan JA. Direct clustering of a data matrix. *J Am Stat Assoc*. 1972;67(337):123–129.
9. Cheng Y, Church GM. Biclustering of expression data. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. ISMB 2000, San Diego, CA; August 19–12, 2000. AAAI Press; 2000:93–103.

10. Murali TM, Kasif S. Extracting conserved gene expression motifs from gene expression data. Proceedings of the Pacific Symposium on Biocomputing. PSB 2003, Lihue, HI, January 3–7, 2003. 2003;8:77–88.
11. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E Stat Nonlin Soft Matter Phys*. 2003;67:031902.
12. Peeters R. The maximum edge biclique problem in NP-complete. *Discrete Appl Math*. 2003;131(3):651–654.
13. Merz P, Zell A. *Genetic Algorithms and Grouping Problems*. Philadelphia, PA: John Wiley & Sons; 1998.
14. Bleuler S, Prelic A, Zitzler E. An EA framework for biclustering of gene expression data. Proceedings of Congress on Evolutionary Computation, Portland, OR, June 19–23. June 19–23, 2004;4: 166–173.
15. Chakraborty A, Maka H. Biclustering of gene expression data using genetic algorithm. Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2005. CIBCB 2005. La Jolla, CA, November 14–14, 2005:1–8.
16. Divina F, Aguilar-ruiz JS. Biclustering of expression data with evolutionary computation. *IEEE Trans Knowl Data Eng*. 2006;18:590–602.
17. Donale K. *The Art of Computer Programming, vol 1: Fundamental algorithms*. 3rd ed Boston, MA: Addison-Wesley. 1997;Section 2.3:318–348.
18. Meffert K, Rotstan N. JGAP-Java Genetic Algorithms and Genetic Programming Package. Available from: http://jgap.sf.nt/. Accessed July 10, 2012.
19. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*. 2002;18:S136–S144.
20. Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A*. 2003;102(9):2981–2986.
21. Prelic A, Bleuler S, Zimmermann P, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*. 2006;22(9):1122–1129.
22. Gasch AP, Spellman PT, Kao CM, et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*. 2000;11:4241–4257.
23. Aguilar-Ruiz J. Shifting and scaling patterns from gene expression data. *Bioinformatics*. 2005;21:3840–3845.
24. Pontes B, Divina F, Giraldez R, Aguilar-Ruiz JS. Virtual error: a new measure for evolutionary biclustering. *Evol Comput, Machine Learning and Data Mining in Bioinformatics*. 2007;4447:217–226.
25. Teng L, Chan L. Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. *J Signal Process Syst*. 2008;50(3):267–280.
26. Ayadi W, Elloumi M, Hao J. A biclustering algorithm based on a bicluster enumeration tree: application to DNA microarray data. *Bio Data Min*. 2009;2:9.
27. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*. 1998;23:324–328.
28. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998;95:14863–14868.
29. Waltman P, Kacmarczyk T, Bate AR, et al. Multi-species integrative biclustering. *Genome Biol*. 2010;11:R96.
30. Berriz GF, King OD, Bryant B, Sander C, Roth FP. Characterizing gene sets with FuncAssociate. *Bioinformatics*. 2003;19:2502–2504.
31. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull*. 1945;1(6):80–83.
32. Warner JR. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci*. 1999;24:437–440.

Supplementary data

Table S1 Parameter determination

	Goodness of biclusters					
	Genes	**Conditions**	**Correlation**	**Biclusters**	**Coverage**	**Overlap**
α						
0.1	72.15	22.84	0.74	111	0.59	3.53
0.3	92.25	23.65	0.71	164	0.69	6.29
0.5	102.22	24.42	0.7	252	0.67	11.82
τ_r						
0.1	81.22	21.51	0.73	355	0.74	11.97
0.15	92.25	23.65	0.71	164	0.69	6.29
0.2	109.86	25.07	0.69	57	0.58	2.59
0.25	128.13	32.5	0.71	8	0.22	0.53
0.3	163	45	0.67	1	0.05	0
τ_c						
0.60	100.62	22.17	0.69	145	0.71	5.9
0.65	92.25	23.65	0.71	164	0.69	6.29
0.70	83.84	22.69	0.74	178	0.61	7.09

Notes: (**A**) Impact of gene-weight parameter on the goodness of biclusters ($\tau_n = 30$, $\tau_m = 10$, $\tau_c = 0.65$, $\tau_r = 0.15$ and $\beta = 0.5$). (**B**) Impact of redundant threshold on the goodness of biclusters ($\tau_n = 30$, $\tau_m = 10$, $\tau_c = 0.65$, and $\alpha = 0.3$, $\beta = 0.5$). (**C**) Impact of correlation threshold on the goodness of biclusters ($\tau_n = 30$, $\tau_m = 10$, $\tau_c = 0.15$, and $\alpha = 0.3$, $\beta = 0.5$).

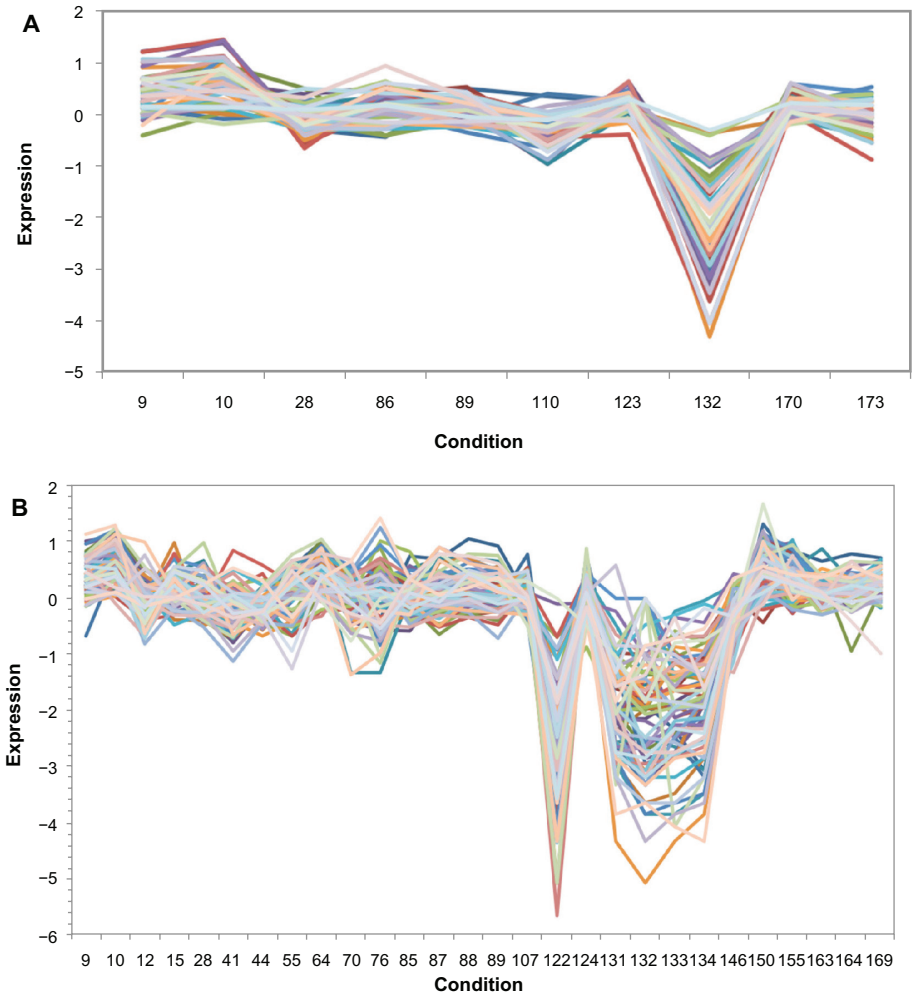

Figure S1 (*Continued*)

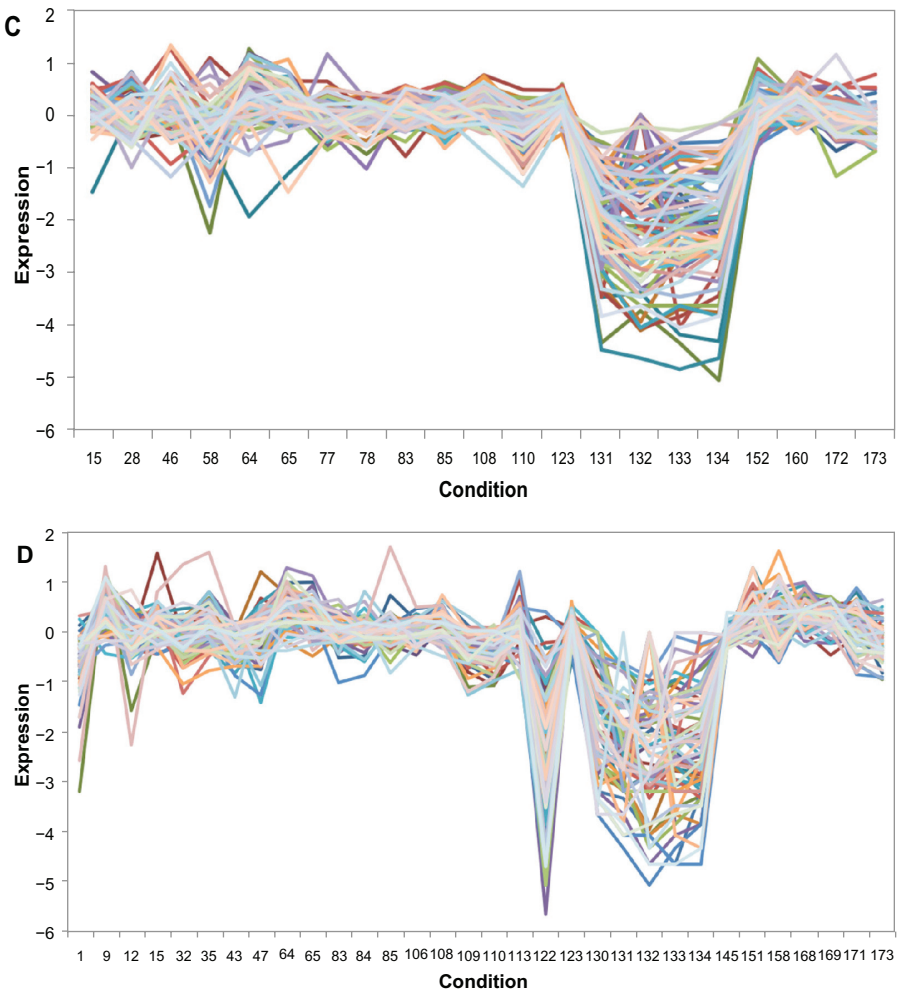

Figure S1 Expression profiles of biclusters 1 (**A**), 2 (**B**), 3 (**C**), and 4 (**D**), in the descending order of the average Pearson's correlation coefficient.
Note: The x-axis represents the series of conditions; eg, the number 8 denotes the 8th condition.

Table S2 Detailed statistics of resulting biclusters (sorted by descending order of average PCC)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
1	47	10	0.87	<0.001	2
2	74	28	0.81	<0.001	3
3	85	21	0.80	<0.001	14
4	71	32	0.80		12
5	74	18	0.80	0.001	1
6	50	7	0.80	–	0
7	79	24	0.80	<0.001	8
8	52	16	0.79	–	0
9	56	4	0.79	–	0
10	87	21	0.79	<0.001	5
11	72	20	0.79	<0.001	5
12	78	26	0.79	<0.001	6
13	74	14	0.79	<0.001	1
14	83	33	0.78	<0.001	19
15	86	23	0.78	<0.001	2
16	49	18	0.78	<0.001	10

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0003674 molecular_function GO:0032991 macromolecular complex	–	0.044
GO:0003674 molecular_function GO:0032991 macromolecular complex GO:0043234 protein complex	–	0.067
GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0044238 primary metabolic process GO:0032991 macromolecular complex GO:0022618 ribonucleoprotein complex assembly	GO:0007114 cell budding GO:0022618 ribonucleoprotein complex assembly GO:0032505 reproduction of a single-celled organism GO:0042257 ribosomal subunit assembly GO:0043933 macromolecular complex subunit organization	0.070
GO:0030529 ribonucleoprotein complex GO:0032991 macromolecular complex GO:0005840 ribosome GO:0044445 cytosolic part GO:0006412 translation	GO:0022625 cytosolic large ribosomal subunit	0.093
GO:0005737 cytoplasm	GO:0005737 cytoplasm	0.050
–	–	0.043
GO:0044238 primary metabolic process GO:0032991 macromolecular complex GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0005840 ribosome	GO:0009072 aromatic amino acid family metabolic process	0.073
–	–	0.032
–	–	0.041
GO:0003674 molecular_function GO:0006412 translation GO:0009987 cellular process GO:0009058 biosynthetic process GO:0044249 cellular biosynthetic process	GO:0044249 cellular biosynthetic process GO:0009058 biosynthetic process	0.068
GO:0032991 macromolecular complex GO:0003674 molecular_function GO:0009987 cellular process GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle	–	0.060
GO:0032040 small-subunit processome GO:0030686 90S preribosome GO:0042254 ribosome biogenesis GO:0030684 preribosome GO:0022613 ribonucleoprotein complex biogenesis	GO:0032040 small-subunit processome GO:0022613 ribonucleoprotein complex biogenesis GO:0042254 ribosome biogenesis GO:0030684 preribosome GO:0030686 90S preribosome	0.074
GO:0003674 molecular_function	–	0.048
GO:0044445 cytosolic part GO:0006412 translation GO:0022625 cytosolic large ribosomal subunit GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle	GO:0015934 large ribosomal subunit GO:0022625 cytosolic large ribosomal subunit GO:0044249 cellular biosynthetic process GO:0009058 biosynthetic process	0.080
GO:0003674 molecular_function GO:0032991 macromolecular complex	–	0.056
GO:0044238 primary metabolic process GO:0016070 RNA metabolic process GO:0044260 cellular macromolecule metabolic process GO:0043283 biopolymer metabolic process GO:0030529 ribonucleoprotein complex	GO:0008152 metabolic process GO:0016070 RNA metabolic process GO:0034960 cellular biopolymer metabolic process GO:0044260 cellular macromolecule metabolic process GO:0044237 cellular metabolic process	0.059

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
17	92	23	0.78	<0.001	12
18	77	25	0.78	<0.001	4
19	77	21	0.78	<0.001	5
20	59	12	0.78	<0.001	1
21	84	30	0.77	<0.001	10
22	53	11	0.77	0.001	1
23	81	28	0.77	<0.001	11
24	61	21	0.77	–	0
25	82	13	0.77	<0.001	1
26	103	24	0.76	<0.001	9
27	93	27	0.76	<0.001	19
28	65	11	0.76	<0.001	1
29	78	32	0.76	<0.001	2
30	62	19	0.76	<0.001	6
31	89	19	0.76	<0.001	12

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0044238 primary metabolic process GO:0032991 macromolecular complex GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0034621 cellular macromolecular complex subunit organization	GO:0034621 cellular macromolecular complex subunit organization GO:0034660 ncRNA metabolic process GO:0006139 "nucleobase, nucleoside, nucleotide and nucleic acid metabolic process" GO:0016070 RNA metabolic process GO:0044237 cellular metabolic process	0.072
GO:0003674 molecular_function GO:0044445 cytosolic part GO:0009987 cellular process GO:0032991 macromolecular complex	–	0.050
GO:0003674 molecular_function GO:0032991 macromolecular complex GO:0044238 primary metabolic process GO:0030529 ribonucleoprotein complex GO:0015935 small ribosomal subunit	GO:0015935 small ribosomal subunit	0.062
GO:0044238 primary metabolic process	–	0.046
GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0005840 ribosome	GO:0005737 cytoplasm	0.073
GO:0044238 primary metabolic process	–	0.058
GO:0032991 macromolecular complex GO:0043283 biopolymer metabolic process GO:0034960 cellular biopolymer metabolic process GO:0043234 protein complex GO:0043170 macromolecule metabolic process	GO:0051246 regulation of protein metabolic process GO:0034960 cellular biopolymer metabolic process GO:0044260 cellular macromolecule metabolic process GO:0032268 regulation of cellular protein metabolic process GO:0043234 protein complex	0.059
–	–	0.039
GO:0003674 molecular_function	–	0.045
GO:0044238 primary metabolic process GO:0003674 molecular_function GO:0009987 cellular process GO:0005840 ribosome GO:0003735 structural constituent of ribosome GO:0045182 translation regulator activity	GO:0003743 translation initiation factor activity GO:0045182 translation regulator activity GO:0008135 "translation factor activity, nucleic acid binding" GO:0032268 regulation of cellular protein metabolic process GO:0043234 protein complex	0.077
GO:0044238 primary metabolic process GO:0003735 structural constituent of ribosome GO:0009987 cellular process GO:0005840 ribosome GO:0003735 structural constituent of ribosome	GO:0015935 small ribosomal subunit GO:0008152 metabolic process GO:0043229 intracellular organelle GO:0043226 organelle GO:0022627 cytosolic small ribosomal subunit	0.098
GO:0003674 molecular_function	–	0.045
GO:0003674 molecular_function GO:0032991 macromolecular complex	–	0.077
GO:0009058 biosynthetic process GO:0044249 cellular biosynthetic process GO:0044238 primary metabolic process GO:0032991 macromolecular complex GO:0044445 cytosolic part	GO:0044249 cellular biosynthetic process GO:0009058 biosynthetic process	0.056
GO:0009058 biosynthetic process GO:0044249 cellular biosynthetic process GO:0043284 biopolymer biosynthetic process GO:0009059 macromolecule biosynthetic process GO:0044238 primary metabolic process	GO:0006139 "nucleobase, nucleoside, nucleotide and nucleic acid metabolic process" GO:0034961 cellular biopolymer biosynthetic process GO:0034645 cellular macromolecule biosynthetic process GO:0016070 RNA metabolic process GO:0009059 macromolecule biosynthetic process	0.063

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
32	91	30	0.76	<0.001	10
33	105	34	0.76	<0.001	8
34	105	28	0.75	<0.001	16
35	110	25	0.75	<0.001	29
36	66	16	0.75	<0.001	8
37	71	10	0.75	0.001	1
38	59	14	0.74	<0.001	3
39	58	16	0.74	<0.001	13
40	83	36	0.74	<0.001	8
41	78	23	0.74	<0.001	5
42	113	26	0.74	<0.001	23

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0017111 nucleoside-triphosphatase activity GO:0016462 pyrophosphatase activity GO:0016817 "hydrolase activity, acting on acid anhydrides" GO:0016818 "hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides" GO:0044238 primary metabolic process	GO:0017111 nucleoside-triphosphatase activity GO:0016462 pyrophosphatase activity GO:0016817 "hydrolase activity, acting on acid anhydrides" GO:0016818 "hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides" GO:0034470 ncRNA processing	0.081
GO:0009058 biosynthetic process GO:0032991 macromolecular complex GO:0009987 cellular process GO:0006412 translation GO:0044445 cytosolic part	GO:0009058 biosynthetic process	0.098
GO:0032991 macromolecular complex GO:0044267 cellular protein metabolic process GO:0006412 translation GO:0009987 cellular process GO:0043234 protein complex	GO:0044444 cytoplasmic part GO:0044424 intracellular part GO:0043234 protein complex GO:0009058 biosynthetic process	0.088
GO:0032991 macromolecular complex GO:0016070 RNA metabolic process GO:0044238 primary metabolic process GO:0009987 cellular process GO:0005198 structural molecule activity	GO:0019438 aromatic compound biosynthetic process GO:0006396 RNA processing GO:0034470 ncRNA processing GO:0034660 ncRNA metabolic process GO:0006139 "nucleobase, nucleoside, nucleotide and nucleic acid metabolic process"	0.085
GO:0032991 macromolecular complex GO:0003735 structural constituent of ribosome GO:0033279 ribosomal subunit GO:0005198 structural molecule activity GO:0006412 translation	GO:0022627 cytosolic small ribosomal subunit	0.069
GO:0044085 cellular component biogenesis	GO:0044085 cellular component biogenesis	0.068
GO:0003674 molecular_function GO:0005198 structural molecule activity GO:0032991 macromolecular complex	–	0.040
GO:0044249 cellular biosynthetic process GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0009058 biosynthetic process GO:0043284 biopolymer biosynthetic process	GO:0000462 "maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)" GO:0030490 maturation of SSU-rRNA GO:0034961 cellular biopolymer biosynthetic process GO:0034645 cellular macromolecule biosynthetic process GO:0022627 cytosolic small ribosomal subunit	0.048
GO:0044445 cytosolic part GO:0006412 translation GO:0043229 intracellular organelle GO:0043226 organelle GO:0043228 nonmembrane-bounded organelle	GO:0043229 intracellular organelle GO:0043226 organelle	0.076
GO:0032991 macromolecular complex GO:0043234 protein complex GO:0003674 molecular_function GO:0044238 primary metabolic process GO:0009987 cellular process	GO:0043234 protein complex	0.069
GO:0044445 cytosolic part GO:0030529 ribonucleoprotein complex GO:0005198 structural molecule activity GO:0033279 ribosomal subunit GO:0006412 translation	GO:0006913 nucleocytoplasmic transport GO:0051169 nuclear transport GO:0005622 intracellular GO:0005737 cytoplasm GO:0010608 posttranscriptional regulation of gene expression	0.080

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
43	90	22	0.74	<0.001	18
44	89	25	0.74	<0.001	6
45	92	28	0.74	<0.001	8
46	106	28	0.74	<0.001	12
47	106	36	0.74	<0.001	14
48	109	25	0.74	<0.001	23
49	99	27	0.74	<0.001	24
50	89	24	0.73	<0.001	10
51	86	15	0.73	<0.001	3
52	141	35	0.73	<0.001	18
53	107	31	0.73	<0.001	20

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0032991 macromolecular complex GO:0022627 cytosolic small ribosomal subunit GO:0030684 preribosome GO:0030686 90S preribosome GO:0030529 ribonucleoprotein complex	GO:0044249 cellular biosynthetic process GO:0009058 biosynthetic process	0.081
GO:0044238 primary metabolic process GO:0032991 macromolecular complex GO:0009987 cellular process GO:0034621 cellular macromolecular complex subunit organization GO:0016070 RNA metabolic process	GO:0034621 cellular macromolecular complex subunit organization GO:0016070 RNA metabolic process	0.061
GO:0019538 protein metabolic process GO:0044267 cellular protein metabolic process GO:0032268 regulation of cellular protein metabolic process GO:0005737 cytoplasm GO:0051246 regulation of protein metabolic process	GO:0005737 cytoplasm GO:0010608 posttranscriptional regulation of gene expression GO:0051246 regulation of protein metabolic process GO:0006417 regulation of translation GO:0032268 regulation of cellular protein metabolic process	0.057
GO:0009987 cellular process GO:0006412 translation GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0044238 primary metabolic process	GO:0022627 cytosolic small ribosomal subunit	0.089
GO:0030529 ribonucleoprotein complex GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0005840 ribosome GO:0032991 macromolecular complex	GO:0016462 pyrophosphatase activity GO:0016817 "hydrolase activity, acting on acid anhydrides" GO:0016818 "hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides"	0.100
GO:0032991 macromolecular complex GO:0044238 primary metabolic process GO:0044445 cytosolic part GO:0009987 cellular process GO:0005840 ribosome	GO:0005622 intracellular GO:0022625 cytosolic large ribosomal subunit GO:0010608 posttranscriptional regulation of gene expression GO:0051246 regulation of protein metabolic process GO:0006417 regulation of translation	0.083
GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0005840 ribosome GO:0005198 structural molecule activity GO:0006412 translation	GO:0034961 cellular biopolymer biosynthetic process GO:0034645 cellular macromolecule biosynthetic process GO:0022627 cytosolic small ribosomal subunit GO:0034960 cellular biopolymer metabolic process GO:0009059 macromolecule biosynthetic process	0.082
GO:0030529 ribonucleoprotein complex GO:0032991 macromolecular complex GO:0044238 primary metabolic process GO:0005840 ribosome GO:0043228 nonmembrane-bounded organelle	GO:0005488 binding	0.074
GO:0003674 molecular_function GO:0009987 cellular process GO:0000166 nucleotide binding	GO:0000166 nucleotide binding	0.065
GO:0006412 translation GO:0032991 macromolecular complex GO:0009058 biosynthetic process GO:0009987 cellular process GO:0044249 cellular biosynthetic process	GO:0006082 organic acid metabolic process GO:0019752 carboxylic acid metabolic process GO:0005737 cytoplasm GO:0009059 macromolecule biosynthetic process GO:0043284 biopolymer biosynthetic process	0.119
GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0005198 structural molecule activity	GO:0007010 cytoskeleton organization GO:0015935 small ribosomal subunit GO:0022627 cytosolic small ribosomal subunit GO:0006417 regulation of translation GO:0032268 regulation of cellular protein metabolic process	0.062

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
54	68	24	0.73	0.001	6
55	128	26	0.73	<0.001	21
56	101	32	0.73	<0.001	15
57	107	32	0.73	<0.001	11
58	111	33	0.72	<0.001	11
59	92	27	0.72	<0.001	11
60	111	33	0.72	<0.001	7
61	76	15	0.72	<0.001	2
62	94	20	0.72	<0.001	6
63	83	24	0.72	<0.001	13
64	126	28	0.72	<0.001	39
65	45	12	0.72	–	0

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0009987 cellular process GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0043229 intracellular organelle GO:0043226 organelle	GO:0043229 intracellular organelle GO:0043226 organelle	0.045
GO:0032991 macromolecular complex GO:0006412 translation GO:0044267 cellular protein metabolic process GO:0019538 protein metabolic process GO:0044238 primary metabolic process	GO:0016043 cellular component organization GO:0065007 biological regulation GO:0050789 regulation of biological process GO:0050794 regulation of cellular process GO:0009059 macromolecule biosynthetic process	0.089
GO:0032991 macromolecular complex GO:0030529 ribonucleoprotein complex GO:0044445 cytosolic part GO:0009987 cellular process GO:0005840 ribosome	GO:0022625 cytosolic large ribosomal subunit GO:0044424 intracellular part	0.099
GO:0032991 macromolecular complex GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0044238 primary metabolic process GO:0009987 cellular process	GO:0043170 macromolecule metabolic process	0.091
GO:0032991 macromolecular complex GO:0009987 cellular process GO:0019538 protein metabolic process GO:0006412 translation GO:0043228 nonmembrane-bounded organelle	GO:0043234 protein complex	0.099
GO:0009987 cellular process GO:0044238 primary metabolic process GO:0032991 macromolecular complex GO:0032268 regulation of cellular protein metabolic process GO:0044445 cytosolic part	GO:0010608 posttranscriptional regulation of gene expression GO:0016070 RNA metabolic process GO:0051246 regulation of protein metabolic process GO:0006417 regulation of translation GO:0044424 intracellular part	0.106
GO:0032991 macromolecular complex GO:0009987 cellular process GO:0044445 cytosolic part GO:0044238 primary metabolic process GO:0006412 translation	–	0.078
GO:0003674 molecular_function GO:0009987 cellular process	–	0.050
GO:0032991 macromolecular complex GO:0032268 regulation of cellular protein metabolic process GO:0044238 primary metabolic process GO:0051246 regulation of protein metabolic process GO:0009987 cellular process	GO:0051246 regulation of protein metabolic process GO:0032268 regulation of cellular protein metabolic process	0.057
GO:0022627 cytosolic small ribosomal subunit GO:0032991 macromolecular complex GO:0015935 small ribosomal subunit GO:0044445 cytosolic part GO:0030686 90S preribosome	GO:0030686 90S preribosome GO:0015935 small ribosomal subunit GO:0044422 organelle part GO:0044446 intracellular organelle part GO:0022627 cytosolic small ribosomal subunit	0.083
GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0044238 primary metabolic process GO:0005840 ribosome GO:0030529 ribonucleoprotein complex	GO:0015934 large ribosomal subunit GO:0044464 cell part GO:0034961 cellular biopolymer biosynthetic process GO:0034645 cellular macromolecule biosynthetic process GO:0022625 cytosolic large ribosomal subunit	0.094
–	–	0.045

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
66	100	32	0.72	<0.001	8
67	124	29	0.72	<0.001	15
68	111	37	0.72	<0.001	9
69	51	21	0.71	–	0
70	106	30	0.71	<0.001	21
71	46	12	0.71	–	0
72	126	36	0.71	<0.001	17
73	87	25	0.71	<0.001	8
74	112	30	0.71	<0.001	18
75	116	31	0.71	<0.001	13
76	68	14	0.71	<0.001	7
77	86	20	0.71	<0.001	3
78	104	39	0.71	<0.001	23

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0005198 structural molecule activity GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0006412 translation GO:0009987 cellular process	–	0.080
GO:0032991 macromolecular complex GO:0043234 protein complex GO:0009058 biosynthetic process GO:0009987 cellular process GO:0043284 biopolymer biosynthetic process	GO:0010608 posttranscriptional regulation of gene expression GO:0006417 regulation of translation GO:0009059 macromolecule biosynthetic process GO:0043284 biopolymer biosynthetic process GO:0044424 intracellular part	0.097
GO:0032991 macromolecular complex GO:0044238 primary metabolic process GO:0006412 translation GO:0009987 cellular process GO:0043228 nonmembrane-bounded organelle	–	0.099
–	–	0.059
GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0044267 cellular protein metabolic process GO:0019538 protein metabolic process GO:0005198 structural molecule activity	GO:0034960 cellular biopolymer metabolic process GO:0009059 macromolecule biosynthetic process GO:0043284 biopolymer biosynthetic process GO:0044260 cellular macromolecule metabolic process GO:0043234 protein complex	0.065
–	–	0.047
GO:0009987 cellular process GO:0044238 primary metabolic process GO:0016462 pyrophosphatase activity GO:0016817 "hydrolase activity, acting on acid anhydrides" GO:0016818 "hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides"	GO:0017076 purine nucleotide binding GO:0032553 ribonucleotide binding GO:0032555 purine ribonucleotide binding GO:0000166 nucleotide binding GO:0017111 nucleoside-triphosphatase activity	0.101
GO:0032991 macromolecular complex GO:0009987 cellular process GO:0044238 primary metabolic process GO:0030529 ribonucleoprotein complex GO:0016070 RNA metabolic process	GO:0016070 RNA metabolic process GO:0043170 macromolecule metabolic process	0.070
GO:0032991 macromolecular complex GO:0006412 translation GO:0044238 primary metabolic process GO:0044424 intracellular part GO:0009058 biosynthetic process	GO:0010468 regulation of gene expression GO:0010556 regulation of macromolecule biosynthetic process GO:0010608 posttranscriptional regulation of gene expression GO:0006417 regulation of translation GO:0044424 intracellular part	0.085
GO:0032991 macromolecular complex GO:0005198 structural molecule activity GO:0044445 cytosolic part GO:0044238 primary metabolic process GO:0009987 cellular process	GO:0005737 cytoplasm GO:0043234 protein complex	0.093
GO:0022627 cytosolic small ribosomal subunit GO:0015935 small ribosomal subunit GO:0006412 translation GO:0044445 cytosolic part GO:0003735 structural constituent of ribosome	GO:0015935 small ribosomal subunit GO:0022627 cytosolic small ribosomal subunit	0.074
GO:0003674 molecular_function GO:0022627 cytosolic small ribosomal subunit GO:0032991 macromolecular complex	GO:0022627 cytosolic small ribosomal subunit	0.052
GO:0032991 macromolecular complex GO:0030529 ribonucleoprotein complex GO:0006412 translation GO:0044238 primary metabolic process GO:0008135 "translation factor activity, nucleic acid binding"	GO:0003743 translation initiation factor activity GO:0045182 translation regulator activity GO:0008135 "translation factor activity, nucleic acid binding" GO:0016070 RNA metabolic process GO:0034960 cellular biopolymer metabolic process	0.108

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
79	90	23	0.71	<0.001	9
80	108	36	0.71	<0.001	7
81	90	24	0.71	<0.001	11
82	106	33	0.71	<0.001	21
83	129	31	0.71	<0.001	18
84	129	28	0.71	<0.001	22
85	77	38	0.71	<0.001	12
86	109	28	0.70	<0.001	6
87	78	21	0.70	0.001	8
88	100	24	0.70	<0.001	19

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0006412 translation GO:0044267 cellular protein metabolic process GO:0019538 protein metabolic process GO:0032991 macromolecular complex GO:0005840 ribosome	–	0.060
GO:0032991 macromolecular complex GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0005198 structural molecule activity GO:0030529 ribonucleoprotein complex	–	0.078
GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0022625 cytosolic large ribosomal subunit GO:0044238 primary metabolic process GO:0043283 biopolymer metabolic process	GO:0022625 cytosolic large ribosomal subunit GO:0043234 protein complex GO:0043170 macromolecule metabolic process	0.067
GO:0044238 primary metabolic process GO:0034960 cellular biopolymer metabolic process GO:0009987 cellular process GO:0043283 biopolymer metabolic process GO:0044260 cellular macromolecule metabolic process	GO:0006139 "nucleobase, nucleoside, nucleotide and nucleic acid metabolic process" GO:0008152 metabolic process GO:0043229 intracellular organelle GO:0043226 organelle GO:0034960 cellular biopolymer metabolic process	0.084
GO:0032991 macromolecular complex GO:0006412 translation GO:0005198 structural molecule activity GO:0005840 ribosome GO:0044445 cytosolic part	GO:0005488 binding GO:0005622 intracellular GO:0022625 cytosolic large ribosomal subunit GO:0044422 organelle part GO:0044446 intracellular organelle part	0.091
GO:0032991 macromolecular complex GO:0044445 cytosolic part GO:0009058 biosynthetic process GO:0044249 cellular biosynthetic process GO:0006412 translation	GO:0009059 macromolecule biosynthetic process GO:0043284 biopolymer biosynthetic process GO:0044424 intracellular part GO:0044237 cellular metabolic process GO:0044249 cellular biosynthetic process	0.098
GO:0030529 ribonucleoprotein complex GO:0044445 cytosolic part GO:0032991 macromolecular complex GO:0033279 ribosomal subunit GO:0043228 nonmembrane-bounded organelle	GO:0005622 intracellular GO:0022625 cytosolic large ribosomal subunit	0.074
GO:0009987 cellular process GO:0006412 translation GO:0044445 cytosolic part GO:0044238 primary metabolic process		0.090
GO:0010468 regulation of gene expression GO:0010556 regulation of macromolecule biosynthetic process GO:0060255 regulation of macromolecule metabolic process GO:0031326 regulation of cellular biosynthetic process GO:0009889 regulation of biosynthetic process	GO:0019222 regulation of metabolic process GO:0060255 regulation of macromolecule metabolic process GO:0009889 regulation of biosynthetic process GO:0031323 regulation of cellular metabolic process GO:0031326 regulation of cellular biosynthetic process	0.055
GO:0032991 macromolecular complex GO:0044238 primary metabolic process GO:0006412 translation GO:0005840 ribosome GO:0003735 structural constituent of ribosome	GO:0044422 organelle part GO:0044446 intracellular organelle part GO:0044260 cellular macromolecule metabolic process GO:0044237 cellular metabolic process	0.073

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
89	82	24	0.70	<0.001	13
90	77	27	0.70	<0.001	5
91	97	22	0.70	<0.001	17
92	110	28	0.70	<0.001	6
93	94	29	0.70	<0.001	15
94	113	34	0.70	<0.001	32
95	94	23	0.70	<0.001	4
96	104	31	0.70	<0.001	10
97	51	13	0.70	<0.001	1
98	154	32	0.70	<0.001	14
99	117	30	0.70	<0.001	11

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0044445 cytosolic part GO:0006417 regulation of translation GO:0010608 posttranscriptional regulation of gene expression GO:0032268 regulation of cellular protein metabolic process GO:0051246 regulation of protein metabolic process	GO:0009889 regulation of biosynthetic process GO:0031323 regulation of cellular metabolic process GO:0031326 regulation of cellular biosynthetic process GO:0010468 regulation of gene expression GO:0010556 regulation of macromolecule biosynthetic process	0.060
GO:0003674 molecular_function GO:0005198 structural molecule activity GO:0009987 cellular process GO:0044238 primary metabolic process GO:0044445 cytosolic part	–	0.050
GO:0044445 cytosolic part GO:0006412 translation GO:0032991 macromolecular complex GO:0009987 cellular process GO:0044238 primary metabolic process	GO:0009059 macromolecule biosynthetic process GO:0043284 biopolymer biosynthetic process GO:0044249 cellular biosynthetic process	0.088
GO:0009987 cellular process GO:0032991 macromolecular complex GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle	–	0.090
GO:0032991 macromolecular complex GO:0006417 regulation of translation GO:0010608 posttranscriptional regulation of gene expression GO:0032268 regulation of cellular protein metabolic process GO:0051246 regulation of protein metabolic process	GO:0005083 small GTPase regulator activity GO:0030695 GTPase regulator activity GO:0005737 cytoplasm GO:0010608 posttranscriptional regulation of gene expression GO:0051246 regulation of protein metabolic process	0.067
GO:0009058 biosynthetic process GO:0044249 cellular biosynthetic process GO:0006412 translation GO:0009987 cellular process GO:0044238 primary metabolic process	GO:0019222 regulation of metabolic process GO:0060255 regulation of macromolecule metabolic process GO:0009889 regulation of biosynthetic process GO:0031323 regulation of cellular metabolic process GO:0031326 regulation of cellular biosynthetic process	0.075
GO:0006412 translation GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle	–	0.062
GO:0006412 translation GO:0009987 cellular process GO:0044238 primary metabolic process GO:0016070 RNA metabolic process GO:0034660 ncRNA metabolic process	GO:0034470 ncRNA processing GO:0034660 ncRNA metabolic process GO:0016070 RNA metabolic process	0.107
GO:0003674 molecular_function	–	0.043
GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0005840 ribosome GO:0032991 macromolecular complex GO:0030529 ribonucleoprotein complex	GO:0005575 cellular_component GO:0044464 cell part GO:0010556 regulation of macromolecule biosynthetic process GO:0005737 cytoplasm GO:0010608 posttranscriptional regulation of gene expression	0.115
GO:0005622 intracellular GO:0009987 cellular process GO:0044238 primary metabolic process GO:0006412 translation GO:0019538 protein metabolic process	GO:0005622 intracellular GO:0022627 cytosolic small ribosomal subunit GO:0032268 regulation of cellular protein metabolic process	0.100

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
100	110	28	0.70	<0.001	8
101	139	34	0.70	<0.001	16
102	98	28	0.69	<0.001	48
103	71	18	0.69	<0.001	1
104	105	21	0.69	<0.001	5
105	140	32	0.69	<0.001	16
106	41	12	0.69	–	0
107	101	25	0.69	<0.001	24
108	99	21	0.69	<0.001	9
109	86	12	0.69	<0.001	7
110	118	30	0.69	<0.001	17
111	98	15	0.69	<0.001	5
112	157	43	0.69	<0.001	38

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0009987 cellular process	GO:0044249 cellular biosynthetic process	0.092
GO:0032991 macromolecular complex		
GO:0044238 primary metabolic process		
GO:0044445 cytosolic part		
GO:0009058 biosynthetic process		
GO:0005198 structural molecule activity	GO:0044422 organelle part	0.100
GO:0032991 macromolecular complex	GO:0044446 intracellular organelle part	
GO:0044445 cytosolic part	GO:0051246 regulation of protein metabolic process	
GO:0006412 translation	GO:0006417 regulation of translation	
GO:0009987 cellular process	GO:0032268 regulation of cellular protein metabolic process	
GO:0032991 macromolecular complex	GO:0006333 chromatin assembly or disassembly	0.085
GO:0044238 primary metabolic process	GO:0006446 regulation of translational initiation	
GO:0006412 translation	GO:0003743 translation initiation factor activity	
GO:0043284 biopolymer biosynthetic process	GO:0019222 regulation of metabolic process	
GO:0005840 ribosome	GO:0045182 translation regulator activity	
GO:0003674 molecular_function	–	0.053
GO:0008150 biological_process	GO:0043234 protein complex	0.058
GO:0009987 cellular process		
GO:0003674 molecular_function		
GO:0032991 macromolecular complex		
GO:0043234 protein complex		
GO:0032991 macromolecular complex	GO:0005575 cellular_component	0.098
GO:0043234 protein complex	GO:0044464 cell part	
GO:0044238 primary metabolic process	GO:0010608 posttranscriptional regulation of gene expression	
GO:0009987 cellular process	GO:0043226 organelle	
GO:0044445 cytosolic part	GO:0051246 regulation of protein metabolic process	
–	–	0.035
GO:0044238 primary metabolic process	GO:0034645 cellular macromolecule biosynthetic process	0.080
GO:0005198 structural molecule activity	GO:0022625 cytosolic large ribosomal subunit	
GO:0032991 macromolecular complex	GO:0034960 cellular biopolymer metabolic process	
GO:0005840 ribosome	GO:0044260 cellular macromolecule metabolic process	
GO:0044445 cytosolic part	GO:0009059 macromolecule biosynthetic process	
GO:0032991 macromolecular complex	GO:0044424 intracellular part	0.080
GO:0019538 protein metabolic process		
GO:0044267 cellular protein metabolic process		
GO:0044238 primary metabolic process		
GO:0006412 translation		
GO:0044267 cellular protein metabolic process	GO:0043229 intracellular organelle	0.039
GO:0009987 cellular process	GO:0043226 organelle	
GO:0019538 protein metabolic process		
GO:0032991 macromolecular complex		
GO:0043229 intracellular organelle		
GO:0043228 nonmembrane-bounded organelle	GO:0044422 organelle part	0.093
GO:0043232 intracellular nonmembrane-bounded organelle	GO:0044446 intracellular organelle part	
GO:0044445 cytosolic part	GO:0044424 intracellular part	
GO:0032991 macromolecular complex		
GO:0009987 cellular process		
GO:0016043 cellular component organization	GO:0006996 organelle organization	0.041
GO:0009987 cellular process	GO:0016043 cellular component organization	
GO:0006996 organelle organization		
GO:0032991 macromolecular complex		
GO:0008150 biological_process		
GO:0044238 primary metabolic process	GO:0015934 large ribosomal subunit	0.108
GO:0030529 ribonucleoprotein complex	GO:0030686 90S preribosome	
GO:0009987 cellular process	GO:0044464 cell part	
GO:0032991 macromolecular complex	GO:0034961 cellular biopolymer biosynthetic process	
GO:0006412 translation	GO:0015935 small ribosomal subunit	

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
113	116	34	0.68	<0.001	21
114	69	13	0.68	0.001	1
115	96	21	0.68	<0.001	5
116	38	9	0.68	–	0
117	109	30	0.68	<0.001	9
118	66	17	0.68	0.001	1
119	104	27	0.68	<0.001	5
120	122	36	0.68	<0.001	38
121	74	16	0.68	0.001	8
122	126	38	0.68	<0.001	35
123	83	18	0.68	<0.001	3
124	119	31	0.67	<0.001	8
125	133	41	0.67	<0.001	27
126	132	25	0.67	<0.001	18

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0009058 biosynthetic process	GO:0000105 histidine biosynthetic process	0.084
GO:0032991 macromolecular complex	GO:0006547 histidine metabolic process	
GO:0044249 cellular biosynthetic process	GO:0009075 histidine family amino acid metabolic process	
GO:0006412 translation	GO:0009076 histidine family amino acid biosynthetic process	
GO:0009987 cellular process	GO:0009059 macromolecule biosynthetic process	
GO:0009987 cellular process	–	0.053
GO:0003674 molecular_function	GO:0022627 cytosolic small ribosomal subunit	0.050
GO:0009987 cellular process		
GO:0022627 cytosolic small ribosomal subunit		
GO:0044267 cellular protein metabolic process		
GO:0019538 protein metabolic process		
–	–	0.041
GO:0032991 macromolecular complex	GO:0043234 protein complex	0.076
GO:0044445 cytosolic part		
GO:0009987 cellular process		
GO:0006412 translation		
GO:0005198 structural molecule activity		
GO:0009987 cellular process	–	0.037
GO:0003674 molecular_function	–	0.072
GO:0009987 cellular process		
GO:0044445 cytosolic part		
GO:0032991 macromolecular complex		
GO:0008150 biological_process		
GO:0044238 primary metabolic process	GO:0022613 ribonucleoprotein complex biogenesis	0.097
GO:0032991 macromolecular complex	GO:0042254 ribosome biogenesis	
GO:0033279 ribosomal subunit	GO:0044085 cellular component biogenesis	
GO:0008152 metabolic process	GO:0034961 cellular biopolymer biosynthetic process	
GO:0043283 biopolymer metabolic process	GO:0015935 small ribosomal subunit	
GO:0022627 cytosolic small ribosomal subunit	GO:0043332 mating projection tip	0.089
GO:0044445 cytosolic part	GO:0044463 cell projection part	
GO:0032991 macromolecular complex	GO:0022627 cytosolic small ribosomal subunit	
GO:0043332 mating projection tip		
GO:0044463 cell projection part		
GO:0032991 macromolecular complex	GO:0008135 "translation factor activity, nucleic acid binding"	0.106
GO:0044445 cytosolic part	GO:0034961 cellular biopolymer biosynthetic process	
GO:0006412 translation	GO:0034645 cellular macromolecule biosynthetic process	
GO:0009987 cellular process	GO:0043229 intracellular organelle	
GO:0043234 protein complex	GO:0044422 organelle part	
GO:0003674 molecular_function	GO:0043234 protein complex	0.053
GO:0043234 protein complex		
GO:0032991 macromolecular complex		
GO:0032991 macromolecular complex	GO:0005488 binding	0.093
GO:0006412 translation	GO:0044422 organelle part	
GO:0009987 cellular process	GO:0044446 intracellular organelle part	
GO:0005488 binding		
GO:0044422 organelle part		
GO:0009987 cellular process	GO:0015935 small ribosomal subunit	0.092
GO:0032991 macromolecular complex	GO:0043229 intracellular organelle	
GO:0033279 ribosomal subunit	GO:0044422 organelle part	
GO:0044238 primary metabolic process	GO:0044446 intracellular organelle part	
GO:0006412 translation	GO:0043226 organelle	
GO:0044238 primary metabolic process	GO:0031125 rRNA 3′-end processing	0.080
GO:0016070 RNA metabolic process	GO:0043628 ncRNA 3′-end processing	
GO:0044237 cellular metabolic process	GO:0034660 ncRNA metabolic process	
GO:0009987 cellular process	GO:0006139 "nucleobase, nucleoside, nucleotide	
GO:0008152 metabolic process	and nucleic acid metabolic process"	
	GO:0008152 metabolic process	

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
127	57	14	0.67	–	0
128	51	18	0.67	<0.001	1
129	77	25	0.67	<0.001	5
130	75	22	0.67	<0.001	4
131	106	26	0.67	<0.001	6
132	133	25	0.67	<0.001	21
133	128	35	0.67	<0.001	22
134	107	28	0.67	<0.001	19
135	109	24	0.66	<0.001	17
136	72	16	0.66	<0.001	9
137	113	24	0.66	<0.001	11
138	48	12	0.66	–	0
139	58	13	0.66	–	0

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
–	–	0.042
GO:0003674 molecular_function	–	0.044
GO:0009987 cellular process	GO:0034622 cellular macromolecular complex assembly	0.048
GO:0043933 macromolecular complex subunit organization	GO:0043933 macromolecular complex subunit organization	
GO:0034621 cellular macromolecular complex subunit organization	GO:0034621 cellular macromolecular complex subunit organization	
GO:0003674 molecular_function		
GO:0034622 cellular macromolecular complex assembly		
GO:0044238 primary metabolic process	GO:0016070 RNA metabolic process	0.067
GO:0016070 RNA metabolic process		
GO:0003674 molecular_function		
GO:0043283 biopolymer metabolic process		
GO:0003674 molecular_function	GO:0043229 intracellular organelle	0.076
GO:0043229 intracellular organelle	GO:0043226 organelle	
GO:0032991 macromolecular complex		
GO:0043226 organelle		
GO:0044238 primary metabolic process		
GO:0032991 macromolecular complex	GO:0005488 binding	0.097
GO:0044238 primary metabolic process	GO:0005622 intracellular	
GO:0009987 cellular process	GO:0044424 intracellular part	
GO:0044445 cytosolic part	GO:0044249 cellular biosynthetic process	
GO:0005198 structural molecule activity		
GO:0032991 macromolecular complex	GO:0034961 cellular biopolymer biosynthetic process	0.096
GO:0006412 translation	GO:0034645 cellular macromolecule biosynthetic process	
GO:0005198 structural molecule activity	GO:0009059 macromolecule biosynthetic process	
GO:0005840 ribosome	GO:0043284 biopolymer biosynthetic process	
GO:0043284 biopolymer biosynthetic process	GO:0043234 protein complex	
GO:0005198 structural molecule activity	GO:0005737 cytoplasm	0.074
GO:0043228 nonmembrane-bounded organelle	GO:0015935 small ribosomal subunit	
GO:0043232 intracellular nonmembrane-bounded organelle	GO:0022627 cytosolic small ribosomal subunit	
GO:0044445 cytosolic part	GO:0032268 regulation of cellular protein metabolic process	
GO:0005840 ribosome		
GO:0009058 biosynthetic process	GO:0003676 nucleic acid binding	0.078
GO:0044238 primary metabolic process	GO:0006139 "nucleobase, nucleoside, nucleotide and nucleic acid metabolic process"	
GO:0044249 cellular biosynthetic process	GO:0008152 metabolic process	
GO:0032991 macromolecular complex	GO:0006417 regulation of translation	
GO:0043284 biopolymer biosynthetic process	GO:0009059 macromolecule biosynthetic process	
GO:0000462 "maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)"	GO:0000462 "maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)"	0.050
GO:0030490 maturation of SSU-rRNA	GO:0030490 maturation of SSU-rRNA	
GO:0022627 cytosolic small ribosomal subunit	GO:0022627 cytosolic small ribosomal subunit	
GO:0006412 translation		
GO:0043228 nonmembrane-bounded organelle		
GO:0044238 primary metabolic process	GO:0008152 metabolic process	0.080
GO:0030529 ribonucleoprotein complex	GO:0044237 cellular metabolic process	
GO:0005840 ribosome		
GO:0043228 nonmembrane-bounded organelle		
GO:0043232 intracellular nonmembrane-bounded organelle		
–	–	0.033
–	–	0.041

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
140	135	37	0.66	<0.001	14
141	103	21	0.66	<0.001	10
142	164	32	0.66	<0.001	26
143	90	18	0.66	<0.001	21
144	101	20	0.66	<0.001	3
145	122	4	0.66	<0.001	2
146	121	32	0.66	<0.001	14
147	121	30	0.66	<0.001	6
148	104	22	0.66	<0.001	23
149	140	19	0.66	<0.001	14
150	116	30	0.65	<0.001	14
151	61	21	0.65	<0.001	1
152	62	15	0.65	<0.001	1

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0044238 primary metabolic process GO:0032991 macromolecular complex GO:0009987 cellular process GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle	GO:0005488 binding GO:0044424 intracellular part GO:0044237 cellular metabolic process GO:0043170 macromolecule metabolic process	0.101
GO:0009987 cellular process GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle GO:0043229 intracellular organelle GO:0043226 organelle	GO:0065007 biological regulation GO:0050789 regulation of biological process GO:0050794 regulation of cellular process GO:0043229 intracellular organelle GO:0043226 organelle	0.063
GO:0044238 primary metabolic process GO:0032991 macromolecular complex GO:0009987 cellular process GO:0006396 RNA processing GO:0016070 RNA metabolic process	GO:0003824 catalytic activity GO:0006396 RNA processing GO:0030684 preribosome GO:0030686 preribosome GO:0034470 ncRNA processing	0.091
GO:0032991 macromolecular complex GO:0019538 protein metabolic process GO:0044238 primary metabolic process GO:0043283 biopolymer metabolic process GO:0044267 cellular protein metabolic process	GO:0008152 metabolic process GO:0034960 cellular biopolymer metabolic process GO:0044260 cellular macromolecule metabolic process GO:0044237 cellular metabolic process GO:0043234 protein complex	0.064
GO:0009987 cellular process GO:0003674 molecular_function GO:0008150 biological_process	–	0.052
GO:0008150 biological_process GO:0003674 molecular_function	–	0.045
GO:0032991 macromolecular complex GO:0044238 primary metabolic process GO:0009058 biosynthetic process GO:0044249 cellular biosynthetic process GO:0009987 cellular process	GO:0009059 macromolecule biosynthetic process GO:0043284 biopolymer biosynthetic process GO:0044249 cellular biosynthetic process	0.061
GO:0003824 catalytic activity GO:0032991 macromolecular complex GO:0044238 primary metabolic process GO:0030684 preribosome	GO:0003824 catalytic activity GO:0030684 preribosome	0.088
GO:0044238 primary metabolic process GO:0034660 ncRNA metabolic process GO:0034470 ncRNA processing GO:0031125 rRNA 3′-end processing GO:0009987 cellular process	GO:0000459 exonucleolytic trimming during rRNA processing GO:0000467 "exonucleolytic trimming to generate mature 3′-end of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)" GO:0000469 cleavages during rRNA processing GO:0006364 rRNA processing GO:0016072 rRNA metabolic process	0.070
GO:0044238 primary metabolic process GO:0019538 protein metabolic process GO:0044267 cellular protein metabolic process GO:0032991 macromolecular complex GO:0005737 cytoplasm	GO:0044464 cell part GO:0005737 cytoplasm GO:0006417 regulation of translation GO:0032268 regulation of cellular protein metabolic process GO:0043170 macromolecule metabolic process	0.069
GO:0019538 protein metabolic process GO:0032991 macromolecular complex GO:0044267 cellular protein metabolic process GO:0044445 cytosolic part GO:0005198 structural molecule activity	GO:0022625 cytosolic large ribosomal subunit GO:0043234 protein complex	0.079
GO:0003674 molecular_function	–	0.051
GO:0003674 molecular_function	–	0.041

(*Continued*)

Table S2 (*Continued*)

Bicluster ID	Number of genes	Number of conditions	Average PCC	The minimum adjusted *P*-value of GO enrichment	Number of enriched GO terms
153	85	27	0.65	<0.001	5
154	142	33	0.65	<0.001	12
155	54	12	0.65	–	0
156	71	15	0.65	<0.001	6
157	103	34	0.65	<0.001	21
158	84	19	0.65	<0.001	6
159	103	20	0.65	<0.001	10
160	74	7	0.65	0.001	3
161	57	7	0.64	<0.001	1
162	87	6	0.63	<0.001	1
163	75	5	0.61	<0.001	2
164	56	10	0.54	–	0

Five most significant GO terms	Five most specific GO terms	Highest pairwise simirarity score
GO:0016070 RNA metabolic process GO:0003674 molecular_function GO:0044238 primary metabolic process GO:0009987 cellular process GO:0034660 ncRNA metabolic process	GO:0034660 ncRNA metabolic process GO:0016070 RNA metabolic process	0.072
GO:0030529 ribonucleoprotein complex GO:0044445 cytosolic part GO:0032991 macromolecular complex GO:0033279 ribosomal subunit GO:0043228 nonmembrane-bounded organelle	GO:0005622 intracellular GO:0043229 intracellular organelle GO:0044422 organelle part GO:0044446 intracellular organelle part GO:0043226 organelle	0.099
–	–	0.039
GO:0043283 biopolymer metabolic process GO:0044238 primary metabolic process GO:0034960 cellular biopolymer metabolic process GO:0043170 macromolecule metabolic process GO:0044260 cellular macromolecule metabolic process	GO:0034960 cellular biopolymer metabolic process GO:0044260 cellular macromolecule metabolic process GO:0043170 macromolecule metabolic process	0.052
GO:0032991 macromolecular complex GO:0009987 cellular process GO:0044445 cytosolic part GO:0043228 nonmembrane-bounded organelle GO:0043232 intracellular nonmembrane-bounded organelle	GO:0015934 large ribosomal subunit GO:0022625 cytosolic large ribosomal subunit GO:0051246 regulation of protein metabolic process GO:0044424 intracellular part GO:0032268 regulation of cellular protein metabolic process	0.079
GO:0005198 structural molecule activity GO:0005488 binding GO:0044445 cytosolic part GO:0009987 cellular process GO:0032991 macromolecular complex	GO:0005488 binding	0.074
GO:0032991 macromolecular complex GO:0034621 cellular macromolecular complex subunit organization GO:0044238 primary metabolic process GO:0009987 cellular process GO:0043933 macromolecular complex subunit organization	GO:0065003 macromolecular complex assembly GO:0034622 cellular macromolecular complex assembly GO:0043933 macromolecular complex subunit organization GO:0034621 cellular macromolecular complex subunit organization	0.063
GO:0044422 organelle part GO:0044446 intracellular organelle part GO:0009987 cellular process	GO:0044422 organelle part GO:0044446 intracellular organelle part	0.037
GO:0003674 molecular_function	–	0.048
GO:0003674 molecular_function	–	0.048
GO:0032991 macromolecular complex GO:0003674 molecular_function	–	0.045
–	–	0.033

Notes: The steps to select specific GO terms from each cluster. (**1**) We hypothesise if a GO term appears on only a small number of biclusters (ie, 1 of 4 biclusters), it is specific for the biclusters. (**2**) We have 164 biclusters. By the proportion test, 1 of 4 biclusters corresponds to 31 of 164 biclusters at 0.05 significance level. (**3**) Therefore, GO terms appear less than 32 times are specific terms.