BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 24 2010, pages 3090–3097
doi:10.1093/bioinformatics/btq592

Gene expression

Advance Access publication October 19, 2010

Model-free unsupervised gene set screening based on information enrichment in expression profiles

Atushi Niida[1,*], Seiya Imoto[1], Rui Yamaguchi[1], Masao Nagasaki[1], André Fujita[2], Teppei Shimamura[1] and Satoru Miyano[1]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639 and [2]Computational Science Research Program, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Associate Editor: Trey Ideker

ABSTRACT

Motivation: A number of unsupervised gene set screening methods have recently been developed for search of putative functional gene sets based on their expression profiles. Most of the methods statistically evaluate whether the expression profiles of each gene set are fit to assumed models: e.g. co-expression across all samples or a subgroup of samples. However, it is possible that they fail to capture informative gene sets whose expression profiles are not fit to the assumed models.

Results: To overcome this limitation, we propose a model-free unsupervised gene set screening method, Matrix Information Enrichment Analysis (MIEA). Without assuming any specific models, MIEA screens gene sets based on information richness of their expression profiles. We extensively compared the performance of MIEA to those of other unsupervised gene set screening methods, using various types of simulated and real data. The benchmark tests demonstrated that MIEA can detect singular expression profiles that the other methods fail to find, and performs broadly well for various types of input data. Taken together, this study introduces MIEA as a broadly applicable gene set screening tool for mining regulatory programs from transcriptome data.

Contact: aniida@ims.u-tokyo.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 16, 2010; revised on September 10, 2010; accepted on October 15, 2010

1 INTRODUCTION

Recently, a number of gene set screening methods have been introduced successfully to analyze regulatory programs hidden in transcriptome data. While classical microarray analyses focus on individual genes, gene set screening methods search a prescribed gene set library for gene sets with informative expression profiles in given transcriptome data. By treating a gene set as a unit, these methods enhance statistical power as well as make results easier to biologically interpret. For example, Gene Set Enrichment Analysis (GSEA), which finds gene sets showing differential expression between two sample groups (Subramanian *et al.*, 2005), has become one of the pivotal tools for transcriptome analysis. A lot of other methods employing similar supervised approach are used to search for gene sets associated with given sample labels (Huang *et al.*, 2009).

However, if we want to globally search for putatively functional gene sets, irrespectively of association with sample labels, we need another approach that does not rely on supervision of sample labels, i.e. the unsupervised approach. For this purpose, Segal *et al.* (2004) proposed a method to find a set of genes whose expressions are induced or repressed in any samples based on a hypergeometric test. Although their pioneering method realizes an unsupervised approach, this method can also be regarded as a kind of supervised method in that it tests differential expression between a single sample and the others. We recently proposed an unsupervised method, Extraction of Expression Module (EEM) to search for gene sets that have significantly large subsets of genes coherently expressed in the input transcriptome data (Niida *et al.*, 2009a). Kim *et al.* (2007) also proposed a method termed Gene Set Expression Coherence Analysis (GSECA), which selects gene sets based on expression coherence. GSECA is different from EEM in that GSECA measures expression coherence, taking into account all the members in the gene set rather than focusing on the coherent subset. Since EEM assumes that a significant gene set is co-expressed across all samples, it is possible that EEM fails to capture gene sets that exhibit coherent expression patterns across only a subset of samples. To overcome this limitation, we also developed an extended version of EEM termed Biclustering-based EEM (BEEM), which takes advantage of a biclustering algorithm (Bergmann *et al.*, 2003) to identify gene sets coherently expressed in a subgroup of samples (Niida *et al.*, 2010).

Compared with the supervised approach, few gene set screening methods have so far employed the unsupervised approach, and also there remain some problems to be solved in the existing methods. For example, although EEM and BEEM evaluate expression coherence of an input gene set based on specific models, i.e. co-expression across all samples or a subgroup of samples, it is possible that such a model-based method fails to capture informative expression profiles that do not fit to the assumed model. To address this problem, we developed a novel model-free method termed MIEA (Matrix Information Enrichment Analysis). Without assuming any specific model, MIEA evaluates information in expression profiles; therefore, it could be used to search for gene sets associated with any types of informative expression profiles.

In order to reveal the properties of MIEA and other unsupervised gene set screening methods, we extensively performed benchmark tests to compare their performances using both simulated and

*To whom correspondence should be addressed.

real data. These benchmark tests showed that MIEA can capture significant gene sets which the other methods miss, and is broadly applicable to various types of transcriptome data and gene sets. Together with the other unsupervised gene set screening methods, MIEA would be a useful gene set screening tool for mining regulatory programs from transcriptome data.

2 METHODS

2.1 MIEA

Let $\mathbf{E}$ and M denote an $n_g \times n_s$ input expression matrix and a gene set of $|M|$ genes, respectively. The rows and columns of $\mathbf{E}$ index genes and samples, and the elements of each row vector are normalized so that the mean is 0 and the variance is 1. From $\mathbf{E}$, we extract the row vectors corresponding to the genes in M and obtain a $|M| \times n_s$ sub-matrix $\mathbf{E}_M$. Generally, to evaluate informative patterns in $\mathbf{E}_M$, unsupervised gene set screening methods employ statistics based on their own specific models. For example, EEM uses the maximal-sized coherent subset of M in $\mathbf{E}_M$ as the test statistic. Namely, EEM assumes a specific model for informative patterns in $\mathbf{E}_M$: informative patterns should be coherent expression of a large subset of M. Note that it is possible that model-based methods like EEM fail to detect informative patterns in $\mathbf{E}_M$ if the patterns do not fit to the assumed model, or are too complex to explain based on one simple model. To overcome this limitation, we proposed a model-free method, MIEA. The MIEA statistic scores information richness of $\mathbf{E}_M$ without assuming any specific model, but taking advantage of singular value decomposition (SVD).

Let $\mathbf{X}$ denote an $n \times m$ matrix and its rank is r. Note that, without loss of generality, we assume $n \geq m$, and therefore $r \leq m$ holds. The SVD theorem states (Press *et al.*, 1992):

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T,$$

where $\mathbf{U}$ is an $n \times m$ matrix whose columns are the left singular vectors $\mathbf{u}_k$, and $\mathbf{V}$ is an $m \times m$ matrix whose column are the right singular vectors $\mathbf{v}_k$. The left and right singular vectors form an orthonormal set, i.e. $\mathbf{v}_i^T\mathbf{v}_j = \mathbf{u}_i^T\mathbf{u}_j = 1$ for $i=j$ and $\mathbf{v}_i^T\mathbf{v}_j = \mathbf{u}_i^T\mathbf{u}_j = 0$ for $i \neq j$. An $m \times m$ diagonal matrix $\mathbf{S}$ has diagonal elements $s_1 \geq s_2 \ldots \geq s_m \geq 0$, which are called singular values. Furthermore, $s_k > 0$ for $1 \leq k \leq r$, and $s_k = 0$ for $r+1 \leq k \leq m$.

One important result of the SVD of $\mathbf{X}$ is that

$$\mathbf{X}^{(l)} = \sum_{k=1}^{l} \mathbf{u}_k s_k \mathbf{v}_k^T$$

is the matrix of rank-l that minimizes the sum of the squares, $\sum_{ij}(X_{ij} - X_{ij}^{(l)})^2$. This result can be used for data compression of $\mathbf{X}$. If the singular values s_j rapidly approach to zero for $j \geq l+1$, we can obtain a good approximation of $\mathbf{X}$ by $\mathbf{X}^{(l)}$. In other words, if the singular values are uneven, we can assume $\mathbf{X}$ has so much redundancy that can be compressed. When applying hierarchical clustering to the matrix, this information can be visualized as non-random pattern in the heatmap. Therefore, by measuring unevenness of the singular values, we can know how much information the gene expression matrix has. To measure unevenness of the singular values, we define an *entropy* of the matrix $\mathbf{X}$ by

$$d(\mathbf{X}) = -\sum_{k=1}^{m} p_k \log(p_k),$$

where $p_k = s_k / \sum_{l=1}^{m} s_l$. Using $d(\mathbf{X})$, we can compare information richness between matrices of the same dimension; i.e. the less $d(\mathbf{X})$ is, the more information $\mathbf{X}$ should have. We can assume that $d(\mathbf{X}) = 0$ corresponds to the most information-rich case that can only be achieved when $\mathbf{X} = \mathbf{X}^{(1)}$, while $d(\mathbf{X}) = \log(m)$ corresponds to the most information-poor case where all singular values are equal and $\mathbf{X}$ has no information to be compressed.

MIEA employs the entropy of $\mathbf{E}_M$, $d(\mathbf{E}_M)$, as a test statistic, and its statistical significance is evaluated by an empirical approach. That is, an empirical null distribution can be generated by repeatedly sampling random gene sets whose sizes are equal to that of M, $|M|$. The P-value of $d(\mathbf{E}_M)$ is obtained by the ratio of the null statistics which are smaller than $d(\mathbf{E}_M)$. Namely, the tested null hypothesis is that $d(\mathbf{E}_M)$ is equal to entropies of equal-sized random sub-matrices from $\mathbf{E}$. However, relying only on this empirical approach leads to intensive computational time. Fortunately, we found that, when $|M|$ is large enough, the empirical null distribution is well approximated by the normal distribution. Based on this observation, we first sample 1000 null statistics and apply the Shapiro test to validate their normality. If the P-value of the Shapiro test is greater than the cutoff of 0.01, we calculate the MIEA P-value by fitting the normal distribution to the 1000 null statistics. Otherwise, we continue resampling until obtaining 10^4 null statistics, and empirically calculate the P-value from them. Note that, even in the latter case, computational time is not serious, because gene sets that reject the null hypothesis of the Shapiro test have relatively small sizes and, for such small gene sets, computation can be finished quickly.

2.2 Other methods compared with MIEA

In this study, we compare the performance of MIEA with those of other gene set screening methods employing unsupervised approach: EEM, BEEM, GSECA and SSA. Here, we give belief explanations of the competitive methods.

2.2.1 EEM From the input gene set, M, EEM finds the maximal-sized subset of M that shows coherent expression across all samples under the given radius parameter r (Niida *et al.*, 2009a). A coherent subset of M is represented as:

$$S_{\mathbf{c}}^M = \{i | i \in M, \sum_{j=1}^{n_s} (E_{ij} - c_j)^2 \leq r\},$$

where $\mathbf{c} = (c_1, \ldots, c_{n_s})^T$ is the vector of the center parameters. The EEM algorithm is designed to find $\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} |S_{\mathbf{c}}^M|$, and the maximal-sized coherent subset is given as $S_{\hat{\mathbf{c}}}^M$. EEM uses $|S_{\hat{\mathbf{c}}}^M|$ as a test statistic to evaluate expression coherence of M. It is clear that the results of EEM depend on the value of the radius parameter r. Assuming $S_{\hat{\mathbf{c}}}^{\text{all}}$, the maximal-sized coherent subset for all n_g genes, r can be converted to the relative radius r_R so that $r_R = |S_{\hat{\mathbf{c}}}^{\text{all}}| / n_g$. We tested relative radius parameters of 0.05, 0.10 and 0.15, chose the one which leads to the smallest P-value, and corrected the P-value for multiple testing.

2.2.2 BEEM While EEM measures expression coherence across all samples, BEEM measures expression coherence in a subset of samples by employing a biclustering algorithm, ISA (Bergmann *et al.*, 2003). We assumed a bicluster as a subset of genes that exhibits higher or lower expressions than a predefined threshold across a subset of samples, and *vice versa*. BEEM obtains a subset of M which is associated with a bicluster in $\mathbf{E}$, and statistically evaluates the size of the bicluster-associated subset. ISA requires three parameters for specifying two thresholds controlling the bicluster size and the targeted bicluster types (i.e. upregulated or downregulated). We performed BEEM with 18 different parameter settings and obtained multiple testing-corrected P-values as described by Niida *et al.* (2010).

2.2.3 GSECA GSECA is similar to EEM in that GSECA tests expression coherence of $\mathbf{E}_M$. However, although EEM uses the size of the maximal-sized coherent subset of M to evaluate expression coherence, GSECA uses the mean of expression correlation values between every pair of genes in M as the test statistic e (Kim *et al.*, 2007):

$$e = \frac{2}{|M|(|M|-1)} \sum_{i,k \in M, i \neq k} \sum_{j=1}^{n_s} E_{ij} E_{kj}.$$

The P-value for this statistic is calculated based on an empirical approach described by Kim *et al.* (2007); an empirical null distribution was produced by randomly sampling 10^5 gene sets whose sizes are equal to $|M|$.

2.2.4 SSA Segal *et al.* (2004) proposed a method that screens gene sets without sample label information, relying on over or underexpression in each sample. However, their method does not explicitly assign *P*-value to a single gene set as the other method does. To make comparison easier, we used a reformulated version of their method termed SSA, which assigns a single *P*-value to each gene set (Niida *et al.*, 2010).

The *j*-th column vector of $\mathbf{E}$, $\mathbf{E}_j=(E_{1j},\ldots,E_{n_gj})^T$ scores how much each gene is over or underexpressed in the *j*-th sample, compared with the average across all samples. Based on E_j, we obtain the top 5% of the upregulated genes in the *j*-th sample, denoted as U_j. SSA tests overlap between the input gene set M and U_j based on the hypergeometric test, and a *P*-value, p_j^u, for upregulation of M in the *j*-th sample is obtained. Similarly, a *P*-value, p_j^d, is also calculated for downregulation of M in the *j*-th sample. By calculating p_j^u and p_j^d for each sample, we obtain a *P*-value vector of length $2\times n_s$, $\mathbf{p}=(p_1^u,p_1^d,\ldots,p_{n_s}^u,p_{n_s}^d)^T$. SSA finally combines $\mathbf{p}$ using Fisher's method (Fisher, 1932) to obtain a single *P*-value to M; if this combined *P*-value is small, it means that M is over or underexpressed in any of the n_s samples.

2.3 Generation of simulated data

We simulated expression matrices and gene set libraries for the input data. Generally, the transcriptome contains sets of co-regulated genes called *expression modules*. Genes that belong to the same expression module behave together at the expression level. We simulated expression matrices of 4000 genes $\times 100$ samples, each of which harbors different types of expression modules. A simulated gene set library contains *positive gene sets*, which have significant overlaps with any of the expression modules, and *negative gene sets*, which were randomly sampled from the 4000 genes. Unsupervised gene set screening can be regarded as a process to search for gene sets that are associated with expression modules; the positive gene sets should be identified by unsupervised gene set screening, while the negative gene sets should not.

To simulate expression matrices, we assumed three different models:

- **Coherent model.** We assumed that an expression matrix has non-overlapping 20 modules, each of which consists of 200 module genes. For each module, we first chose one gene and generated its expression values across samples by the standard Gaussian distribution. That is, assuming that we chose gene k, we have $E_{kj}\sim \mathrm{N}(0,1)$ for $j=1,\ldots,100$. The other module genes were generated so that they gather around gene k. The expression value of gene i who is a member of the module generated from gene k is

$$E_{ij}=s_cE_{kj}+(1-s_c)\eta_{ij},$$

  where $\eta_{ij}\sim \mathrm{N}(0,1)$ and s_c is a parameter specifying signal strength.
- **Bicluster model.** We assumed that an expression matrix has 50 modules, each of which consists of 200 module genes, and is allowed to overlap with each other. We randomly selected 200 genes from the 4000 genes to define module genes of each expression modules. We assumed the 200 module gene constitute a bicluster, and randomly chose $100r_s$ samples as samples that constitute the bicluster in the module. Here, r_s is a parameter specifying the ratio of the samples that constitute a bicluster. Let B_{ij} be an indicator variable, where B_{ij} takes 1 if and only if the expression value of gene i in sample j, E_{ij}, belongs to any of the defined bicluster, or 0 otherwise. We set

$$E_{ij}=s_bB_{ij}+\zeta_{ij},$$

  where $\zeta_{ij}\sim \mathrm{N}(0,1)$ and s_b is a parameter specifying signal strength.
- **Sine-wave model.** We assumed that an expression matrix has non-overlapping 20 modules, each of which consists of 200 module genes. For each module, we randomly shuffled indices of samples. The expression value of the *i*-th module gene and the sample associated with the shuffled sample index j was then defined as follows:

$$E_{ij}=s_c\sin\{\frac{\pi}{8}(i+j)\}+(1-s_c)\epsilon_{ij},$$

  where $\epsilon_{ij}\sim \mathrm{N}(0,1)$, and s_c is a parameter specifying signal strength.

We next generated gene set libraries that make pairs with each of the expression matrices. Each gene set library includes 10 positive and 10 negative gene sets. A positive gene set includes $200r_g$ genes sampled from one expression module, and randomly sampled $200(1-r_g)$ genes. Here, r_g is a given parameter specifying the ratio of module genes in the positive gene set. Ten negative gene sets were prepared by randomly sampling 200 genes.

We additionally simulated another type of gene set library for the expression matrices of the coherent and bicluster models. For a given parameter K, the alternative gene set library was prepared so that the $200r_g$ genes in a positive gene set consist of equal-sized K groups whose members were sampled from different K modules; i.e. a positive gene set with a larger K value has a more complex expression profiles. Hereafter, we refer to the paired set of these alternative gene set libraries and the expression matrices from the coherent and bicluster models as *the composite coherent model* and *the composite bicluster model*, respectively. Note that the models with $K=1$ are identical to the normal coherent and bicluster models.

2.4 Preparation of real data

We also measured performance of each method using real data. The real data used in this study include seven expression datasets and five gene set libraries.

To prepare the expression datasets, we obtained seven datasets from public resources. The information of the seven datasets is summarized in Table 1. The data preprocessing was done as described by Niida *et al.* (2009a). Briefly, they were scaled to the logarithmic scale, and normalized in each sample. After the probe set IDs were converted to gene symbols, expression profiles of 6000 or 8000 genes (depending on the original numbers of genes) with the largest variance were extracted and normalized across samples.

The five gene set libraries have various types of sources as listed in Table 2. The PWM gene set library was prepared by predicting downstream target genes of *cis*-regulatory motifs, using the proximal promoter sequences and position weight matrices (PWMs) obtained from the TRANSFAC database (Matys *et al.*, 2006; Niida *et al.*, 2010). The locus gene set library was prepared by collecting genes located next to each other on the chromosomes. To collect such genes, we set a window containing 100 genes adjacent on the genome, moved the window by 10 genes and kept genes covered by the window at each position (Niida *et al.*, 2009a). The miRtarget gene set library was prepared from the MicroCosm Targets database (Griffiths-Jones *et al.*, 2008); genes in each gene set were predicted to have miRNA binding sites in their 3′UTR with $P<10^{-3}$. The curated and GO gene set libraries are downloaded from MSigDb, a database used in GSEA (Subramanian *et al.*, 2005). Note that not all genes in each gene set were used for our analysis; the intersection of each gene set and genes in each expression dataset was used as an actual input gene set. In each gene set screening, we also discarded gene sets for which the intersection is less than 10.

3 RESULTS

3.1 Simulated data test

We performed a simulation study to compare the performance of MIEA with those of the four other unsupervised gene set screening methods: EEM, BEEM, GSCEA and SSA. We simulated input data based on the five models: the coherent, composite coherent, bicluster, composite bicluster and sine-wave models (See examples in Supplementary Figure S1). Each of which model has several arbitrary parameters: i.e. r_g and s_c for the coherent and sine-wave models; r_g, s_c and K for the composite coherent model;

A The coherent model

r_g	s_c	MIEA	EEM	BEEM	GSECA	SSA
0.2	0.4	0.838	0.998	0.985	0.999	0.927
0.2	0.2	0.746	0.998	0.833	0.933	0.626
0.2	0.1	0.563	0.820	0.555	0.720	0.513
0.1	0.4	0.757	0.996	0.952	0.798	0.655
0.1	0.2	0.596	0.952	0.595	0.653	0.582
0.1	0.1	0.570	0.549	0.490	0.592	0.530
0.05	0.4	0.646	0.898	0.596	0.615	0.574
0.05	0.2	0.618	0.714	0.540	0.524	0.546
0.05	0.1	0.512	0.573	0.459	0.497	0.529

B The composite coherent model (K=2)

r_g	s_c	MIEA	EEM	BEEM	GSECA	SSA
0.2	0.4	0.878	0.996	0.966	0.933	0.798
0.2	0.2	0.708	0.980	0.611	0.806	0.546
0.2	0.1	0.571	0.574	0.499	0.574	0.504
0.1	0.4	0.783	0.922	0.718	0.674	0.571
0.1	0.2	0.616	0.840	0.576	0.635	0.497
0.1	0.1	0.545	0.585	0.439	0.545	0.550
0.05	0.4	0.762	0.719	0.579	0.534	0.506
0.05	0.2	0.577	0.692	0.525	0.533	0.563
0.05	0.1	0.516	0.562	0.446	0.541	0.447

C The composite coherent model (r_g, s_c) = (0.2, 0.4)

K	MIEA	EEM	BEEM	GSECA	SSA
1	0.866	0.998	0.980	0.998	0.931
2	0.896	0.997	0.984	0.921	0.780
5	0.955	0.926	0.688	0.790	0.625
10	0.998	0.776	0.486	0.732	0.471
20	0.995	0.733	0.489	0.745	0.356

D The bicluster model

r_g	r_s	s_b	MIEA	EEM	BEEM	GSECA	SSA
0.2	0.3	2.0	0.927	0.993	0.985	0.964	0.837
0.2	0.3	1.0	0.869	0.943	0.985	0.943	0.708
0.2	0.1	2.0	0.920	0.962	0.985	0.971	0.901
0.2	0.1	1.0	0.798	0.857	0.933	0.905	0.740
0.1	0.3	2.0	0.736	0.821	0.963	0.757	0.599
0.1	0.3	1.0	0.652	0.669	0.562	0.720	0.507
0.1	0.1	2.0	0.731	0.722	0.745	0.760	0.675
0.1	0.1	1.0	0.619	0.647	0.583	0.656	0.576

E The composite bicluster model (K=2)

r_g	r_s	s_b	MIEA	EEM	BEEM	GSECA	SSA
0.2	0.3	2.0	0.915	0.918	0.958	0.909	0.609
0.2	0.3	1.0	0.828	0.812	0.691	0.840	0.600
0.2	0.1	2.0	0.880	0.832	0.893	0.907	0.739
0.2	0.1	1.0	0.692	0.738	0.705	0.857	0.674
0.1	0.3	2.0	0.720	0.658	0.608	0.703	0.504
0.1	0.3	1.0	0.695	0.629	0.468	0.688	0.455
0.1	0.1	2.0	0.709	0.677	0.573	0.705	0.502
0.1	0.1	1.0	0.600	0.587	0.519	0.654	0.574

F The composite bicluster model (r_g, r_s, s_b) = (0.2, 0.3, 2.0)

K	MIEA	EEM	BEEM	GSECA	SSA
1	0.912	0.996	0.955	0.983	0.824
2	0.889	0.932	0.973	0.916	0.660
5	0.882	0.749	0.677	0.834	0.556
10	0.810	0.612	0.529	0.746	0.523
20	0.851	0.577	0.518	0.765	0.457

G The sine-wave model

r_g	s_c	MIEA	EEM	BEEM	GSECA	SSA
0.2	0.8	0.996	0.998	0.741	0.500	0.432
0.2	0.4	0.845	0.927	0.451	0.510	0.546
0.2	0.2	0.516	0.455	0.479	0.447	0.480
0.1	0.8	0.835	0.880	0.582	0.507	0.466
0.1	0.4	0.645	0.610	0.470	0.519	0.501
0.1	0.2	0.526	0.479	0.504	0.484	0.481
0.05	0.8	0.661	0.652	0.536	0.586	0.478
0.05	0.4	0.556	0.561	0.523	0.495	0.539
0.05	0.2	0.489	0.500	0.482	0.493	0.485

Fig. 1. The AUCs of MIEA, EEM, BEEM, GSECA and SSA when applied to data from the five simulation models. For each of the five simulation models, data were generated using several different parameter settings. The AUCs computed by applying the five methods to them are presented.

r_g, r_s and s_b for the bicluster model; and r_g, r_s, s_b and K for the composite bicluster model. We used various combinations of parameter values for each model; therefore, in total, we assumed 53 types of data simulators employing different models and different parameter settings. We generated 20 Monte Carlo datasets from each data simulator, and applied each of the five methods to each of the 20 datasets. From the results pooled across the 20 simulations, we calculated specificity and sensitivity for each method over the whole range of significance cutoffs to depict receiver operating characteristic (ROC) curves. Specificity was calculated as the proportion of negative gene sets whose significance is below the cutoff, while sensitivity is calculated as the proportion of positive gene sets whose significance is above the cutoff. We then computed the area under the curves (AUCs) as a measure of the performance of each method, since the AUC assesses the overall discriminative ability of the method at determining whether a given gene set is associated with an expression module. The results are summarized as follows:

- **Coherent model.** This model assumes that genes in the same expression module are coherently expressed across all samples, and a subset of a positive gene set overlaps with any single expression module. This type of expression modules should be efficiently captured by EEM. Expectedly, EEM performs best among the five methods, while BEEM, GSECA and MIEA follow EEM (Fig. 1A).
- **Composite coherent model.** This model also assumes coherent expression modules, but differs in that a positive gene set is assumed to harbor K subsets overlapping with different expression modules. Also for the composite coherent model with $K=2$, EEM works best among the five methods; however, the performance is attenuated compared with that for the coherent model. For BEEM and GSECA, we also observed similar trends; their performances are worse than those for the coherent model. On the other hand, MIEA does not show any distinct difference between the results for the two models, and keeps relatively good performance (Fig. 1B). This observation prompted us to inspect dependency of these method on the value of parameter K. We found that the performance of MIEA is the most robust to the changes of K, and MIEA performs best for larger K ($K=5$, 10 and 20 in Fig. 1C).
- **Bicluster model.** In this model, module genes are assumed to be overexpressed in a subset of samples. Since this type of expression module is a target of BEEM, BEEM performs best for this model. The other methods except SSA also show good performance (Fig. 1D).

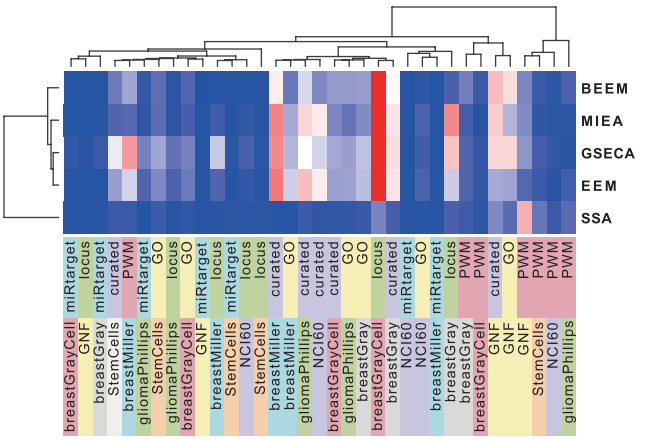

Fig. 2. The performances of MIA, EEM, BEEM, GSECA and SSA when applied to various types of real data. We applied the five methods to the combinations of the seven microarray datasets and the five gene set libraries. The ratios of significant gene sets in all the 175 gene set screenings are visualized as a heatmap (high ratio: red, low ratio: blue).

Table 1. Expression datasets

ID	The number of genes	The number of samples	Description	Reference
breastGray	6000	118	Breast tumors	Chin *et al.* (2006)
breastGrayCell	6000	54	Breast cancer cell lines	Neve *et al.* (2006)
breastMiller	8000	251	Breast tumors	Miller *et al.* (2005)
gliomaPhillips	8000	100	Glioma tumors	Phillips *et al.* (2006)
NCI60	8000	60	Various cancer cell lines	Shankavaram *et al.* (2007)
GNF	8000	79	Various tissues	Su *et al.* (2004)
StemCells	6000	291	Various stem cells	Müller *et al.* (2008)

Table 2. Gene set libraries

ID	The number of gene sets	Description	Reference
PWM	200	Having common TF binding motifs	Matys *et al.* (2006), Niida *et al.* (2010)
locus	1813	Adjacent on the genome	Niida *et al.* (2009a)
GO	1454	Associated with common GO terms	Subramanian *et al.* (2005)
curated	1892	Curated from the literature	Subramanian *et al.* (2005)
miRtarget	851	Having common miRNA binding motifs	Griffiths-Jones *et al.* (2008)

- **Composite bicluster model.** This model is based on the same expression module model as the bicluster model, but assumes that positive gene sets have two subsets associated with different expression modules. As compared with the bicluster model, the performance of BEEM is substantially impaired for the composite bicluster model with $K=2$. On the other hand, MIEA, EEM and GSECA remain good choices, although the performance of EEM is slightly attenuated (Fig. 1E). We also confirmed that MIEA performs best for large K values in this model (Fig. 1F).
- **Sine-wave model.** In this model, expression values of module genes are defined by the sine function whose phase depends on the sum of gene and sample indices. A singular expression module generated from this model seems a difficult target if the method is based on simple expression module model like expression coherence and bicluster. A model-free method, MIEA, successfully retrieves a gene set associated with this singular expression module. EEM also performs comparably well, suggesting that EEM is powerful enough to capture coherent patterns hidden in singular expression modules. The other three methods hardly work at all (Fig. 1G).

Taken together, MIEA demonstrates relatively good performance for all the five models. Especially, it should be noted that MIEA performs better than the other methods in complex expression profiles generated from the sine-wave model or the composite models with large K values. This observation suggests that MIEA is broadly applicable for various types of input data, which is what we would expect for a model-free method. EEM shows superior performance for all the five models, especially for the coherent and composite coherent models. GSECA is also a good choice except for the sine-wave model. BEEM is effective for the bicluster model, as expected. SSA shows the worst performance for all the five models.

3.2 Real data test

Next, we evaluated the practical performance of the above methods using real data. The real data include seven microarray datasets and five gene set libraries listed in Tables 1 and 2. We performed gene set screenings by applying MIEA EEM, BEEM, GSCEA and SSA to every combination of the expression datasets and the gene set libraries: i.e. (5 methods) $\times$ (7 expression datasets) $\times$ (5 gene set libraries) $=$ (175 gene set screenings) were performed.

In each gene set screening, we considered gene sets whose P-values are less than a cutoff of 10^{-4} as significant gene sets. To evaluate the performance of each method, we calculated the ratio of significant gene sets in each gene set screening (Supplementary Fig. S2). A heatmap of the performance profiles across input data (Fig. 2) reveals that the performance of the four methods except SSA has similar dependency on the input data while, on average, GSECA has slightly better performance than others. The four methods other than SSA show outstandingly high performance for the combination of the breastGrayCell expression dataset and the locus gene set library. This observation presumably reflects the fact that in breast cancer cell lines there are some chromosomal regions whose copy numbers are frequently altered, and the genes residing on such regions behave together in the transcriptome. Note that some methods take substantial ratios of the locus gene set library for expression data from clinical tumor samples, i.e. breastGray breastMiller and gliomaPhillips. However, it is reasonable that the data from monoclonal cell lines have much higher signal strength than those from the mixture of heterogeneous tumor tissues.

On average, the curated and GO gene set libraries tend to contain more significant gene sets across the seven expression datasets. Although substantial ratios were observed for the PWM gene set library with some expression datasets, the miRtarget gene set library contains few significant gene sets in most of the expression datasets. This observation suggests that expression regulations by miRNAs

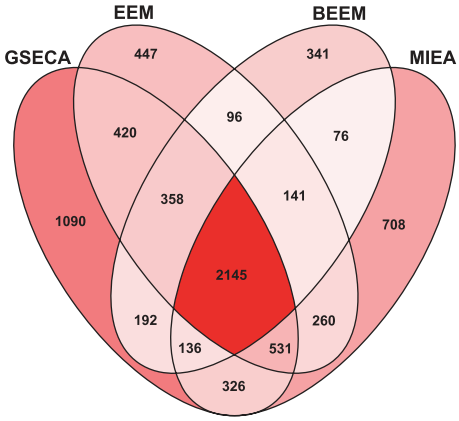

Fig. 3. The 15 groups of expression-gene set pairs taken as significant by MEIA, EEM, BEEM and GSECA. A total of 7267 expression-gene set pairs significant in any of the four methods were divided into 15 groups based on which method takes them as significant, and visualized using a Venn diagram.

are hard to be detected at the transcriptome level, consistent with the previous reports that miRNAs regulate their target genes at both the transcriptional and post-transcriptional levels (Carthew and Sontheimer, 2009). Among the five methods, SSA shows the worst performance in most of the gene set screenings. Although it shows high performance for PWM gene set library, this observation might be only a statistical artifact caused by gene set-size dependency, as described previously (Niida *et al.*, 2010).

The breastMiller dataset whose matrix size is the largest (8000 genes $\times$ 251 samples) is most associated with significant results. Based on this observation, we checked how much the matrix size affects the performance of each method. For the breastMiller and GNF datasets, we downsized expression matrices by randomly sampling 50, 25 and 12.5% of genes or samples. We then applied the five methods to the combination of the curated gene set library and each of the downsized matrices, and compared the results with those from the original matrices. This analysis demonstrated that the performance of each method is relatively robust to the sample-wise downsizing while the gene-wise downsizing decreases the performance more severely, especially for SSA (Supplementary Fig. S3). Among the five methods, EEM seems the most robust to both types of the downsizing. Thus, the performance variance is partially caused by the different sizes of input expression matrices, and also would reflect data qualities and biological properties of them.

Next, we tried to characterize performance difference by focusing on each gene set. We omitted SSA from this analysis, because it shows the worst performance in the above analyses. Since each library includes hundreds to thousands of gene sets, all the combinations of the seven expression datasets and five gene set libraries yields 35026 pairs of expression datasets and gene sets. Hereafter, we refer to each of the pairs simply as an *expression-gene set pair*. When we applied MIEA, EEM, BEEM and GSECA to the 35026 expression-gene set pairs, 7267 pairs scored P-values less than a cutoff of 10^{-4} in any of the four methods. We divided the 7267 pairs into 15 groups based on in which methods they are significant. The Venn diagram visualizing the 15 groups demonstrates that expression-gene set pairs targeted by the four methods are largely overlapping, while each method, especially MIEA and GSECA, has unique expression-gene set pairs which are retrieved by only one method, but not others (Fig. 3). We also examined which expression dataset and gene set library is associated with the 15 group of expression-gene set pairs. For each group, we calculated a hypergeometric P-value for enrichment of the expression-gene set pairs associated with each expression dataset or each gene set library. Namely, the P-value for the enrichment was calculated as follows:

$$p = 1 - \sum_{i=0}^{|F\cap G|-1} \binom{N-|F|}{|G|-i}\binom{|G|}{i} \Big/ \binom{N}{|G|},$$

where G is the group of expression-gene set pairs, F is a set of expression-gene set pairs associated with a particular expression dataset or gene set library, and N is the total number of expression-gene set pairs, i.e. 35026. Figure 4 shows the association P-value p in minus log scale. Negative associations (i.e. depletion of the expression-gene set pairs associated with each expression dataset or each gene set library) can be measured by $1-p$, which are shown in Supplementary Figure S4. We found that the degree of association to each expression dataset or each gene set library substantially varies among the 15 groups. For example, expression profiles of gene sets in the GNF expression dataset tend to be targeted uniquely by BEEM. For the gliomaPhillips dataset, MIEA performs relatively poor while EEM performs better. For the breastGray expression datasets, MIEA performs well; it retrieves gene sets that the other methods fail to find. Taken together, these observations suggest that the four methods have different characteristics to target different types of expression profiles of gene sets.

4 DISCUSSION

In this study, we introduced a new unsupervised model-free gene set screening method, MIEA, and extensively compared its performance with those of the four existing unsupervised gene set screening methods: SSA, EEM, BEEM and GSECA.

SSA is a reformulated version of the classical gene set screening method proposed by Segal *et al.* (2004). SSA shows by far the worst performance in all the benchmark tests. This result seems reasonable when considering that SSA tests over or underexpression of a gene set in each single sample individually, while the other methods take into account expression profiles of a gene set across multiple samples.

On the other hand, EEM, which tests expression coherence across all samples, performs broadly well for various types of input data. BEEM, which tests over or underexpression in a subgroup of samples, works well for the GNF expression dataset, confirming our previous report (Niida *et al.*, 2010). However, in the other expression datasets, EEM outperforms BEEM; bicluster-type expression modules targeted by BEEM are less common than coherent expression modules targeted by EEM in most real expression datasets. The feature of these methods is that they are based on the rigid model assumptions. There are pros and cons about this feature. A notable advantage of this feature is that it can be utilized for not only gene set screening, but also expression module discovery. If a gene set is taken as significant by EEM and BEEM, a subset of genes should be co-expressed across all samples or in a subgroup of samples. EEM and BEEM can extract the subset as an expression module. This expression module discovery approach

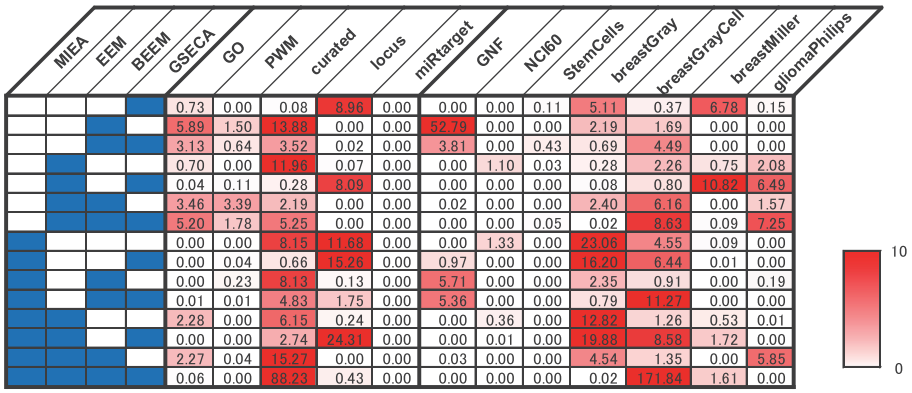

Fig. 4. The association of the 15 expression-gene set pair groups with the types of input data. We measured associations of each group of expression-gene set pairs with the seven expression datasets or the five gene sets using hypergeometric tests. Each row of the table represents each group and colored cells indicate which methods take it as significant. *P*-values from the association tests are presented in minus log scale.

brings us much biological knowledge, especially when combined with additional analyses of the extracted module information (Niida *et al.*, 2009a, b, 2010). However, note that EEM and BEEM potentially fail to capture expression profiles which do not fit to their expression module models.

Similarly to EEM, GSECA assumes expression coherence of input gene sets as an indication of its functionality. However, while EEM tests existence of a coherent subset in the input gene set, GSECA evaluates overall coherence by calculating the sum of correlations for all gene pairs in the gene set (Kim *et al.*, 2007). In the real data test, this approach retrieves the largest number of gene sets as significant, possibly by capturing loosely coherent patterns that cannot be detected by EEM. An example of such an expression profile is given in Supplementary Figure S5A. However, note that GSECA cannot capture a gene set containing subsets whose expressions are anti-correlated, because the coherence signals cancel each other out. Supplementary Figure S5B shows an example taken as significant by MIEA, EEM and BEEM, but not GSECA. This case was also shown in the simulated data test using the sine-wave model (Supplementary Fig. S1E and G) .

The newly introduced method, MIEA, is designed to detect any informative expression profiles of the input gene sets without assuming any specific models. This notable feature endows the method with the broad applicability so that it can capture complex expression profiles that the other model-based methods fail to detect. Supplementary Figure 5C and D shows such complex expression profiles, which are hard to be captured by any simple models. However, note that this broad applicability comes at cost: in case where a model-based method would be appropriate, the model-free method has less power, demonstrated by the simulated data test using the coherent and bicluster model. (Fig. 1A and D) .

Collectively, we conclude that MIEA is broadly applicable to various types of gene sets and expression datasets, while EEM and GSECA also have comparably broad applicability. Since EEM and GSECA have biased preferences for their targets, they might perform better than MIEA when applied to their favorite target. Although each method alone seems to have enough performance, combining these different methods would enable more comprehensive screenings for functional gene sets. This study not only introduced MIEA but also revealed unique characteristics of the novel and existing unsupervised gene set screening methods by performing extensive benchmark tests. Taken together with our previous studies (Niida *et al.*, 2009a, b, 2010), this study provides a foundation for unsupervised gene set screening analysis, a novel paradigm in transcriptome data analysis.

ACKNOWLEDGEMENTS

Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo.

Funding: This work was supported by the Grant-in-Aid for the Global COE Program 'Center of Education and Research for the Advanced Genome-Based Medicine' from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

Conflict of Interest: none declared.

REFERENCES

Bergmann,S. *et al*. (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **67**, 031902.

Carthew,R. and Sontheimer, E. (2009) Origins and mechanisms of mirnas and sirnas. *Cell*, **136**, 642–655.

Chin,K. *et al*. (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**, 529–541.

Fisher,R. (1932) *Statistical Methods for Research Workers.* Oliver and Boyd, London.

Griffiths-Jones,S. *et al*. (2008) mirbase: tools for microrna genomics. *Nucleic Acids Res.*, **36**, D154–D158.

Huang,D.W. *et al*. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Kim,T. *et al*. (2007) Inferring biological functions and associated transcriptional regulators using gene set expression coherence analysis. *BMC Bioinformatics*, **8**, 453.

Matys,V. *et al*. (2006) Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Miller,L. *et al*. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl Acad. Sci. USA*, **102**, 13550–13555.

Müller,F. *et al*. (2008) Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, **455**, 401–405.

Neve,R. *et al*. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, **10**, 515–527.

Niida,A. *et al*. (2009a) Gene set-based module discovery in the breast cancer transcriptome. *BMC Bioinformatics*, **10**, 71.

Niida,A. *et al*. (2009b) A novel meta-analysis approach of cancer transcriptomes reveals prevailing transcriptional networks in cancer cells. *Genome Informatics*, **22**, 121–131.

Niida,A. *et al*. (2010) Gene set-based module discovery decodes cis-regulatory codes governing diverse gene expression across human multiple tissues. *PLoS One*, **5**, e10910.

Phillips,H. *et al*. (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*, **9**, 157–173.

Press,W. *et al*. (1992) *Numerical Recipes in c*, 2nd edn. *Cambridge University Press*, Cambridge.

Segal,E. *et al*. (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.

Shankavaram,U. *et al*. (2007) Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study. *Mol. Cancer Ther.*, **6**, 820–832.

Su,A. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15278–15279.