Gene 518 (2013) 179–186

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Random forests-based differential analysis of gene sets for gene expression data

Huey-Miin Hsueh [a], Da-Wei Zhou [a], Chen-An Tsai [b,*]

[a] Department of Statistics, National Chengchi University, Taiwan
[b] Department of Agronomy, National Taiwan University, Taiwan

ARTICLE INFO

Article history:
Accepted 27 November 2012
Available online 6 December 2012

Keywords:
DNA microarray
Gene-set analysis
Random Forests

ABSTRACT

In DNA microarray studies, gene-set analysis (GSA) has become the focus of gene expression data analysis. GSA utilizes the gene expression profiles of functionally related gene sets in Gene Ontology (GO) categories or priori-defined biological classes to assess the significance of gene sets associated with clinical outcomes or phenotypes. Many statistical approaches have been proposed to determine whether such functionally related gene sets express differentially (enrichment and/or deletion) in variations of phenotypes. However, little attention has been given to the discriminatory power of gene sets and classification of patients.

In this study, we propose a method of gene set analysis, in which gene sets are used to develop classifications of patients based on the Random Forest (RF) algorithm. The corresponding empirical *p*-value of an observed out-of-bag (OOB) error rate of the classifier is introduced to identify differentially expressed gene sets using an adequate resampling method. In addition, we discuss the impacts and correlations of genes within each gene set based on the measures of variable importance in the RF algorithm. Significant classifications are reported and visualized together with the underlying gene sets and their contribution to the phenotypes of interest.

Numerical studies using both synthesized data and a series of publicly available gene expression data sets are conducted to evaluate the performance of the proposed methods. Compared with other hypothesis testing approaches, our proposed methods are reliable and successful in identifying enriched gene sets and in discovering the contributions of genes within a gene set. The classification results of identified gene sets can provide an valuable alternative to gene set testing to reveal the unknown, biologically relevant classes of samples or patients.

In summary, our proposed method allows one to simultaneously assess the discriminatory ability of gene sets and the importance of genes for interpretation of data in complex biological systems. The classifications of biologically defined gene sets can reveal the underlying interactions of gene sets associated with the phenotypes, and provide an insightful complement to conventional gene set analyses.

1. Introduction

Biological phenomena often occur through the interactions of multiple genes via signaling pathways, genetic networks, or other functional relationships. In DNA microarray studies, single gene analyses can only take into account a small portion of genetic variation in the complex biological system. In contrast to single gene analyses, a gene-set analysis (GSA) is used to evaluate the association between the expression of biological pathways, or a priori defined gene sets, and a particular phenotype. Genes that serve a common molecular function, a biological process, or a cellular component are annotated to the same term and grouped together into sets. The annotation terms can be obtained from public-domain web-libraries such as Gene Ontology (GO), KEGG, BioCarta and the Broad Institute. See Pang et al. (2006) and Delongchamp et al. (2006). This helps biologists to interpret the selected sets of genes in a manner of gene regulation mechanism from the microarray data.

Many statistical methods are proposed for gene-set data analysis in literatures. Mootha et al. (2003) and Subramanian et al. (2005), first proposed the Gene Set Enrichment Analysis (GSEA), in which they consider the distributions of entire genes in a gene set, rather than a subset from the list of differential expression genes, and then use some statistic to assess the significance of predefined gene sets. These existing approaches are not only distinct in terms of the test statistic used, but also differ in terms of the null hypothesis and hence differ in the problem of research interest. Tian et al. (2005) and Goeman and Buhlmann (2007) summarize the methods into two types: competitive and self-contained tests. The null hypothesis of a competitive test is that the specific gene set is not differentially expressed when compared to other gene sets. This method involves not only the gene set of research interest but also the full data set. The sampling unit in construction of the empirical null distribution for calculating a *p*-value is the gene. A

Abbreviations: GSA, gene-set analysis; GO, Gene Ontology; RF, Random Forest algorithm; OOB, out-of-bag error rate.

* Corresponding author. Tel.: +886 2 3366 4775; fax: +886 2 2362 0879.
E-mail addresses: hsueh@nccu.edu.tw (H.-M. Hsueh), 98354014@nccu.edu.tw (D.-W. Zhou), catsai@ntu.edu.tw (C.-A. Tsai).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.034

positive finding can be obtained only when the gene set contributes to the statistical variation of a phenotype of interest more than other gene sets. On the other hand, the self-contained test is utilized to determine whether the gene set is differentially expressed. The analysis is conducted with respect to the specific gene set data alone. This approach evaluates *p*-values by permuting samples using the conventional approach. Following the idea of GSEA, many statistical methods have been proposed, such as the global test (Goeman et al., 2004), approaches similar to the two-sample t-test (Tian et al., 2005), the ANCOVA test (Mansmann and Meister, 2005), the Hotelling's T2 test (Kong et al., 2006), the MaxMean approach (Efron and Tibshirani, 2007), the SAM-GS test (Dinu et al., 2007), the global statistics approach (Chen et al., 2007), the Random-sets method (Newton et al., 2007), the Logistic Regression (LRpath) approach (Sartor et al., 2009), and the MANOVA test (Tsai and Chen, 2009) amongst others. These approaches rely on different statistical assumptions and consider different data structures, which then usually leads to different findings even when they are applied to the same data set. A comprehensive review of these methodologies can be found in, for example, Goeman and Buhlmann (2007) and Nam and Kim (2008). Fridley et al. (2010) also provided intensive empirical comparisons on the self-contained analysis. As referenced above, none of their methods have addressed the feasibility of a pre-defined gene set in discriminating different phenotypes.

On the other hand, various machine learning-type algorithms, which take various biological information into consideration, have been developed from a classification perspective. For example, Lin et al. (2006) demonstrated that accuracy and robustness of a classification in analyzing microarray data can be improved by considering the existing biological annotations. Wei and Li (2007) applied a boosting-based method for a nonparametric pathways-based regression (NPR) analysis. Although the NPR generates an improved prediction, there is not a selection criterion to identify differential gene sets. Tai and Pan (2007a, 2007b) proposed a group penalization method that incorporates biological information to build a penalized classifier. Lottaz and Spang (2005) provided a biologically focused classifier, such as StAM, based on the GO hierarchical structure. This method has a limitation that only the genes annotated in the leaf nodes of the GO tree can be used as the predictors, while other genes (relevant, but not annotated yet) cannot be used. However the biological information of gene sets may come from different databases, such as KEGG or BioCarta, and are not limited to the GO annotation only.

Recently, the Random Forest algorithm, developed by Breiman (2001), has gained popularity for use in microarray data analysis due to its flexibility in terms of the type and the dimension of the input data, the absence of overfitting, and a predictive performance comparable to other machine learning methods. See Huang et al. (2005), Díaz-Uriarte and Alvarez de Andrés (2006), Statnikov et al. (2008), and Boulesteix et al. (2008). Pang et al. (2006) and Pang and Zhao (2008) used the Random Forest classification and regression based on pathway information. They proposed a rank analysis of pathways in terms of the predictive performance of the Random Forest built on the pathway. However, no biological variation is taken into account, and hence no confirmatory conclusion can be made based on the evidence. Here, the Random Forest algorithm will be employed to link a gene set and the phenotypic response. The correspondent predictive performance will be used to reveal the strength of the association between the gene set and the phenotype. In addition, the resultant statistical evidence, considering the biological variation, will be obtained.

In this paper, we propose a self-contained GSA method that can not only identify differential gene sets, which are significantly associated with the variation of phenotypes, but can also assess the impacts of individual genes on a prediction model. The Random Forests algorithm will be applied to develop a classifier based on the gene set. The empirical *p*-value of the performance of the classifier will be obtained by using the permutation test to evaluate the statistical significance of the gene set. In addition, during the analysis, we integrate the classification results from the identified gene sets to uncover potential associations between gene sets and phenotypes. Our proposed approach is compared with some existing GSA approaches based on the performance of synthesized data sets and a series of publicly available microarray data.

2. Materials and methods

Consider a microarray study of size n and one k-class phenotype. Assume the gene set or pathway S including m genes is of interest. In contrast to the competitive test, where a relative conclusion is made upon a comparison with the whole gene set, the self-contained test, which seeks an absolute association between the gene set and the phenotype, is studied here. The null hypothesis of a self-contained test of the gene set S is given as

$\mathbf{H_0^S}$. The gene set S is independent with the phenotype variable.

To collect more information on multiple genes in a gene set, a complex classifier is constructed and its test set error rate is recorded. The lower the error rate, the more the evidence shows that the gene set is associated with the phenotypes. Hence the test set error rate is utilized as a test statistic of the self-contained test, and the correspondent *p*-value is applied to draw a statistical conclusion.

We consider using the Random Forests classifier (Breiman, 2001). The Random Forest is based on an ensemble of many classification trees, in which every one of them is constructed based on a bootstrap sample out of the original dataset, which is then split. For each tree, the algorithm randomly selects input variables as potential predictors. The observations outside the bootstrap sample are called the out-of-bag (OOB) data and are used for calculating a test set error rate of the tree. Every subject is likely to be OOB in one-third of the bootstrappings and is predicted under those circumstances. When the specified numbers of trees are added to the forest, there is a final prediction for each subject by aggregating these predictions. Typically, the classification with the most votes (majority vote) over all the trees in the forest is considered. Summarizing the deviations between the observed phenotypes and their predictions produces the OOB test set error rate. This error rate reveals the association between the gene set and the phenotype. A gene-set with a lower error rate is regarded to have a better predicting power with regard to the phenotype variable and hence has a greater significance. Breiman (2001) indicated that unlike the cross-validation, the OOB error rate provides an unbiased estimate of the error rate. Moreover, applying classification trees makes the method time-efficient.

Given an observed OOB test set error rate e_0 in the Random Forest, a permutation-based *p*-value can be obtained as following,

$$p-\text{value}=\frac{\sum_{k=1}^{N} I\left\{e^{(k)}\le e_0\right\}}{N}, \tag{1}$$

Table 1
Type I error rate comparisons in the simulation study. Type I error rates of eight GSA methods: RF, Hotelling's T2, PCA, SAM-GS, ANCOVA, Global, GSEA, and MaxMean tests.

Method	$\rho=0$	$\rho=0.3$	$\rho=0.5$	$\rho=0.9$
Hotelling's T^2	0.050	0.039	0.038	0.050
PCA	0.053	0.042	0.052	0.062
SAM-GS	0.046	0.042	0.038	0.055
ANCOVA	0.042	0.038	0.034	0.052
Global	0.001	0.009	0.016	0.034
GSEA	0.059	0.058	0.052	0.048
MaxMean	0.093	0.094	0.107	0.098
RF	0.040	0.034	0.027	0.036

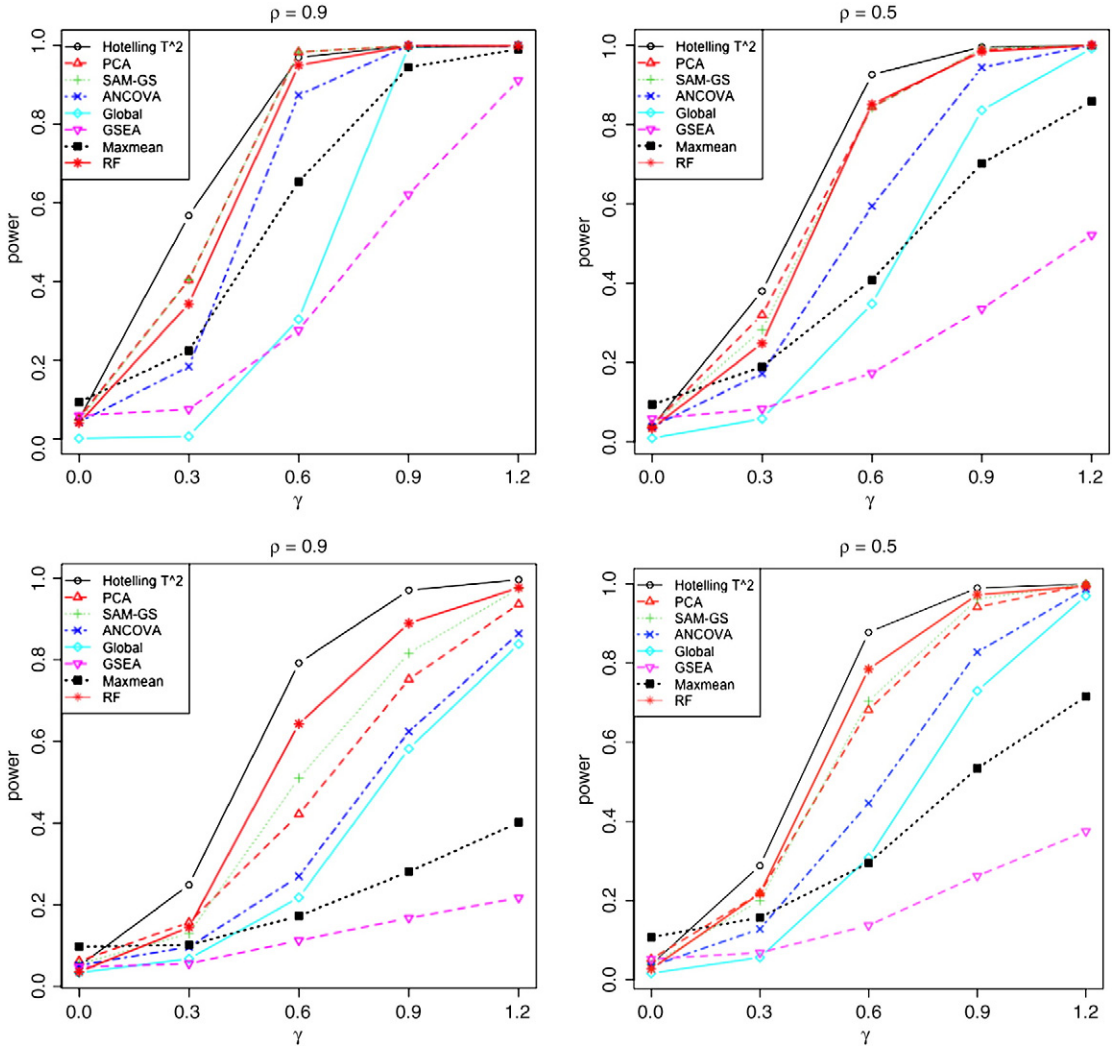

Fig. 1. Power analysis of simulation data. Power analysis of eight self-contained tests: Hotelling's T2, PCA, SAM-GS, ANCOVA, Global, GSEA, MaxMean and our proposed method, RF.

where $e^{(k)}$ is the test set error rate of the Random Forest classifier based on the k-th permutation sample and N is the number of permutations. The way of permutation is dependent on the null hypothesis to be tested. For a self-contained test of the gene set S, the null distribution is produced by shuffling the phenotype labels of the n samples in each run. Along with the gene set, the randomized labels are used for construction of a Random Forest classifier and generation of an OOB test set error rate. Repeat the process for N runs. The p-value against H_0^S can be found by Eq. (1) and the statistical conclusion can be drawn. The gene set is then said to be significantly expressed if its correspondent p-value is less than or equal to the predetermined significance level α. In this study, the calculations are based on the randomForest package (Liaw and Wiener, 2002) in R language.

To measure the importance of predictor variable X_j, the Random Forest algorithm assesses the importance of a variable by looking at how much prediction error increases when that variable of the (OOB) data is permuted while remaining variables are left unchanged. If the original variable X_j is associated with the phenotypes, the prediction

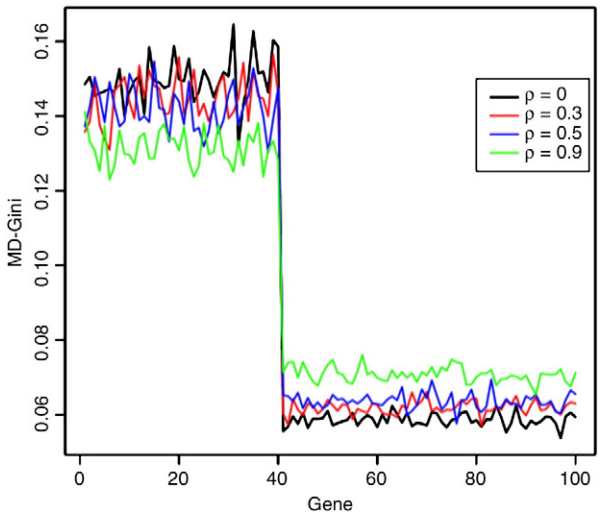

Fig. 2. Importance plot. Graphical representation of MDG importance measures of 100 genes in the simulation study.

Table 2
Summary of datasets. Data summary of the four microarray datasets from the GSEA website.

Dataset	No. of subjects	Genes	No. of gene sets[a]	Phenotype
Breast cancer	49	22,215	444	Three tumor types
P53	50	12,625	440	p53+/p53 mutant
Gender	32	22,283	480	Male/female
Lung_a	86	7,129	419	Normal/tumor

[a] Gene sets were retrieved from various sources of pathway databases, including KEGG and BioCarta.

Table 3
Significant pathways in Breast cancer data. The results of the top 20 significant pathways in Breast cancer data.

Pathway	Gene set size	Observed error rate	Null error rate mean (Std)	*p*-Value
BC-regulation of BAD phosphorylation	24	0.0816	0.5098 (0.0582)	<0.0005
BC-GATA3 participate in activating the Th2 cytokine genes expression	21	0.0816	0.5147 (0.0622)	<0.0005
BC-CARM1 and regulation of the estrogen receptor	24	0.0816	0.5088 (0.0584)	<0.0005
Jak-STAT signaling pathway	71	0.0816	0.5035 (0.0519)	<0.0005
Fructose and mannose metabolism	39	0.0816	0.5052 (0.0572)	<0.0005
Glycolysis-gluconeogenesis	68	0.0816	0.5018 (0.0556)	<0.0005
Carbon fixation	25	0.1020	0.5175 (0.0628)	<0.0005
Estrogen-response	10	0.1020	0.5244 (0.0669)	<0.0005
Downregulated of MTA-3 in ER-negative breast tumors	19	0.1020	0.5203 (0.0639)	<0.0005
Pentose phosphate pathway	22	0.1020	0.5118 (0.0621)	<0.0005
Valine, leucine and isoleucine degradation	46	0.1020	0.5047 (0.0528)	<0.0005
mCalpain and friends in Cell motility	33	0.1020	0.5043 (0.0562)	<0.0005
Sulfur metabolism	9	0.1224	0.5334 (0.7260)	<0.0005
Phenylalanine metabolism	22	0.1224	0.5109 (0.0595)	<0.0005
Map kinase inactivation of SMRT corepressor	16	0.1224	0.5193 (0.6710)	<0.0005
Interleukin-2 receptor β-chain in T-cell activation	45	0.1224	0.5072 (0.0580)	<0.0005
Role of ERBB2 in signal transduction and oncology	29	0.1224	0.5119 (0.0599)	<0.0005
Trefoil factors initiate mucosal healing	33	0.1224	0.5056 (0.0573)	<0.0005
Tyrosine metabolism	37	0.1224	0.5036 (0.0583)	<0.0005
Tryptophan metabolism	94	0.1224	0.4979 (0.0491)	<0.0005

error increases substantially. Therefore, the mean decrease in accuracy (MDA) and the mean decrease in the Gini index (MDG) are respectively defined as the difference in prediction accuracy, or the node impurity before and after permuting X_j, normalized by the number of trees in the Random Forest. The Gini index describes the node impurities as a splitting criterion of trees in order to measure how well a potential split does in separating the samples of the two classes in this particular node. Here, we use the MDG as the importance measure of genes across different gene sets.

3. Results

3.1. Simulation study

We conducted extensive simulations to investigate the performance of our proposed approach when compared to the existing methods. The simulation design was similar to that considered by Liu et al. (2007) and Tsai and Chen (2009). Consider 100 genes in each gene set ($m=100$) and one phenotype variable with a binary category, say, normal and diseased. In each phenotype group, there are 10 samples of gene expression. The gene expressions of a sample are generated from a multivariate normal distribution. Regarding the synthetic distribution, the means of the 100 genes in the normal group are first iid produced from a uniform (0, 10) distribution. After that, the means of the 100 genes in the diseased group are consecutively determined by adding $2r$ in means of the first 20 genes; subtracting $2r$ in the means of the next 20 genes, and not altering in the remaining 60 genes, where r ranges from zero to 1.2. That is, in the gene set of size 100, forty of them are differentially expressed, of which half are up-regular expressed and the other half are down-regular expressed. For simplicity, the gene expressions of the two groups are assumed to share the same covariance matrix, where the variances of the 100 genes are iid from the uniform (0.1, 10) distribution, and the pairwise correlation coefficient between the 100 genes has the following form:

$$\rho_{jl} = \begin{cases} \rho, & 1 \le j \ne l \le 20 \\ \rho, & 21 \le j \ne l \le 40 \\ \rho, & 41 \le j \ne l \le 120 \end{cases},$$

where $\rho = 0$, 0.3, 0.5 and 0.9. In the following, 5,000 classification trees are added in every Random Forest classifier, and 1,000 permutations are used for every *p*-value. The empirical performance is based on 1,000 repetitions. Tsai and Chen (2009) provided the empirical results of seven existing self-contained methods. To compare these seven methods with ours, Table 1 presents their empirical type I error rates, and Fig. 1 shows their empirical powers at significance level 5%. Our method is denoted as RF.

From Table 1, we find that our proposed method controls the type I error rate well. Indeed, the error rate is far below the nominal level in all cases. It indicates that our proposed method gives a conservative conclusion. Such conservativeness may lead to power loss in detecting a difference. The power analysis in Fig. 1 shows that our method is not much more inferior than other methods while adequately controlling the type I error rate. It is frequently seen that Hotelling's T2 has a greater power. However, the empirical type I error rate of Hotelling's T2 is exactly 5% in some scenarios, as seen in Table 1, and it leads us to believe that the true type I error rate is likely to exceed the nominal level. This method tends to be liberal, but only to a slight extent. On the other hand, PCA, which is more powerful when no or small correlations exist between genes, is too liberal as well. The fact that Hotelling's T2 and PCA are liberal produces an optimistic performance in power analysis. When the correlations between genes are zero or small, our method is comparable to SAM-GS; whereas in other case, it is more powerful than these methods. Note that in this simulation, the genes

Table 4
Significant pathways in P53 data. The results of the top 20 pathways in P53 data.

Pathway	Gene set size	Observed error rate	Null error rate mean (Std)	*p*-Value
SA_G1_AND_S_PHASES	24	0.12	0.3894 (0.0468)	<0.0005
g2 Pathway	44	0.14	0.3815 (0.0425)	<0.0005
SA PROGRAMMED CELL DEATH	24	0.14	0.3883 (0.0470)	<0.0005
Mitochondria pathway	33	0.16	0.3821 (0.0436)	<0.0005
DNA_DAMAGE_SIGNALLING	49	0.16	0.3698 (0.0359)	<0.0005
p53hypoxia pathway	40	0.16	0.3814 (0.0441)	<0.0005
Bad pathway	41	0.18	0.3785 (0.0412)	<0.0005
bcl2family and reg. network	59	0.18	0.3777 (0.0426)	0.0005
p53 Pathway	40	0.18	0.3843 (0.0459)	0.0005
Chemical pathway	44	0.20	0.3800 (0.0423)	<0.0005
Cell cycle pathway	47	0.20	0.3824 (0.0439)	<0.0005
P53_signalling	153	0.20	0.3635 (0.0299)	<0.0005
Calcineurin pathway	36	0.20	0.3856 (0.0462)	0.0005
Cell_cycle_arrest	46	0.20	0.3783 (0.0402)	0.0010
Cell_cycle	130	0.22	0.3722 (0.0360)	0.0005
g1 Pathway	63	0.24	0.3750 (0.0506)	<0.0005
pml Pathway	37	0.24	0.3848 (0.0449)	0.0020
P53_UP	56	0.24	0.3775 (0.0388)	0.0035
eponfkb Pathway	21	0.24	0.3914 (0.0516)	0.0060
hivnef Pathway	111	0.26	0.3674 (0.0345)	0.0040

Table 5
Significant pathways in gender data. The results of the top 20 pathways in gender data.

Pathway	Gene set size	Observed error rate	Null error rate mean (Std)	*p*-Value
Testis genes from XHX and NETAFFX	111	0.0312	0.5543 (0.1202)	<0.0005
GNF female genes	116	0.0625	0.5421 (0.1192)	<0.0005
ST dictyostelium discoideumcAMP chemotaxis pathway	55	0.1562	0.5467 (0.1195)	<0.0005
WILLARD_INACT	31	0.1875	0.5427 (0.1149)	<0.0005
XINACT	34	0.1875	0.5374 (0.1188)	0.0010
SIG_Regulation_of_the_actin_cytoskeleton_by_Rho_GTPases	67	0.1875	0.5442 (0.1223)	0.0015
MAP00252 alanine and aspartate metabolism	36	0.2500	0.5394 (0.1192)	0.0100
MAP00910 nitrogen metabolism	36	0.2812	0.5432 (0.1177)	0.0150
SIG CHEMOTAXIS	85	0.2812	0.5475 (0.1229)	0.0155
RAP DOWN	434	0.3125	0.5485 (0.1247)	0.0220
rb Pathway	28	0.3125	0.5400 (0.1174)	0.0325
cdc25 Pathway	20	0.3125	0.5320 (0.1189)	0.0345
rab Pathway	29	0.3125	0.5304 (0.1172)	0.0400
set Pathway	20	0.3125	0.5318(0.1205)	0.0425
INSULIN 2F UP	405	0.3438	0.5435 (0.1255)	0.0715
Intrinsic pathway	38	0.3750	0.5432 (0.1148)	0.0910
vegf Pathway	49	0.3750	0.5412 (0.1234)	0.1100
MAP00970 aminoacyl tRNA biosynthesis	33	0.3750	0.5355 (0.1163)	0.1120
pepi Pathway	11	0.3750	0.5303 (0.1159)	0.1165
Androgen genes from NETAFFX	125	0.3750	0.5544 (0.1189)	0.1400

are differentially expressed in terms of having different class means. This setup favors those methods, of which the test statistic emphasizes deviations in means, such as Hotelling's T2 and SAM-GS. However, our method, involving a complex classifier, is expected to be capable to obtain more sophisticated and subtle associations between gene sets and phenotype variables.

As an example, Fig. 2 is the plot of the average values of mean decrease in the Gini (MDG) for evaluating the importance of genes with $\rho = 0$, 0.3, 0.5 and 0.9, when $r = 0.6$. Genes with large values of MDG indicate strong associations with phenotypes. As expected, the informative genes (the first 40 genes) appear to have much higher MDG values than the non-informative genes. Also, we observe that the difference of MDG between informative and non-informative genes decreases as the correlation between genes increases. In summary, the importance measure of MDG is helpful in discovering differentially expressed genes and in obtaining the order of the association effect.

3.2. Application to gene set data

The proposed self-contained test is applied to four publicly available gene expression data sets: Breast cancer, P53, Gender and Lung_a data sets. The data are available in http://bioinformatics.med.yale.edu/pathway-analysis/rf.htm. Table 2 reports the summary of every data set. The Breast cancer data set consists of forty-nine patients, which can be classified into three tumor classes based on steroid receptor activity: luminal, basal and molecular apocrine, see Farmer et al. (2005). In P53 data set, 17 of 50 NCI-60 cell lines are classified as normal and 33 are classified as carrying mutations in the gene. The Gender data set compares 15 male lymphoblastoid cell lines with 17 female cell lines. In a lung cancer study in Michigan (Subramanian et al. (2005)), 86 gene-expression profiles in tumor samples from patients with lung adenocarcinomas are reported along with their clinical outcomes. The research interest is to determine the significance of more than four hundreds pathways from KEGG, BioCarta or manually. In applying our proposed test, the tree size of the Random Forest is 50,000. All pathways are first ranked by their observed OOB test set error rate. The top 20 pathways are further analyzed for their null distribution of the test statistic by using 2000 permutations. Tables 3–6 report the mean, standard deviation of the null distribution and the *p*-value of the 20 pathways.

Pang et al. (2006) had built Random Forest classifiers for the pathways of these data sets. However, they only provided the observed OOB test set error rates of the classifiers in their article, and hence only a rank analysis of pathways was performed. No conclusions of statistical significance are available from their results. Furthermore,

Table 6
Significant pathways in lung cancer data. The results of the top 20 pathways in lung cancer data.

Pathway	Gene set size	Observed error rate	Null error rate mean (Std)	*p*-Value
Protein export	7	0.2326	0.3213 (0.0314)	0.0065
TSP-1 induced apoptosis in microvascular endothelial cell	10	0.2326	0.3160 (0.0284)	0.0145
Proepithelin conversion to epithelin and wound repair control	11	0.2326	0.3162 (0.0284)	0.0150
Actions of nitric oxide in the heart	34	0.2442	0.2957 (0.0190)	0.0095
Phenylalanine and tyrosine catabolism	9	0.2442	0.3176 (0.0300)	0.0205
Caspase cascade in apoptosis	28	0.2558	0.2981 (0.0208)	0.0170
Cycling of Ran in nucleocytoplasmic transport	5	0.2558	0.3320 (0.0345)	0.0245
TGF-beta signaling pathway	71	0.2558	0.2892 (0.0148)	0.0267
Blood group glycolipids	15	0.2558	0.3074 (0.0234)	0.0285
Glycolysis – glucone	68	0.2558	0.2920 (0.0162)	0.0290
Control of gene expression by vitamin D receptor	21	0.2558	0.3016 (0.0218)	0.0295
Ascorbate and aldarate metabolism	8	0.2558	0.3140 (0.0270)	0.0300
Phospholipid degrada	8	0.2558	0.3131 (0.0268)	0.0300
IL 4 signaling pathway	12	0.2558	0.3117 (0.0262)	0.0325
Nuclear receptors in lipid metabolism and toxicity	33	0.2558	0.2944 (0.0187)	0.0340
Carbon fixation	21	0.2558	0.3019 (0.0213)	0.0345
CARM1 and regulation of the estrogen receptor	24	0.2558	0.2997 (0.0207)	0.0350
CD40L signaling pathway	13	0.2558	0.3086(0.0254)	0.0370
Platelet amyloid precursor protein pathway	19	0.2558	0.3057 (0.0241)	0.0410
TNFR1 signaling pathway	38	0.2558	0.2967 (0.0202)	0.0450

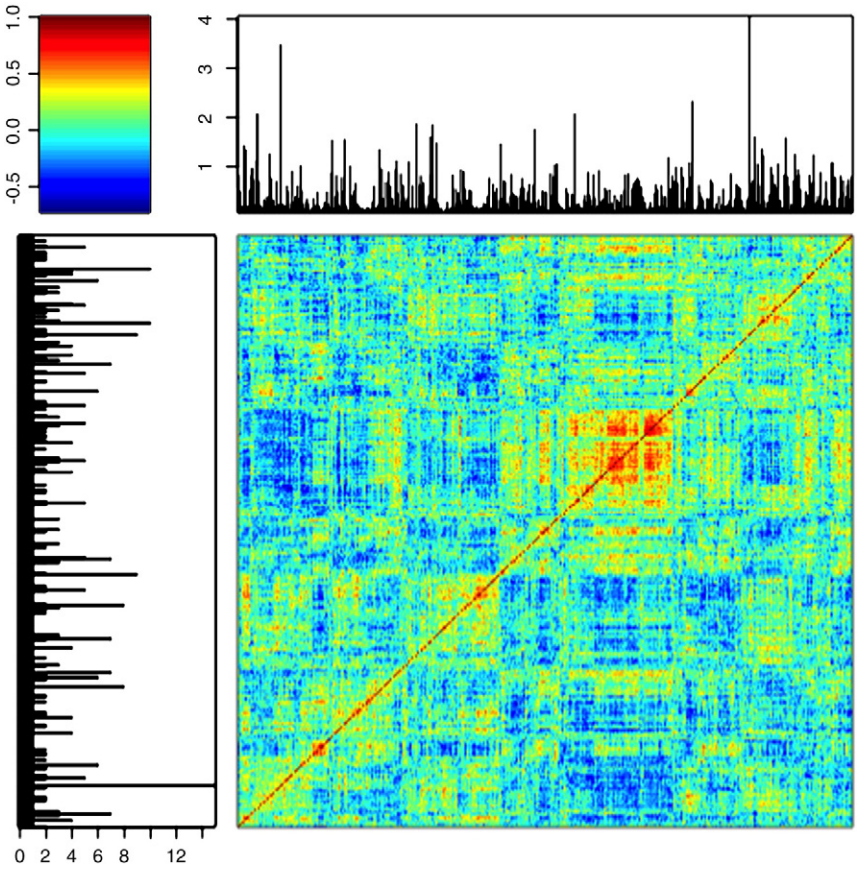

Fig. 3. Correlation plot. Heatmap of pairwise correlation coefficients for 461 genes. The upper panel displays the MDG for each gene. The barplot in the lower left panel displays the number of functional groups that each gene appears in.

since ties occur frequently among the observed error rates of the pathways, only a crude comparison between the pathways can be made. In contrast, the null distributional information in Tables 3–6 gives more details in distinguishing between the significance levels of pathways. By taking the data variation into account, we see that pathways of the same observed error rate can have varied *p*-values and different statistical conclusions.

In performing multiple hypothesis tests, the statistical conclusion is drawn under a guarantee of either a per-comparisonwise or familywise control on the false positive rate. To control a familywise error rate or a false discovery rate, some multiplicity adjustment is always required and is employed on the *p*-values. Here for simplicity, we consider controlling a per-comparisonwise error rate. At significance level 5%, all the 20 pathways have significant association with the phenotype in Breast cancer, P53 and the Lung_a data set. On the other hand, only 14 pathways in the Gender data set are significant. The permutation *p*-value takes into account the biological variation, which reflects the dependence structure between the genes, in each gene set. Therefore, it is not a monotone function of the observed error rate due to distinct biological variations. That is, one gene set may have a smaller *p*-value than the other, while its observed error rate is higher. For example, in P53, the observed error rate of the cell_cycle_arrest pathway is 0.2, smaller than that of the g1Pathway (which is 0.24). However, the latter is more significant in terms of *p*-value. The permutation test constructs the null distribution under a certain null hypothesis in order to evaluate how likely it is that the observed error rate would be obtained by chance.

In a biological system, genes or gene sets do not regulate independently and there is a wide range of interactions between them. Here we combine the importance measure with the correlation between genes to investigate whether correlated genes have an effect on the importance measures. Fig. 3 shows the pairwise correlation coefficients in a color scheme across 461 genes obtained from the top 20 highly enriched gene sets in the P53 dataset. We observe that there is a clear clustering pattern between the correlation and importance measure. Most genes are not highly involved in common functional groups. We are mostly interested in those genes in the enrichment gene sets with high importance measures. In Table 7, we identify five "most important" genes that cause at least a 2.0% decrease in the Gini importance (MDG) when information from that gene is removed. Also, we observe that they all are not highly correlated, as shown in Fig. 3, and the CDKN1A gene appears in 15 pathways out of the 20 significant gene sets. The p53 protein plays an important role in tumor suppression and controls multiple cellular functions, including induction of apoptosis, growth arrest, and regulation of angiogenesis. Several p53-responsive genes have been verified to be regulated by p53 (Riley et al., 2008). As referenced and confirmed in this study, the BAX and CDKN1A genes respectively induce apoptosis and cell cycle arrest in response to DNA damage. The results

Table 7
High MDG genes in P53 data. The MDG rankings for the 5 genes in P53 data.

Gene ID	Number of functional groups involved	MDG
BAX	9	4.027
CDKN1A	15	3.526
FAS	1	2.515
IL1A	1	2.157
CDKN2D	5	2.151

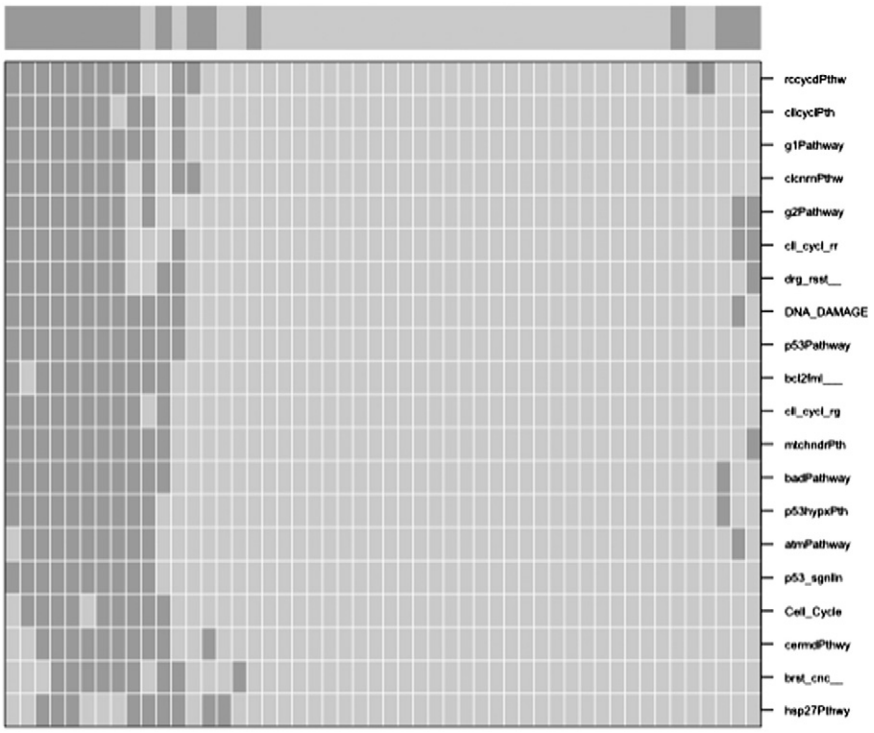

Fig. 4. Annotation-based classifications for the top 20 significant pathways in the P53 dataset. Columns correspond to samples and rows correspond to gene sets. The sample classes are color coded, dark gray for normal and light gray for mutation samples. Each row represents the classification of the corresponding gene set. The image is clustered in both columns and rows in order to bring similar classifications and sample profiles close together. The top horizontal side bar indicates the original status of samples.

indicate that the importance measure is useful for understanding biological mechanisms induced by p53 activation.

3.3. Annotation-based classification

Previous analyses have demonstrated the power of our proposed approach to simultaneously identify differentially expressed gene sets and to assess the impacts of individual genes. We observe that the classifications based on the annotated gene sets reveal interactions between gene sets and phenotypes. Of course, gene sets with functional annotations provide a better understanding for classification of samples than randomly selected sets of genes. In particular, we are interested in the classifications corresponding to the differentially expressed gene sets for the functional mechanism of their supporting genes.

We use the p53 microarray data set for exploring the annotation-based classification. The p53 dataset identifies targets of the transcription factor p53 from 10,100 gene expression profiles in the NCI-60 collection of cancer cell lines. The mutation status of the p53 gene has been reported for 50 of the NCI-60 cell lines with 17 normal and 33 mutation samples. Fig. 4 displays the resulting classifications of phenotype groups based on the top 20 significant gene sets. After clustering the classifications, we can better understand the patterns of gene set-based classifications. We see that all annotation-based classifications separate two classes of samples well, and most classifications have a higher accuracy on mutation samples. Interestingly, two normal samples are consistently classified as mutation samples by 20 significant gene sets. The results indicate that the deferential gene sets justify the splitting of samples into two phenotype groups.

4. Discussion

In this article, we propose a self-contained test by using the performance of the Random Forest classifier built on specific sets of functionally related genes to determine whether the set has a significant association with a phenotypic-class variable. The simulation studies show that our method is conservative in the sense of a well-controlled type I error rate. Further, the power loss brought by being conservative is not severe; the method has comparable performance to other methods in power analysis. Most existing approaches focus on identifying differentially expressed gene sets associated with phenotypes of interest based on the framework of statistical tests. The point is that any informative pathway should have the ability not only to differentiate subtle expression level changes, but also to generate a good prediction of phenotypes. From a practical perspective, the Random Forest algorithm seems to perform well for classification purposes and can provide measures of gene importance for each gene set. Overall, our approach provides a complementary strategy for identifying gene sets associated with phenotypes of interest. Several real examples are analyzed to illustrate the applicability of our proposed methods.

Ideally, a self-contained test reveals a marginal association of the gene set with the phenotype, and a competitive test seeks for the conclusion with some baseline adjustment. However, the definition of a competitive test is not clear. The null hypothesis given above indicates that there exists at least one other gene set that is more differentially expressed than the specific gene set. Consequently, one will not obtain a significant result unless the specific gene set has the strongest association with the phenotype. In other words, the aim is to determine whether the specific gene set is the best among all possible sets. It is a too stringent requirement. A more reasonable alternative hypothesis may suppose that the specific gene set is belonging to the top group, which includes the gene sets with the highest association with the phenotype. However, since there is an enormous amount of possible gene sets (2k-1), the hypothesis testing is a difficult task. This study only focuses on the self-contained test.

Our method makes use of the Random Forest ensemble learning. As this article focuses on a categorical phenotype, this method is

also applicable to a continuous-type phenotype. When the phenotype is continuous, the test statistic becomes the OOB mean of squared residuals (MSE), given as $\text{MSE} = \frac{1}{n}\sum_1^n \left(y_i - \hat{y}_i^{\text{OOB}}\right)^2$, where $\hat{y}_i^{\text{OOB}}$ is the average of the OOB predictions. See Breiman (2001) and Liaw and Wiener (2002). Breiman (2001) showed that Random Forests has a convergent error rate as the number of trees in the forest becomes large and hence does not have a problem of overfitting. The tree sizes considered here are 5,000 and 50,000 respectively for the simulation and the real examples. Liaw and Wiener (2002) suggest determining an appropriate tree size by comparing predictions of a forest to predictions of a subset of the forest. To determine the adequacy of the tree sizes, we have conducted another experiment on several gene sets of real examples. The observed OOB test set error rates are calculated for various tree sizes. It is found that the required tree size for a stable error rate depends more on the strength of association between the gene set and the phenotype, rather than the size of the gene set. Using the tree sizes that we have selected, most gene sets have a convergent error rate, but the error rates of some gene sets do not converge to a single value, and instead tend to go up or down when the tree size is relatively large. We find that the magnitudes of vibrations are limited, because the vibrations come from varied predictions of one or two subjects. This situation occurs when there is (are) outlier(s) in the data set. With the limited effect, we conclude that the tree sizes used are adequate.

5. Conclusions

In summary, our proposed method allows one to simultaneously assess the discriminatory ability of gene sets and the importance of genes for interpretation of data in complex biological systems. The classifications of biologically defined gene sets can reveal the underlying interactions of gene sets associated with the phenotypes, and provide an insightful complement to conventional gene set analyses. In addition, when using the measures of variable importance in the RF algorithm, we explore its ability to assess the impacts and correlations of genes within each gene set. Annotation-based classifications are reported and visualized together with the underlying gene sets and their contribution to the phenotypes of interest, which thus far cannot be found using conventional GSA methods.

Conflict of interest statement

We declare that we have no conflict of interest.

Authors' contributions

H.M.H. and C.A.T. initiated this research and outlined the general idea. C.A.T. and D.W.Z. wrote the programs and performed the analysis. H.M.H. and CAT wrote the article, read and approved the final version of the article.

Acknowledgments

This work was supported partially by the National Science Council of Taiwan, R.O.C. under the grants (NSC 100-2118-M-002-007 and NSC 99-2118-M-004-005).

References

Boulesteix, A.-L., Porzelius, C., Daumer, M., 2008. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. Bioinformatics 24, 1698–1706.
Breiman, L., 2001. Random forests. J. Mach. Learn. 45, 5–32.
Chen, J.J., Lee, T., Delongchamp, R., Chen, T., Tsai, C.A., 2007. Significance analysis of groups of genes in expression profiling studies. Bioinformatics 23, 2104–2112.
Delongchamp, R., Lee, T., Velasco, C., 2006. A method for computing the overall statistical significance of a treatment effect among a group of genes. BMC Bioinformatics 7 (Suppl. 2), S11.
Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3.
Dinu, I., et al., 2007. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics 8, 242.
Efron, B., Tibshirani, R., 2007. On testing the significance of sets of genes. Ann. Appl. Biol. 1, 107–129.
Farmer, P., et al., 2005. Identification of molecular apocrine breast tumours by microarray analysis. Oncogene 24, 4660–4671.
Fridley, B.L., Jenkins, G.D., Biernacka, J.M., 2010. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. PLoS One 5, e12693.
Goeman, J., Buhlmann, P., 2007. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23, 980–987.
Goeman, J., van de Geer, S., de Kort, F., van Houwelingen, H., 2004. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20, 93–99.
Huang, X., et al., 2005. A comparative study of discriminating human heart failure etiology using gene expression profiles. BMC Bioinformatics 6, 205.
Kong, S., Pu, W., Park, P., 2006. A multivariate approach for integrating genome wide expression data and biological knowledge. Bioinformatics 22, 2373–2380.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.
Lin, S., Devakumar, J., Kibbe, W., 2006. Improved prediction of treatment response using microarrays and existing biological knowledge. Pharmacogenomics 7, 495–501.
Liu, Q., Dinu, I., Adewale, A.J., Potter, J.D., Yasui, Y., 2007. Comparative evaluation of gene-set analysis methods. BMC Bioinformatics 8, 431.
Lottaz, C., Spang, R., 2005. Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. Bioinformatics 21, 1971–1978.
Mansmann, U., Meister, R., 2005. Testing differential gene expression in functional groups: Goeman's global test versus an ANCOVA approach. Methods Inf. Med. 44, 449–453.
Mootha, V., et al., 2003. PGC-1 α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 34, 267–273.
Nam, D., Kim, S., 2008. Gene-set approach for expression pattern analysis. Brief. Bioinform. 9, 189–197.
Newton, M., Fernando, A., Johan, A., Srikumar, S., Paul, A., 2007. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann. Appl. Stat. 1, 85–106.
Pang, H., Zhao, H., 2008. Building pathway clusters from Random Forests classification using class votes. BMC Bioinformatics 9, 87.
Pang, H., et al., 2006. Pathway analysis using random forests classification and regression. Bioinformatics 22, 2028–2036.
Riley, T., Sontag, E., Chen, P., Levine, A., 2008. Transcriptional control of human p53-regulated genes. Nat. Rev. Mol. Cell Biol. 9, 402–412.
Sartor, M., Leikauf, G., Medvedovic, M., 2009. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics 25, 211–217.
Statnikov, A., Wang, L., Fliferis, C., 2008. A comprehensive comparisons of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 3.
Subramanian, A., et al., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.
Tai, F., Pan, W., 2007a. Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. Bioinformatics 23, 1775–1782.
Tai, F., Pan, W., 2007b. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. Bioinformatics 23, 3170–3177.
Tian, L., Greenberg, S., Kong, S., Altschuler, J., Kohane, I., Park, P., 2005. Discovering statistically significant pathways in expression profiling studies. Proc. Natl. Acad. Sci. U. S. A. 102, 13544–13549.
Tsai, C.A., Chen, J., 2009. Multivariate analysis of variance test for gene set analysis. Bioinformatics 25, 897–903.
Wei, Z., Li, H., 2007. Nonparametric pathway-based regression models for analysis of genomic data. Biostatistics 8, 265–284.