Gene 655 (2018) 71–83

Contents lists available at ScienceDirect

Gene

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

Research paper

Statistical approach for selection of biologically informative genes

Samarendra Das[a,c], Anil Rai[b], D.C. Mishra[b], Shesh N. Rai[c,*]

[a] Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012, India
[b] Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012, India
[c] Biostatistics Shared Facility, JG Brown Cancer Center and Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, 40202, KY, USA

ARTICLE INFO

Keywords:
Informative genes
Bootstrap
Boot-MRMR
Gene Set Enrichment with QTLs
Gene sampling
Subject sampling

ABSTRACT

Selection of informative genes from high dimensional gene expression data has emerged as an important research area in genomics. Many gene selection techniques have been proposed so far are either based on relevancy or redundancy measure. Further, the performance of these techniques has been adjudged through post selection classification accuracy computed through a classifier using the selected genes. This performance metric may be statistically sound but may not be biologically relevant. A statistical approach, *i.e.* Boot-MRMR, was proposed based on a composite measure of maximum relevance and minimum redundancy, which is both statistically sound and biologically relevant for informative gene selection. For comparative evaluation of the proposed approach, we developed two biological sufficient criteria, *i.e.* Gene Set Enrichment with QTL (GSEQ) and biological similarity score based on Gene Ontology (GO). Further, a systematic and rigorous evaluation of the proposed technique with 12 existing gene selection techniques was carried out using five gene expression datasets. This evaluation was based on a broad spectrum of statistically sound (*e.g.* subject classification) and biological relevant (based on QTL and GO) criteria under a multiple criteria decision-making framework. The performance analysis showed that the proposed technique selects informative genes which are more biologically relevant. The proposed technique is also found to be quite competitive with the existing techniques with respect to subject classification and computational time. Our results also showed that under the multiple criteria decision-making setup, the proposed technique is best for informative gene selection over the available alternatives. Based on the proposed approach, an R Package, *i.e.* BootMRMR has been developed and available at https://cran.r-project.org/web/packages/BootMRMR. This study will provide a practical guide to select statistical techniques for selecting informative genes from high dimensional expression data for breeding and system biology studies.

1. Introduction

Genome wide expression studies are powerful genomic approaches, which have ability to capture expression dynamics of several thousand (s) of genes in a cell (Lai et al., 2006; Trevino et al., 2007). Among these thousands of expressed genes, all may not be required for classification, gene regulation modeling, modules detection, *etc.* (Guyon et al., 2002; Lai et al., 2006; Díaz-Uriarte and de Andrés, 2006; Wang et al., 2013; Das et al., 2017a). There is a need to select few genes or set of genes which are highly relevant for particular condition/trait, *i.e.* informative genes (Golub et al., 1999; Wang et al., 2013). These informative genes are used as predictors for diagnosing a disease (Golub et al., 1999; Guyon et al., 2002; Lai et al., 2006; Trevino et al., 2007) or to understand the stress response mechanism in plants (Wang et al., 2013; Das et al., 2017a). Further, in order to develop statistical models for Gene Expression (GE) data having large number of genes as predictors as compared to small number samples/subjects, leads to large p small n class of problems and consequently raises several statistical issues *like* stability, power and feasibility of the model (Kursa, 2014). Therefore, it is inevitable to reduce the dimensionality of GE data, which is often

Abbreviations: GE, Gene Expression; MRMR, Maximum Relevance and Minimum Redundancy; CA, Classification Accuracy; Boot-MRMR, Bootstrap-MRMR; MCDM, Multiple Criteria Decision Making; SVM-RFE, Support Vector Machine- Recursive Feature Elimination; RF, Random Forest; FC, Fold Change; Wilcox, Wilcoxon's statistic; IG, Information Gain; GR, Gain Ratio; SU, Symmetric Uncertainty; PCF, Pearson's Correlation Filter; SRC, Spearman's Rank Correlation; Al, Aluminum; QTL, Quantitative Trait Loci; GO, Gene Ontology; GSEQ, Gene Set Enrichment with QTLs; NP, Non-Parametric; FCD, F-test with Correlation Difference; FCQ, F-test with Correlation Quotient; rv, random variable; iid, independently and identically distributed; MF, Molecular Function; BP, Biological Process; CC, Cellular Component; SVM-LBF, SVM classifier with Linear Basis Function; SVM-RBF, SVM classifier with Radial Basis Function; SVM-PBF, SVM classifier with Polynomial Basis Function; SE, Standard Error; FDR, False Discovery Rate; FPR, False Positive Rate; FNR, False Negative Rate; ACC, Accuracy; MCC, Mathew's Correlation Co-efficient; TOPSIS, Technique for Order Performance by Similarity to Ideal Solution

* Corresponding author.
E-mail addresses: samarendra.das@louisville.edu (S. Das), anilrai@iasri.res.in (A. Rai), dcmishra@iasri.res.in (D.C. Mishra), shesh.rai@louisville.edu (S.N. Rai).

https://doi.org/10.1016/j.gene.2018.02.044
Received 24 May 2017; Received in revised form 26 November 2017; Accepted 14 February 2018
Available online 16 February 2018
0378-1119/ Published by Elsevier B.V.

achieved by informative gene selection.

In a gene selection technique, it is desirable to have two important features, *i.e.* minimum redundancy among the selected genes and maximum relevance of these genes with the experimental condition/trait (Ding and Peng, 2005; Peng et al., 2005; Mundra and Rajapakse, 2010). Several gene selection techniques have been proposed to select only pertinent genes from thousand(s) of genes with the help of limited available experimental samples based on either relevance or redundancy measure (Saeys et al., 2007). In this regard, volcano plot method is quite popular among biologists (Cui and Churchill, 2003) where, genes are selected by considering their relevance within a given level of experimental conditions under which data is being generated. But volcano plot is a graphical method and is not sufficient to discover some complex relationships among genes for a certain condition/trait (Liang et al., 2011). Besides, several statistical and machine learning algorithms have also been proposed in literature (Inza et al., 2004; Saeys et al., 2007). Further, these methods select genes by only considering their relevance within a level of conditions of the class/trait. However, in these computational techniques, there is a possibility of selection of spuriously associated genes as they failed to consider redundancy measure. Then, Maximum Relevance and Minimum Redundancy (MRMR) technique has been developed to select cancer responsible genes by considering both relevancy and redundancy measures (Ding and Peng, 2005; Peng et al., 2005). Here, both the measures are computed using mutual information by discretizing the continuous GE data. In this case, also there is a chance of losing information and selection of spuriously associated genes (Mundra and Rajapakse, 2010).

It has been observed that most of the available gene selection techniques were used to select cancer responsible genes from human GE data and subsequently used for patient classification (*e.g.* with and without cancer) (Golub et al., 1999; Guyon et al., 2002; Lai et al., 2006; Díaz-Uriarte and de Andrés, 2006). Therefore, it is important and highly pertinent to systematically explore these techniques in the context of plant genomics. Usually, Classification Accuracy (CA)/error rate computed at the post gene selection phase has been used as major criterion to evaluate the performance of gene selection technique(s) (Golub et al., 1999; Guyon et al., 2002; Ding and Peng, 2005; Peng et al., 2005; Lai et al., 2006; Díaz-Uriarte and de Andrés, 2006; Mundra and Rajapakse, 2010; Kursa, 2014). It may be noted that this traditional criterion may be statistically sound but may not be biologically relevant for performance evaluation. For instance, a gene selection technique may lead to identification of a set of genes which predicts the classes of subjects more accurately, but these selected genes may or may not be biologically relevant for that particular condition/trait. Hence, it is important to assess the performance of a gene selection technique based on both statistically sound and biologically relevant criteria.

Therefore, in this study a statistical approach, *i.e.* Bootstrap-MRMR (Boot-MRMR) is developed for selection of biologically relevant informative genes from high dimensional GE data. This proposed approach is based on a composite measure considering both gene relevancy and redundancy, where informative genes are selected after minimising the effects of spurious associations among genes under a sound statistical framework. Further, the proposed approach of gene selection is found to be competitive and even better than the existing techniques for subject classification while its performance was evaluated on five different crop GE datasets. Besides this, two biologically relevant criteria are also developed based on Quantitative Trait Loci (QTL) and Gene Ontology (GO) information for comparative performance analysis of the proposed Boot-MRMR approach and it was tested on four rice GE datasets. It was found that the genes selected through the proposed approach are more biologically relevant when compared to existing techniques. Also, a systematic and rigorous comparative evaluation of existing 12 gene selection techniques with the proposed Boot-MRMR approach has been carried out on four GE datasets related to various stresses in rice under a Multiple Criteria Decision Making (MCDM) set up. These existing techniques are Support Vector Machine-Recursive Feature Elimination (SVM-RFE) (Guyon et al., 2002; Liang et al., 2011; Wang et al., 2013), Random Forest (RF) (Díaz-Uriarte and de Andrés, 2006; Diaz-Uriarte, 2007; Kursa, 2014), Fold Change (FC) (Das et al., 2017b), t-score (Cui and Churchill, 2003; Das et al., 2017b), F-score (Lazar et al., 2012), Wilcoxon's statistic (Wilcox) (Hossain et al., 2013), MRMR (Ding and Peng, 2005; Peng et al., 2005), information theoretic measures, *i.e.* Information Gain (IG), Gain Ratio (GR), Symmetric Uncertainty (SU) (Forman, 2003; Liu et al., 2005; Mao et al., 2006; Cheng et al., 2012), Pearson's Correlation Filter (PCF) (Golub et al., 1999) and Spearman's Rank Correlation (SRC) (Saeys et al., 2007; Cheng et al., 2012). The results showed that the performance of Boot-MRMR approach is better for most of the cases under this MCDM environment.

2. Materials and methods

2.1. Data collection

The GE experimental datasets of rice and soybean were collected from Gene Expression Omnibus database of NCBI for platforms GPL2025 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2025) and GPL4592 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL4592) respectively. These GE datasets were collected under biotic (*Xanthomonas* bacteria) and abiotic stresses (salinity, cold and drought) for rice and Aluminum (Al) stress for soybean. The summary and details of these datasets are given in Table 1 and Supplementary Table S1 respectively. Further, a detail description about the collection, pre-processing and meta-analysis of these datasets is given in Supplementary Document S1. The QTL datasets for these stresses *viz.* salinity, drought, cold and bacteria for rice were collected from the Gramene QTL database (http://www.gramene.org/qtl/) (Ni et al., 2009) and are given in Supplementary Document S2.

2.2. Boot-MRMR approach

Here, we proposed an improvised gene selection approach, *i.e.* Boot-MRMR for selection of informative genes from high dimensional GE data. The proposed approach can minimize the spurious associations among the genes while informative gene selection. Further, this approach is based on a Non-Parametric (NP) test statistic for informative gene selection. In usual MRMR technique, genes are ranked by optimizing the combination of relevance and redundancy measures under two schemes *i.e.* F-test with Correlation Difference (FCD) and F-test with Correlation Quotient (FCQ) (Ding and Peng, 2005; Peng et al., 2005). However, in case of Boot-MRMR approach, we used the FCQ scheme, as it outperformed the FCD scheme on continuous GE data (Ding and Peng, 2005).

Let, the MRMR objective function (J) for the gene selection problem is given as:

$$J = \max(V/W) \tag{1}$$

where, V and W indicates the relevance and redundancy measures respectively. Further, the near optimal solution of J for the continuous GE data was obtained by Ding and Peng (2005) and given as:

$$w_i = \max_{i\in\Omega}\left\{F(i,y)/\left(\frac{1}{|\Omega|}\sum_{i\neq j=1}^{|\Omega|} |\,C(i,j)\,|\right)\right\} \tag{2}$$

where, $\boldsymbol{\Omega}$ is the gene space (*e.g.* number of probes in a microarray chip), w_i is the weight associated with *i-th* gene, y is the class label of a subject, $F(i, y)$ is the F-score between the *i-th* gene in y class and $C(i, j)$ is correlation measure between *i-th* and *j-th* genes in GE dataset. In this technique, w_i was used as criterion for gene ranking (Ding and Peng, 2005; Peng et al., 2005). However, many methods of gene selection are susceptible to small permutation of experimental conditions (Guoyon

Table 1
Gene expression studies used in the study.

Sl. no.	Descriptions	GSE accessions	#Genes	#Samples	#Class
D1	Salinity stress in rice	GSE13735, GSE14403, GSE21651, GSE28209, GSE16108, GSE6901	6637	70	2
D2	Cold stress in rice	GSE38023, GSE31077, GSE33204 GSE37940, GSE6901	8840	100	2
D3	Drought stress in rice	GSE6901, GSE26280, GSE21651, GSE23211, GSE24048, GSE25176	9078	90	2
D4	Aluminum stress in soybean	GSE18423, GSE18517, GSE18518	8416	76	2
D5	Bacterial stress in rice	GSE16793, GSE19239, GSE19844 GSE32426, GSE33411, GSE36093 GSE36272	8356	221	2

GSE Accessions: Accession numbers of Gene Expression studies; #Genes: Number of genes; #Samples: Number of GEO samples; #Class: Number of classes (*e.g.* 2 in control *vs.* stress genomic study); D1–D4: GEO datasets from abiotic stresses; D5: GEO datasets from biotic stress.

and Elisseeff, 2003; Das et al., 2017a). Further, the ranking of genes was done using a single high dimensional GE data, which leads to the selection of spuriously associated genes.

Therefore, in the proposed Boot-MRMR technique, a modified bootstrap based subject sampling model was used. In this model, the GE experimental samples, *i.e.* subjects are taken as sampling units and these subjects are randomly taken from the population. Each subject has GE measurements for same set of genes of $\boldsymbol{\Omega}$. Moreover, the replicated GE samples were taken as new sampling units under this sampling model, which may have different GE profiles as compared to other replicates (Goeman and Buhlmann, 2007). Let, M denotes population size, *i.e.* total number of GE profiles for different subjects in the experiment and each subject is treated as an independent unit in the population. The relation of each subject with its class can be shown as:

$$(X_1, y_1),\ (X_2, y_2),\ \ldots, (X_s, y_s),\ \ldots, (X_M, y_M) \tag{3}$$

where, X_s represents the N-dimensional vector (N is total number of genes in $\boldsymbol{\Omega}$) of GE levels for *s-th* subject and y_s is the corresponding class label (*e.g.* stress: $+1$ *vs.* control: -1), $s = 1, 2, \ldots, M$. Therefore, M expression levels of different subjects are independently and identically distributed (iid), but expression levels of genes within the same subject may be correlated for a given condition. Let, m units of realizations were randomly selected out of M population units (as represented by Eq. (3)) ($m \leq M$) with replacement to construct one bootstrap sample. Then, the standard MRMR algorithm was applied on this bootstrap sample to get one list of the genes along with their ranks (say one gene list). This procedure was repeated S times to get S gene lists. Here, S, *i.e.* number of bootstrap samples must be sufficiently large (Wang et al., 2013). It has been empirically established that value of S should be around 200 to ensure all desirable features of bootstrapping (Efron and Tibshirani, 1993). Accordingly, the value of S was set as 200 in this study.

Further, genes in every gene list will have ranks between 1 to N. Then a function, *i.e. Rank Score* for *k-th* gene list, *i.e.* R_k ($R_k : \Omega \rightarrow [0, 1]$) was defined to map ranks of genes in Ω to corresponding scores. The *Rank Score* ($R_k^{(i)}$) for *i-th* gene in *k-th* ($k = 1, 2, \ldots, S$) gene list can be defined as:

$$R_k{}^{(i)} = f(p_{ik}) = \frac{N - p_{ik} + 1}{N} \tag{4}$$

where, p_{ik} ($1 \leq p_{ik} \leq N$) is the rank of *i-th* gene in *k-th* gene list. It can be noted that, for *i-th* gene, $R_k^{(i)}$ ($N^{-1} \approx 0 \leq R_k^{(i)} \leq 1$) is a random variable (rv). So, without loss of generality, another rv $r_k{}^{(i)}$ can be defined as:

$$r_k^{(i)} = R_k^{(i)} - Q_2 \tag{5}$$

where, Q_2 is second quartile (median) of rank scores, *i.e.* $R_k^{(i)}$ (here any quartile can be taken). It may be noted that the rank scores of genes in each gene list will be symmetrically distributed around Q_2 (*i.e.* 0.5) (as rank scores are function of gene ranks). Further, to select informative genes, the following hypothesis needs to be tested for each gene in Ω successively.

H_0. *i-th* gene is not informative for a given condition/trait, *i.e.* $W_i \leq 0$.

H_1. *i-th* gene is informative for a given condition/trait, *i.e.* $W_i > 0$.

where, W_i be the median deviated expected rank score for *i-th* gene over all possible bootstrap samples. The bootstrap procedure coupled in the subject sampling model was used to ensure the iid assumptions of the rank scores. The test statistics for testing the above hypothesis can be obtained as:

Let for gene i, the r_k's (from Eq. (5)) are arranged in ascending order of their magnitude. Subsequently, the ranks 1, 2, …, S are assigned, keeping in mind their original signs. Let, A^+ be sum of the ranks of positive r_k's and A^- be the sum of the ranks of negative r_k's. Thus, for the computation of distribution of A^+, another rv $B_{(l)}$ is defined as:

$$B_{(l)} = \begin{cases} 1 & \text{if the } |r_k| \text{ has rank } l\ (>0) \\ 0 & \text{else} \end{cases} \tag{6}$$

Now, the variables $B_{(l)}$ are independent Bernouli variates and its mean and variance can be obtained as:

$$E(B_{(l)}) = \frac{S}{2}\binom{S-1}{l-1} B(l, S-l+1) \tag{7}$$

$$\text{Var}\,(B_{(l)}) = E\{B_{(l)}(1 - E(B_{(l)}))\} \tag{8}$$

Further, the mean and variance of the test statistic (A^+) can be obtained as:

$$E(A^+) = \sum_{l=1}^{S} lE(B_{(l)}) \tag{9}$$

$$\text{Var}(A^+) = \frac{1}{4}\sum_{l=1}^{S} l^2\{E(B_{(l)})(1 - E(B_{(l)}))\} \tag{10}$$

Under the simple null hypothesis $H_0 : W_i = 0$, the Eqs. (9) and (10) can be expressed as:

$$E_{H_0}(A^+) = \frac{1}{2}\sum_{l=1}^{S} l = \frac{S(S+1)}{4} \tag{11}$$

$$\text{Var}_{H_0}(A^+) = \frac{1}{4}\sum_{l=1}^{S} l^2 = \frac{S(S+1)(2S+1)}{24} \tag{12}$$

As the number of bootstrap samples are quite large ($S = 200$), then under Lindeberg's central limit theorem (Ash, 2000; Rohatgi and Ehsanes Saleh, 2011), the test statistic (A^+) follows normal distribution asymptotically, *i.e.*

$$\frac{A^+ - E_{H_0}(A^+)}{\sqrt{\text{Var}_{H_0}(A^+)}} \sim N(0,\ 1) \tag{13}$$

Based on the above test statistic, the statistical significance value for *i-th* gene (p_i-*value*) was computed. Further, to control the type I error in cases of multiple tests of genes, we applied the False Discovery Rate (FDR) method (Benjamini and Hochberg, 1999) to compute adjusted *p*-

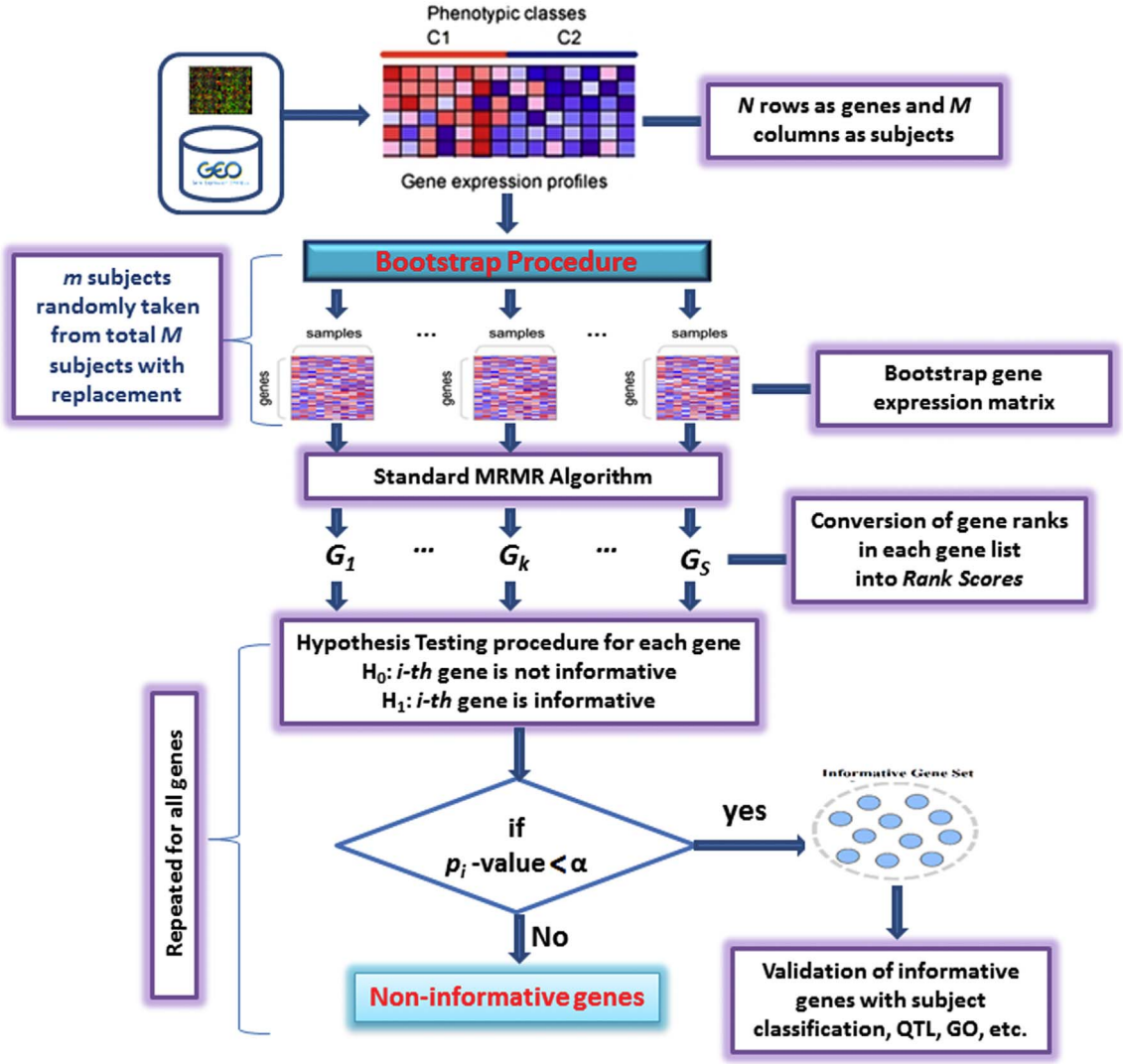


Fig. 1. Flowchart depicting the implemented algorithm of Boot-MRMR approach. G_k's are N-dimensional vectors of gene lists having gene rank scores. MRMR is Maximum Relevance and Minimum Redundancy algorithm. p_i-*value* is statistical significance value for *i-th* gene. α is the desired level of statistical significance.

values for selection of informative genes. The implemented algorithm of the Boot-MRMR approach is given in Fig. 1. Moreover, the comparative performance analysis of the proposed Boot-MRMR approach with respect to 12 existing gene selection techniques (Supplementary Table S2) was carried out on five different real crop GE datasets (Table 1).

2.3. Biologically relevant criteria for performance analysis of gene selection techniques

In this section, two biologically relevant criteria were reported for performance evaluation of gene selection techniques. The first criterion is based on the extent of statistical association of selected informative gene set with related QTLs, *i.e.* Gene Set Enrichment with QTLs, of the target trait. The second criterion is about the development of indices based on associated GO terms with selected informative gene set.

2.3.1. Gene Set Enrichment with QTLs (GSEQ)

QTLs are segments of DNA or genomic region either containing or linked to genes for a given quantitative trait (Tiwari et al., 2016) and are widely used in plant breeding experiments (Du et al., 2016). In order to assess the performance of gene selection techniques, the trait specific QTLs can be used as biologically relevant criteria for cross validating the selected gene sets (obtained through gene selection methods).

Let G be the selected gene set obtained from Ω using any gene selection technique (Supplementary Table S2) for a given condition and Q be the set of associated QTLs for that condition/trait. Let, *i-th* gene in G is represented as $g_i^c\,[a, b] \in G$, where a and b represent start and stop positions (in terms of base pairs) of the gene g_i on chromosome c. For a QTL, $q_t^c[d, e] \in Q$, where d and e represents the start and stop positions of the QTL q_t on chromosome c. Let, the QTL hit by the gene g_i^c can be expressed by using an indicator function, as:

$$I_{q_t}(g_i) = \begin{cases} 1 & \text{if } g_i^c[a,] \geq q_t^c[d,] \text{ and } g_i^c[,b\,] \leq q_t^c[,e\,] \\ 0 & else \end{cases} \tag{14}$$

In other words, if a selected gene is completely overlapped with the QTL regions for a given trait on the same chromosome, then it can be said that the gene got a QTL hit. Therefore, a statistic (*NQTL*), *i.e.* total number of QTL hits by the informative genes in G can be used to compute total number of genes in G overlapped with the QTLs in Q and can be expressed as:

$$NQTL = \sum_{t=1}^{|Q|} \sum_{i=1}^{|G|} I_{q_t}(g_i) \tag{15}$$

Besides this, a test statistic is also used for GSEQ test to assess the performance of a gene selection technique. If a gene set G is enriched

Table 2
2 × 2 contingency table for gene set enrichment with QTL test.

	Overlapped with QTL regions	Not overlapped with QTL regions	Total
Selected gene set	n_{GQ}	n_{GQ^c}	n_G
Not in gene set	n_{G^cQ}	$n_{G^cQ^c}$	n_{G^c}
Total	n_Q	n_{Q^c}	N

G: selected gene set; G^c: not selected gene set; n_G: number of genes in the gene set.

with the underlying QTL information, then member genes in G should have higher proportion of QTL hits as compared to that of genes in G' (*i.e.* $\Omega - G$). For testing the trait specific QTLs enrichment of gene set G, let us define following null hypothesis:

H_0. QTLs enrichment of G and G' are same.

H_1. G is more QTLs enriched as compared to G'.

The constructed null hypothesis is a competitive one, not self-contained, as it considers the genes from both the gene sets (G and G') (Goeman and Buhlmann, 2007). For this GSEQ test, we used 2 × 2 contingency table and gene sampling procedure. The 2 × 2 contingency table method was extensively used in gene set enrichment test with known pathways or GO based information (Goeman et al., 2004; Barry et al., 2005; Al-Shahrour et al., 2005). In this gene sampling procedure, n genes were randomly selected from the informative gene set (G) ($n \leq |G|$) without replacement to construct one random gene sample. This procedure was repeated p times to get p random gene samples. For each gene sample, a 2 × 2 contingency table, as shown in Table 2, was constructed. Therefore, total p 2 × 2 contingency tables were obtained. Here, the value of p was taken as 100.

To test the null hypothesis for its possible rejection, we proposed the following procedure. Let N is total number of genes in Ω, N_Q is total number of QTL hits in Ω and n is the size of each random gene sample. The underlying hypergeometric distribution for each 2 × 2 contingency table can be given as:

$$P[X = x] = \frac{\binom{N_Q}{x}\binom{N - N_Q}{n - x}}{\binom{N}{n}} \tag{16}$$

where, X is random variable representing value of QTL hits in *i-th* gene sample ($i = 1, 2, \ldots, p$). The corresponding statistical significance value for *i-th* gene sample (P_i) can be computed by:

$$P_i = P[X_i \geq x \mid H_0] = 1 - P[X_i \leq x \mid H_0] \tag{17}$$

Similarly, statistical significance values were calculated for each p random gene samples using the above procedure. It may be noted that these P_i's ($i = 1, 2, \ldots, p$) are iid's and each one follows a uniform distribution with parameter [0, 1] (Bland, 2013) (*i.e.* $P_i \sim$ U [0, 1]). Let's define (without loss of generality) a rv Z_i, *i.e.* $Z_i = -2 \log P_i$ which follows a chi-squared distribution with 2 df, *i.e.* $Z_i \sim \chi_2$. Here, Z_i ($i = 1, 2, \ldots, p$) are also iid's with chi-squared distribution. Therefore, we can test H0 against H1 through the following test statistic:

$$L = 2\log\left(1/\prod_{i=1}^{p} P_i\right) = \sum_i Z_i \sim \chi^2_{2p} \tag{18}$$

The *p-value* obtained from the above test statistic is an indicator of the extent of QTL enrichment for G. The lesser *p-value* indicates better informative gene selection technique through which G is selected and *vice-versa*.

It may be noted that, in single gene analysis, *e.g.* Boot-MRMR, each gene is tested for its involvement in the trait under consideration. Therefore, there is a problem of multiple hypotheses testing, which was addressed by computing FDR adjusted *p-values*. But in case of GSEQ, gene sets (polygenes) are taken as input for testing their involvement in QTL enrichment. Hence, the multiple testing of hypothesis problem is well tackled in GSEQ, as it takes gene set as a functional unit for enrichment analysis.

2.3.2. Biological similarity score based on GO terms

Under the GO term enrichment analysis, the functions of the genes are annotated under three taxonomies, *i.e.* Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) (GO Consortium, 2015). This analysis helps in evaluating the functional similarities among the genes in G (Mazandu and Mulder, 2014), as there exists a direct relationship between semantic similarity of gene pairs with their structural (sequence) similarity (Lord et al., 2003; Wang et al., 2007). To assess the performance of gene selection techniques with respect to biological relevancy of selected genes, we used the GO based semantic similarity measure developed by Wang et al. (2007) to compute the biological similarity score for the selected gene set (G) thorough a particular technique.

Let, $GO_i = \{go_{i1}, go_{i2}, \ldots, go_{iK}\}$ and $GO_j = \{go_{j1}, go_{j2}, \ldots, go_{jL}\}$ be the two sets of GO terms that annotate two genes g_i and g_j in G respectively. Then the functional semantic similarity (ρ_{ij}) between g_i and g_j can be expressed as:

$$\rho_{ij} = \frac{|GO_i \cap GO_j|}{|GO_i \cup GO_j|} \quad \forall\ i \neq j = 1, 2, \ldots, |G| \tag{19}$$

Further, the biological similarity score for G based on the above semantic similarity measure (Eq. (19)) can be expressed as:

$$\rho(G) = \frac{2}{|G|(|G| - 1)} \sum_{i \neq j=1}^{|G|} \rho_{ij} \tag{20}$$

where, $\rho(G)$ is the proposed biological similarity score for G. Using the Eq. (20), the biological similarity scores under MF, BP and CC taxonomies were computed for each of the selected informative gene sets. The higher value of $\rho(G)$ indicates more informative-ness of the selected gene set and *vice-versa*. This is due to the fact that for a given trait the genes present in G are likely to share common GO terms.

2.4. Performance analysis of gene selection techniques based on classification

The performance of the proposed and existing gene selection techniques (Supplementary Table S2) was evaluated based on subject classification *i.e.* CA and Standard Error (SE) in *CA*. These performance metrics were computed through a sliding varying window size technique (Das et al., 2017a). Here, the window size refers to number of ranked genes obtained by a gene selection technique. The window sizes were taken as 50, 100, 150, …, 950, 1000 with a sliding length of 50. Moreover, the number of top ranked genes (equal to the window sizes) selected through the proposed and other existing techniques were then used in SVM classifier to discriminate the class labels of samples between stress (+1) and control (−1). In the SVM classifier, three basis functions, *i.e.* linear (SVM-LBF), radial (SVM-RBF) and polynomial (SVM-PBF) were used to compute the CA and SE in *CA*. Further, the techniques which provide maximum discrimination between the two groups through classification will be the better technique for informative gene selection and *vice-versa*. The performance of these techniques was adjudged on the basis of CA and SE in *CA*.

The other criteria *viz.* sensitivity, specificity, False Discovery Rate (FDR), False Positive Rate (FPR), False Negative Rate (FNR), Accuracy (ACC), F1-Score and Mathew's Correlation Co-efficient (MCC) were also used in this performance evaluation. The expressions for these criteria are given in Supplementary Table S3. Usually, the genes were selected from Ω based on the training GE data and these selected genes were used as predictors in classifiers to test their ability to differentiate the class labels of the samples. Then, the performance of that technique was evaluated on test samples (Guyon et al., 2002; Ding and Peng, 2005;

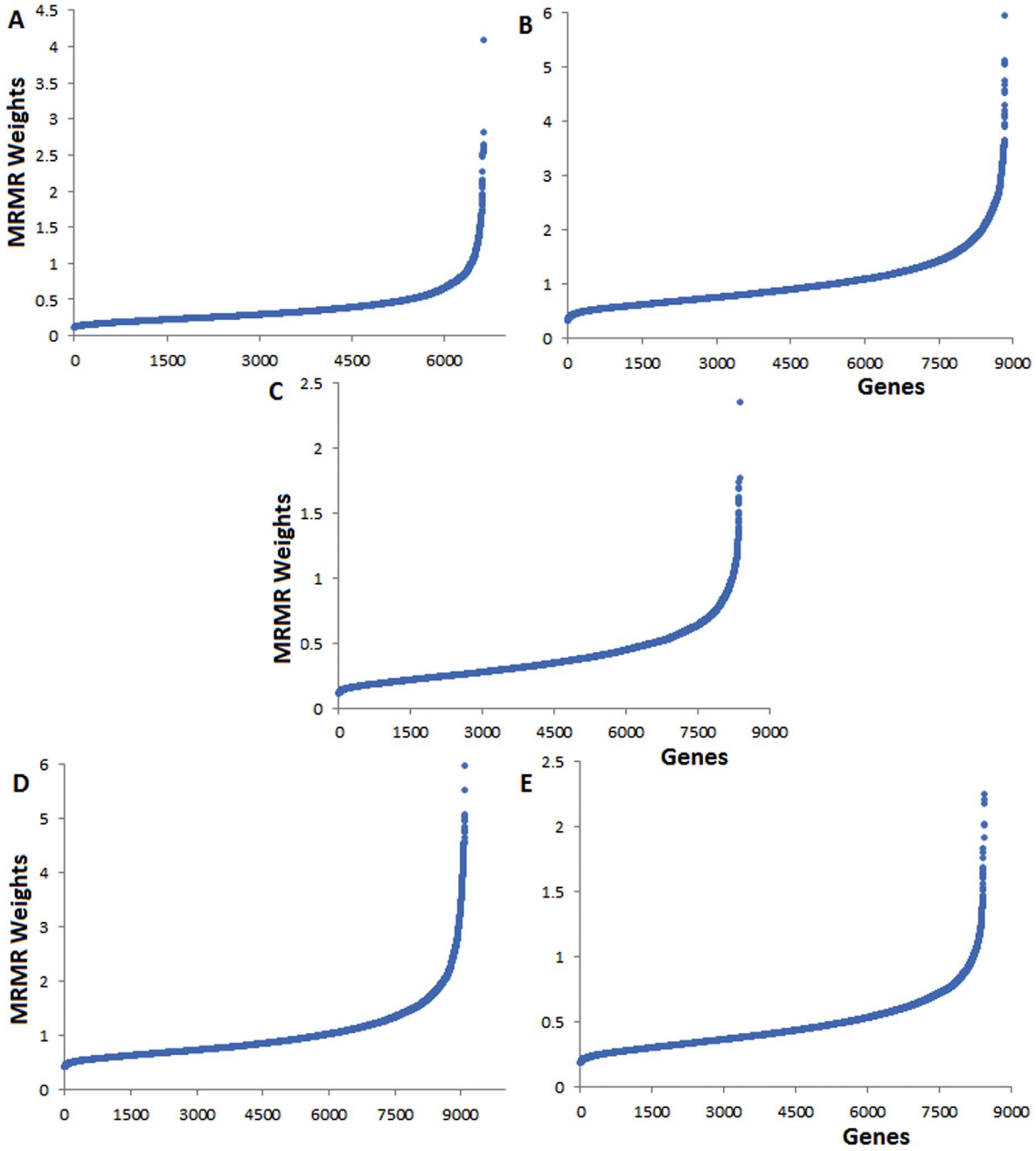


Fig. 2. Distribution of weights computed from MRMR algorithm. The horizontal axis represents the genes. The vertical axis shows weights computed through MRMR technique. Distribution of *MRMR weights* are shown for (A) salinity stress, (B) cold stress, (C) aluminum stress, (D) drought stress and (E) bacterial stress.

Peng et al., 2005; Lai et al., 2006; Díaz-Uriarte and de Andrés, 2006; Kursa, 2014). This procedure is sometimes unreliable because of the small number of test samples and imbalances (unequal representation of samples from both classes) of training and testing data (Wei and Dunbrack Jr, 2013).

Therefore, to avoid this limitation, we used the following procedure in classification based performance evaluation of the gene selection techniques, *i.e.* (i) the training and testing datasets were merged, (ii) the bootstrap samples were drawn from this merged data and, (iii) then bootstrap samples were then partitioned for training and testing sets. Then, the SVM-LBF, SVM-RBF and SVM-PBF classifiers were trained on the training dataset for a bootstrap sample. Further, a 2 × 2 confusion matrix was constructed from the SVM results for each testing dataset for that bootstrap sample. From that confusion matrix, the performance criteria, *viz.* sensitivity, specificity, FDR, FPR, FNR, ACC, F1 and MCC (Supplementary Table S3) were computed for each bootstrap sample. This procedure was repeated over 100 times for 100 bootstrap samples and then these performance criteria were averaged over 100 trials.

2.5. Performance analysis of gene selection techniques based on computational time

The computational time required to select the informative gene set of a fixed size from the high dimensional GE dataset was also used as a performance metric. Here, the gene set of size 1000 was considered. Through this, the method which requires less runtime is better and *vice-versa*. To measure this, we ran the code written in R (*v* 3.3.1) 100 times for each gene selection technique to select the informative gene set of size 1000 and measured the required average CPU time. All these analyses were performed on a 4 GB RAM computer with Windows 10

OS and CPU clock rate as 2.93 GHz.

2.6. Comprehensive performance analysis of gene selection techniques

Here, we emphasized to compare the performance of the proposed Boot-MRMR approach with respect to 12 existing techniques (Supplementary Table S2) based on 16 criteria (Supplementary Table S3) on four real crop datasets (Table 1). In operational research, such problems are called as MCDM problem (Hwang and Yoon, 1981), where the main goal is to consider a set of criteria and choose the best performing option over a list of options (Khezrian et al., 2014). Under this MCDM set up, Technique for Order Performance by Similarity to Ideal Solution (TOPSIS) (Ahn, 2011) has been extensively used in areas of research, *viz.* human resources management, product design, manufacturing, water management, quality control, location analysis, DNA extraction analysis and e-tendering (Kwong and Tam, 2002; Chen and Tzeng, 2004; Milani et al., 2005; Srdjevic et al., 2004; Yang and Chou, 2005; Wang et al., 2015). However, this is being used for the first time in gene selection problems. Here, the basic idea is to choose the best gene selection technique out of 13 gene selection techniques (Supplementary Table S2) based on the 16 decision criteria (C1–C16) (Supplementary Table S3). Further, for the criteria; C2, C5–C7, C12 and C16, lower value indicates better performance of gene selection techniques and *vice-versa*. For C1, C3, C4, C9–C11, C13–C15 criteria, higher value stands for better performance and *vice-versa*. The comparative performance analysis of these techniques under the MCDM setup was carried out using TOPSIS approach and the major steps for this process are given in Supplementary Document S3. Through this, the gene selection techniques with higher R_r (R_r: TOPSIS score and $0 \leq R_r \leq 1$) are preferred and considered as better technique over these multiple criteria and *vice-versa*.

3. Results

The distributions of weights and *p-values* for genes obtained from the existing MRMR algorithm and proposed Boot-MRMR approach are shown in Figs. 2 and 3 respectively. The distributions of MRMR weights of genes for salinity, cold, drought and bacterial stresses in rice and Al stress in soybean contains lower and upper values, which are not so widely separated (Fig. 2). On the contrary, from the distribution of *p-values* of the genes, it was found that these *p-values* are well separated and relatively small number of genes were found to be statistically significant (Fig. 3). The distinction between informative genes (*i.e.* genes with lower *p-values*) and non-informative genes can be better identified from Fig. 3 as compared to Fig. 2. This comparative graphical analysis showed the improvement of Boot-MRMR over MRMR algorithm (Figs. 2, 3). Hence, the informative genes selection based on *p-values*, computed through the NP test seems to be more statistically sound and meaningful as compared to gene ranking methods *like* MRMR. Further, the performance analysis of the proposed Boot-MRMR technique was done using rice GE data for salinity, cold, drought and bacterial stresses and soybean data for Al stress through subject/samples classification. However, for QTL and GO based approach, performance evaluation was carried out on rice data for the four stresses as rice genome is well annotated.

3.1. Performance analysis of Boot-MRMR technique based on classification

The informative gene sets of size 1500 were selected by using each of the gene selection techniques as given in Supplementary Table S2. These selected genes were then used as predictors in SVM-LBF, SVM-RBF and SVM-PBF classifiers. The post selection CA and SE in CA for different sliding window sizes over fivefold cross validation for different gene selection techniques with respect to five different stress scenarios are also shown in Fig. 4. For better performance analysis, mean CA and SE in CA were computed for each gene selection techniques and the results are given in Supplementary Tables S4–S6. For cold and drought stress using SVM-LBF classifier, it was observed that the CA of Boot-MRMR was higher than that of other gene selection techniques followed by SVM-RFE (Fig. 4, Supplementary Table S4). Further, for these stresses, *i.e.* cold and drought, the SE values in CA are lowest for Boot-MRMR as compared to other techniques followed by SVM-RFE, which indicates that the genes selected by this proposed technique is highly informative and robust (Supplementary Table S4). In case of salinity, bacterial and Al stresses, the CA of SVM-RFE was found to be higher than that of other gene selection techniques followed by Boot-MRMR (Fig. 4, Supplementary Table S4). For these stresses, the SE values of CA for SVM-RFE are lower followed by Boot-MRMR. Similar interpretations about the performance of these gene selection techniques can be made for SVM-PBF and SVM-RBF classifiers (Fig. 4, Supplementary Tables S5, S6). Further, results obtained through varying window size technique for the Boot-MRMR are given in Supplementary Table S7.

The other classification based performance metrics, *viz.* sensitivity, specificity, FDR, FPR, FNR, ACC, F1-measure and MCC for each gene selection techniques in five different stress scenarios of rice and soybean are given in Supplementary Tables S4–S6. For salinity, bacterial and Al stresses with SVM-LBF classifier, the sensitivity, specificity, ACC, F1 and MCC measures of SVM-RFE were found to be higher than that of all other gene selection techniques followed by Boot-MRMR (Supplementary Table S4). Further, for this classifier, the values of FDR, FPR and FNR for SVM-RFE were lowest among other techniques followed by Boot-MRMR (Supplementary Table S4). These results indicated that SVM-RFE performed better followed by Boot-MRMR for salinity, bacterial and Al stresses. But, for SVM-LBF classifier in cold and drought stresses, the values of Sensitivity, Specificity, ACC, F1 measure and MCC of Boot-MRMR were found to be higher than that of other techniques of gene selection (Supplementary Table S4). Moreover, the values of FDR, FPR and FNR were lowest for Boot-MRMR as compared to other existing techniques. The comparative analysis based on subject classification indicated that out of the five stresses, in two cases, the performance of Boot-MRMR was better followed by SVM-RFE (Supplementary Tables S4-S6). In other cases, SVM-RFE outperformed other techniques followed by Boot-MRMR for gene selection (Fig. 4, Supplementary Tables S4-S6).

3.2. Performance analysis based on QTL based criteria

We mapped the QTLs and informative genes obtained by using 13 gene selection techniques (for rice) in the whole genome using MSU rice genome browser (Ouyang et al., 2007). The list of QTLs along with their genomic regions for different stresses, *viz.* salinity, bacterial, drought and cold are given in Supplementary Document S2.

The GSEQ analysis was used to test whether the informative gene sets (considering gene set of size 1000) selected for each stress through the gene selection techniques (Supplementary Table S2) is enriched with the underlying QTL information or not. In other words, the performance of Boot-MRMR and other techniques was assessed by using the GSEQ test as a biological relevant criterion to evaluate the reliability of selected genes. Further, the *p-values* along with the *NQTL* computed for each gene selection technique are given in Table 3. For salinity stress, the value of *NQTL* for Boot-MRMR (125) was higher than that of other gene selection techniques followed by SVM-RFE (120) and SRC (120) (Table 3). It was observed that more number of informative genes selected by Boot-MRMR were overlapped with salinity responsive QTL regions. Moreover, the *p-value* (1.02 E-77) from GSEQ test for the Boot-MRMR was much lesser than that of other techniques for salinity stress. It was found that informative genes selected by Boot-MRMR were more enriched with the underlying salinity responsive QTLs as compared to other 12 techniques. Similar interpretations can be made for cold and drought stresses in rice except bacterial stress (Table 3). Further, the QTL wise distributions of informative genes selected

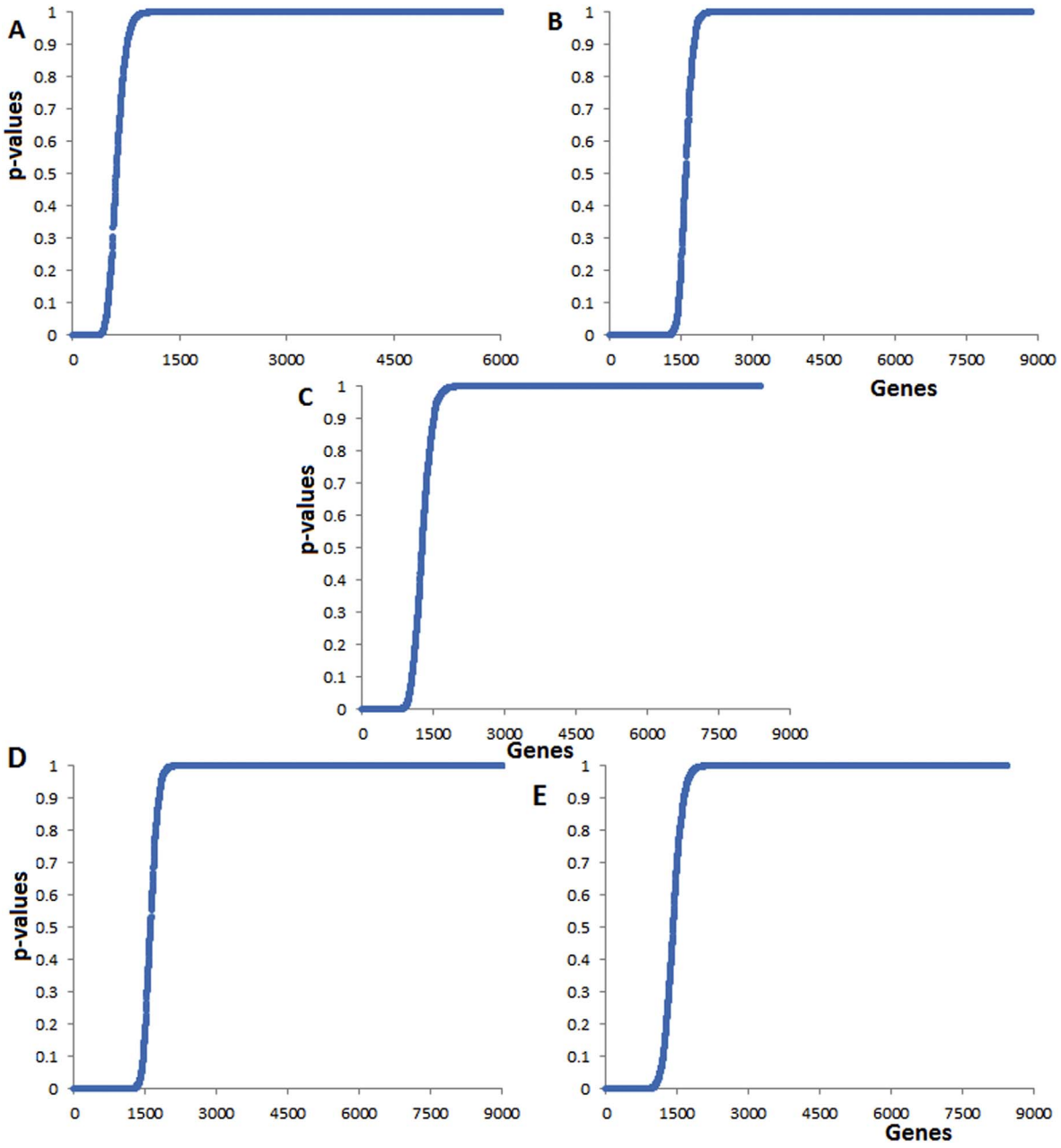


Fig. 3. Distribution of *p-values* obtained from the Boot-MRMR approach. The horizontal axis represents the genes. The vertical axis shows the computed statistical significance values (*p-values*) associated with the genes obtained from Boot-MRMR technique. Distributions of *p-values* are shown for (A) salinity stress, (B) cold stress, (C) aluminum stress, (D) drought stress and (E) bacterial stress.

through Boot-MRMR for all the stresses are given in Supplementary Document S4. However, for the bacterial stress, GSEQ test was found to be non-significant for all the gene selection techniques, which is due to the fact that only 4 QTLs are found to get QTL hit genes out of 24 QTLs (Fig. S5 in Supplementary Document S4).

3.3. Performance analysis based on GO terms

The GO based similarity analysis was performed on selected informative gene sets of size 1000 by each of the gene selection techniques (Supplementary Table S2) to evaluate functional similarities among these genes. The results from the GO based similarity analysis are given in Table 4. In salinity and bacterial stresses, out of three in two taxonomy categories, the magnitude of the developed biological similarity score for the Boot-MRMR was higher than other 12 techniques (Table 4). In other words, Boot-MRMR technique selects more functional similar genes for these stresses as compared to other competitive techniques. But for other two stresses, the biological similarity score of Boot-MRMR in one taxonomy category was higher than other contemporary techniques (Table 4). This analysis indicated that the proposed Boot-MRMR is competitive with other techniques of informative genes selection in terms of functional similarity. Further, associated GO terms with gene sets of size 1000 selected through proposed Boot-MRMR for all the stresses are given Supplementary Table S8.

3.4. Performance analysis based on computational time

Based on the average runtime required to select informative gene sets of size 1000, the 13 gene selection techniques were ranked. The results are given in Supplementary Table S9. The slowest method was SVM-RFE followed by RF, with computational training time up to

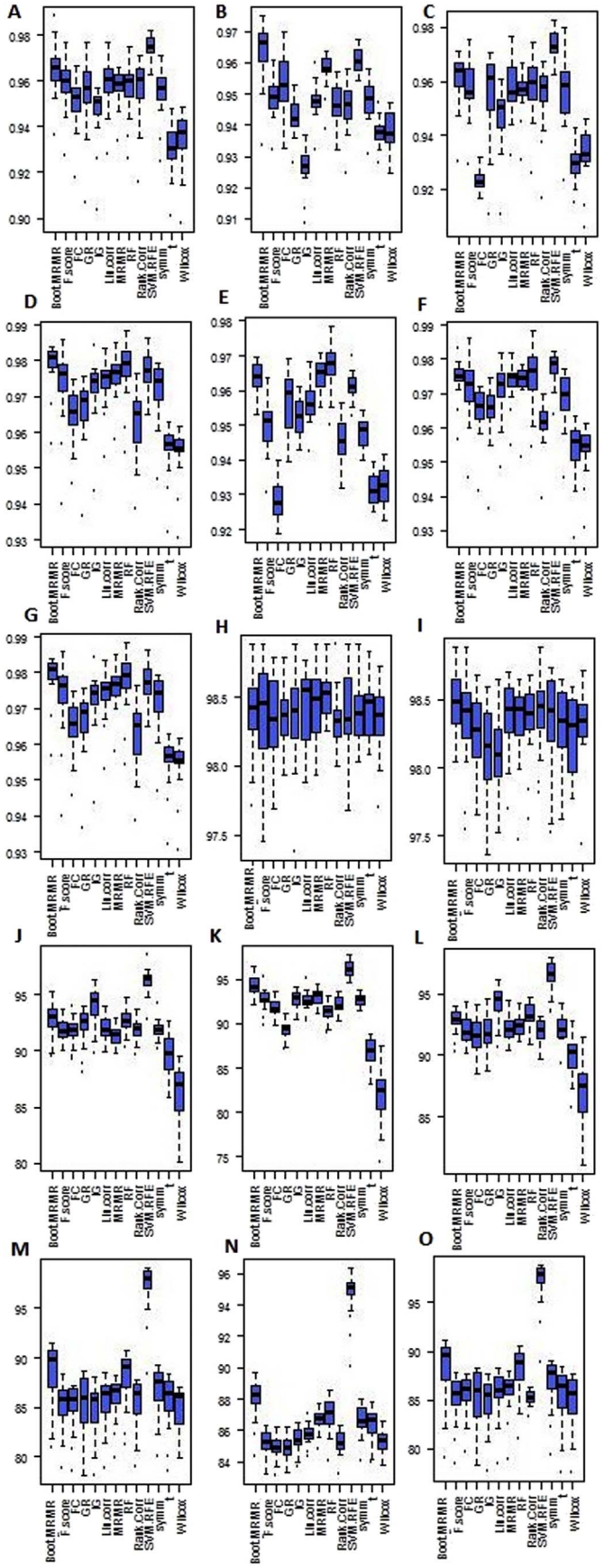


Fig. 4. Post selection classification accuracy and its standard error. The horizontal axis represents the gene selection techniques. The vertical axis represents post selection classification accuracy and its standard error obtained by using sliding window size technique. The classification accuracies over the window sizes are presented as boxes and standard error is shown in the form of bars on every boxes. The distributions of classification accuracy are shown for (A–C) salinity stress with SVM-LBF, SVM-PBF and SVM-RBF classifiers; (D–F) cold stress with SVM-LBF, SVM-PBF and SVM-RBF classifiers; (G–I) drought stress with SVM-LBF, SVM-PBF and SVM-RBF classifiers; (J–L) aluminum stress with SVM-LBF, SVM-PBF and SVM-RBF classifiers; (M–O) bacterial stress with SVM-LBF, SVM-PBF and SVM-RBF classifiers.

several hours, especially for larger datasets irrespective of all stresses. The proposed Boot-MRMR technique required far less computational time up to several minutes as compared to these two techniques (Supplementary Table S9) in all the datasets. However, simple univariate gene selection techniques *like* FC, t-score and F-score required less computational time for selection of informative genes.

3.5. Revealing conflicts among the criteria: one-criterion-at-a-time analysis

The performance of the 13 gene selection techniques under these 16 decision criteria (Supplementary Table S3) for rice GE datasets under various stresses *viz.* salinity, cold, drought and bacteria are given in Tables 3, 4 and Supplementary Tables S4–S6, S9. Subsequently, these techniques were evaluated individually by each of these 16 decision criteria for multiple datasets and are ranked in Supplementary Table S10. These rankings indicated that, Boot-MRMR was quite competitive with SVM-RFE and better than other existing techniques under the classification based criteria (C1–C10) (Supplementary Table S10). On the contrary, for QTL based criteria (C11 and C12), the proposed Boot-MRMR was found to be best for all the four stresses followed by SVM-RFE (Tables 3, S10). Further, through BP and CC based biological similarity analysis, the Boot-MRMR performed better as compared to other techniques for salinity stress (Tables 4, S10). But for MF based biological similarity score for same stress, SVM-RFE was ranked one followed by Boot-MRMR. Similar interpretations can be made about the ranking of gene selection techniques for cold, drought and bacterial stress datasets. Further, when computational time was taken as individual criterion, the simple technique *like* FC performed better followed by t-score (Supplementary Table S9).

The above comparative analysis of the gene selection techniques under each criterion individually (Supplementary Table S10) clearly showed the presence of conflicts among criteria in the given decision-making problem. For instance, SVM-RFE performed well for most of the datasets, when classification based criteria (C1–C10) were considered (Supplementary Tables S4-S6, S10), but performed poor under runtime (C16) and biology based criteria (C11–C15) (Tables 3, 4, S9). While the developed Boot-MRMR approach was found to be better when the biology based criteria, *i.e.* C11–C15 were considered but performed poor under runtime (Tables 3, 4, S10). Due to such conflicts in the performance of these techniques, the MCDM Entropy-TOPSIS approach was deemed necessary to choose the best gene selection option over the list of 13 options under these 16 decision criteria for each stress.

3.6. Selection of best gene selection technique: the TOPSIS approach

The TOPSIS scores and ranking of the gene selection techniques under three different classifiers for each stress are shown in Fig. 5. For salinity stress under SVM-LBF classifier, it was found that the proposed Boot-MRMR technique has highest TOPSIS score and subsequently found best for informative gene selection (Fig. 5). Similarly, for SVM-PBF and SVM-RBF classifiers under the same stress, the similar findings were observed (Fig. 5). For this stress, the performance of Boot-MRMR was found to be superior followed by SVM-RFE, IG, SU, RF, GR, FC, t-score, Wilcox, SRC, F-score, PCF and MRMR irrespective of the classifiers (Fig. 5). Similar interpretations can be made for cold and bacterial

Table 3
Performance evaluation of gene selection techniques based on GSEQ analysis.

Methods	Salinity stress		Cold stress		Drought stress		Bacteria stress	
	NQTL	*p-Value*	NQTL	*p-Value*	NQTL	*p-Value*	NQTL	*p-Value*
BMRMR	125	1.02E-77	149	1.06E-62	110	7.24E-55	62	0.9078
MRMR	117	1.34E-42	136	1.56E-34	107	4.04E-51	50	0.9227
SVMR	120	1.28E-60	121	2.71E-07	108	2.22E-54	54	0.9245
t	113	2.50E-41	137	1.98E-45	106	2.31E-39	50	0.9221
F	115	1.05E-37	134	7.34E-27	106	1.76E-44	46	0.9220
FC	110	6.43E-14	130	1.96E-43	102	3.47E-45	58	0.9206
PCF	117	1.00E-37	127	8.16E-10	105	1.03E-31	42	0.9290
SRC	120	4.73E-56	140	1.43E-29	103	2.00E-28	49	0.9221
IG	89	0.099	124	1.85E-10	93	1.17E-05	58	0.9207
GR	101	0.0245	97	0.001	98	5.56E-14	44	0.9237
RF	109	4.25E-11	131	1.36E-08	100	3.54E-21	43	0.9222
SU	108	4.44E-09	139	2.30E-20	99	8.36E-17	51	0.9229
Wilcox	116	1.45E-35	143	2.70E-50	103	4.19E-24	43	0.9221

Methods: Codes of the gene selection methods codes as given in Supplementary Table S2; NQTL: number of QTL hits within the selected gene set; *p-value*; statistical significance value from GSEQ test.

stresses with SVM-LBF, SVM-PBF and SVM-RBF classifiers (Fig. 5). But for drought stress, the TOPSIS score of SVM-RFE was found to be higher than that of other techniques and ranked top in the list followed by Boot-MRMR (Fig. 5). Moreover, three out of four stresses, the performance of Boot-MRMR technique was superior as compared to other gene selection techniques, whereas, for drought stress it is found to be quite competitive with SVM-RFE.

4. Boot-MRMR R software package

To facilitate the use of proposed informative gene selection approach, we have developed an R software package which includes BootMRMR R package accompanying documentation and model real data example. This package can be freely downloaded from https://cran.r-project.org/web/packages/BootMRMR. This software is capable of computing weights and *p-values* for genes using Bootstrap and modified bootstrap based resampling procedure. It also able to identify group of informative genes of given size based on the proposed approaches. Further, it provides function to compare the performance of gene selection methods under MCDM setup using TOPSIS technique.

5. Discussion

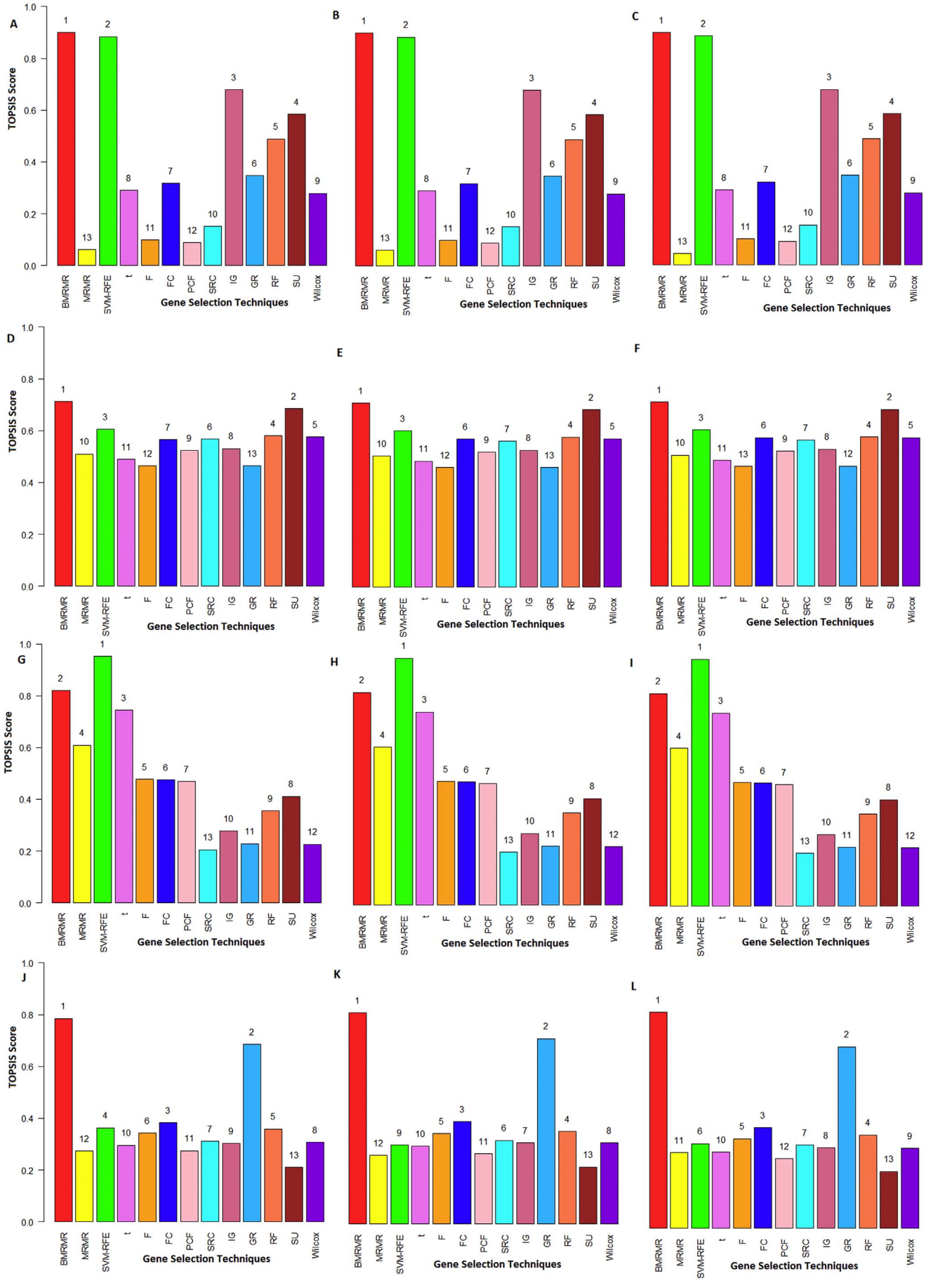
We proposed Boot-MRMR technique for selection informative genes from high dimensional GE data, which is not only effective to remove redundancy or collinearity among genes, but also improves relevancy of genes with the target trait/condition. Through this technique, informative genes were selected based on an assessment of the statistical significance of the hypothesis under consideration, which is more statistically convincing as compared to other gene ranking methods. Here, a *p-value* was assigned to each gene by using the NP test statistic and informative genes were selected based on these *p-values*. The *p-values* associated with genes represent more scientifically calculated interpretable values to genome researchers and experimental biologists (*e.g.* values between 0 and 1 with well-defined statistical meaning) as compared to other gene ranking techniques. Moreover, Boot-MRMR has advantage over classical techniques *like t*-test, F-score, PCF, *etc.* as it does not require Gaussian distributional assumption of the data. The performance of Boot-MRMR technique was found to be better over other techniques *like* MRMR, as the bootstrap procedure used in this approach was able to remove the spurious associations of genes with the target trait/condition as well as among other genes. The performance analysis showed that Boot-MRMR is either competitive or better with respect to other contemporary techniques *like* SVM-RFE, information theoretic measures and RF for development classification models.

The GSEQ analysis is a new way to validate the findings of the gene selection techniques, which is more biologically appealing than the traditional criterion based on classification. Through the GSEQ analysis, a meaningful measure, *i.e. p-value* was computed for the selected gene

Table 4
Performance evaluation of gene selection techniques based on biological similarity analysis.

Methods	Salinity stress			Cold stress			Drought stress			Bacteria stress		
	MF	BP	CC	MF	BP	CC	MF	BP	CC	MF	BP	CC
BMRMR	0.124	**0.156**	**0.187**	0.121	**0.166**	0.068	**0.173**	0.145	0.112	0.128	**0.189**	**0.116**
MRMR	0.085	0.126	0.059	0.129	0.151	0.053	0.132	0.135	0.116	0.096	0.114	0.096
SVMR	**0.152**	0.112	0.120	0.113	0.158	0.068	0.170	0.129	**0.132**	0.089	0.128	0.074
t	0.119	0.125	0.076	0.117	0.138	0.058	0.146	0.130	0.122	0.098	0.118	0.103
F	0.094	0.129	0.062	0.133	0.140	0.051	0.122	0.143	0.104	0.097	0.128	0.113
FC	0.114	0.114	0.079	0.125	0.155	0.058	0.104	0.117	0.122	0.120	0.109	0.086
PCF	0.102	0.134	0.057	**0.147**	0.154	0.054	0.116	0.138	0.109	0.092	0.117	0.111
SRC	0.088	0.140	0.067	0.126	0.155	0.057	0.093	0.141	0.086	0.093	0.127	0.108
IG	0.117	0.133	0.107	0.096	0.141	0.070	0.118	**0.161**	0.074	0.102	0.101	0.112
GR	0.101	0.143	0.082	0.112	0.103	**0.076**	0.116	0.131	0.064	**0.145**	0.177	0.075
RF	0.099	0.145	0.093	0.117	0.132	0.068	0.116	0.156	0.091	0.086	0.144	0.101
SU	0.098	0.143	0.101	0.133	0.151	0.064	0.096	0.142	0.115	0.075	0.120	0.104
Wilcox	0.094	0.134	0.077	0.125	0.155	0.058	0.089	0.146	0.090	0.095	0.123	0.111

Methods: gene selection methods codes as given in Supplementary Table S2; Values in table represents biological similarity score among genes within the selected gene set; MF: Molecular function GO terms; BP: Biological process; CC: Cellular component GO; values in bold indicate highest value.



(caption on next page)

Fig. 5. MCDM-TOPSIS analysis of gene selection techniques. The horizontal axis represents the gene selection techniques. The vertical represents TOPSIS score and the values shown on the top of the bars represent the ranks of the gene selection techniques. The figures are shown for shown for (A–C) salinity stress with SVM-LBF, SVM-PBF and SVM-RBF classifiers; (D–F) cold stress with SVM-LBF, SVM-PBF and SVM-RBF classifiers; (G–I) drought stress with SVM-LBF, SVM-PBF and SVM-RBF classifiers; (J–L) bacterial stress with SVM-LBF, SVM-PBF and SVM-RBF classifiers.

set. Based on the computed *p-value*, the performance of the gene selection techniques was inferred, as it is biological relevant criterion under a sound statistical framework. The GSEQ test and GO based biological similarity score provided several biologically reliable criteria for performance analysis of gene selection techniques. Further, through this performance analysis, it was found that Boot-MRMR approach selects more biologically relevant informative genes as compared to other existing techniques.

The interpretation of statistical significance values for both the approaches, *i.e.* Boot-MRMR and GSEQ were quite different and greatly depends on the sampling scheme which forms the basis of the test (Goeman et al., 2004). For Boot-MRMR, the *p-values* were computed for each gene (single gene testing) based on a self-contained null hypothesis and bootstrap procedure coupled in subject sampling model. While for GSEQ analysis the *p-values* were computed for each gene set (gene set testing) based on a competitive null hypothesis with the use of 2×2 contingency table method and gene sampling procedure. Further, this subject sampling model was the mirror image of the gene sampling procedure (with different sampling units), which reverses the meaning of statistical significance values in these two approaches. For Boot-MRMR approach, a significant *p-value* gives confidence that the given gene is biologically informative for the target condition/trait. On the contrary, for the GSEQ test, the significant *p-value* provides the strength of the gene selection technique, which selected biologically informative gene set.

The performance of the gene selection techniques was demonstrated on a broad spectrum of comparative metrics. These metrics include biological relevant criteria, *i.e.* GSEQ as well as biological similarity analysis, statistically meaningful criteria, *i.e.* performance metrics based on classification and computational time. Further, adjudging the performance of these techniques based on only CA, might lead to the selection of biologically irrelevant genes. Through the GSEQ analysis, it was evident that informative genes selected by the developed Boot-MRMR technique, were more enriched with the QTLs, *i.e.* more genes are associated with the QTL regions as compared to other gene selection techniques. Further, the GO based biological similarity analysis showed that the functional similarities exist among the informative genes selected by Boot-MRMR, which were comparable to those of popular techniques *like* SVM-RFE, information theoretic measures. However, the Boot-MRMR selects genes which are more biologically informative for the target trait/condition (which may have biological functions important to stress tolerance) as compared to other techniques, *viz.* t-score, F-score, RF, MRMR and correlation based approaches by eliminating redundant and spuriously associated genes. It may also be noted that, the proposed Boot-MRMR was not so computationally expensive and required less execution time to provide statistically and biologically informative minimal gene set as compared to other existing competitive techniques.

After informative gene set selection using classification based criteria, the selected gene set need to be scrutinized based on functional similarity and enrichment with the trait specific QTLs to assess its biological relevance. Ideally, a gene selection technique should consider both biological relevance measures and traditional statistical criteria to select an informative gene set. Such type of challenges could be easily solved by MCDM based approach of operation research. Therefore, in this case, we used MCDM-TOPSIS approach to identify best gene selection technique based on these diversified criteria on multiple crop datasets. This approach was also able to avoid the existing conflicts among the criteria to select the best gene selection technique. This was the first systematic and rigorous study to evaluate the performance of gene selection techniques under MCDM setup on multiple crop GE datasets. The MCDM-TOPSIS analysis revealed that for most of the datasets, the proposed Boot-MRMR was found to be best for informative gene selection over other existing techniques.

6. Conclusions

Selection of informative genes from available high dimensional GE data is a challenging task. In this study, we proposed a statistical approach for informative gene selection from such GE data by considering gene relevance and redundancy simultaneously. Here, the informative genes were selected based on the statistical criterion, which is more convincing as compared to other competitive gene selection techniques. Further, the GSEQ analysis provided two innovative biologically relevant criteria for performance analysis of gene selection technique(s). Through this, it was observed the gene set obtained by Boot-MRMR are more enriched with the underlying QTLs and has more functional similarity as compared to other techniques. Further, the systematic MCDM-TOPSIS analysis of the gene selection techniques revealed that Boot-MRMR approach is better method over the available alternatives with respect to a broad spectrum of criteria. The Boot-MRMR approach is independent of platforms on which gene expressions are measured. However, results may have little variations based on platforms as this method is based on resampling procedure, *i.e.* bootstrapping in which distributional properties of the data is being taken in to account. Hence, all major informative genes will remain same irrespective of platform. Further, the proposed approaches can be used for other case *vs.* control genomic studies including GE based on NGS data. The findings of this study will guide the genome researchers and experimental biologists to select informative gene set scientifically and objectively.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2018.02.044.

Availability of data and material

All the secondary data used in this study are available in the NCBI database. The proposed methods are implemented in the developed R package, which may be available form https://cran.r-project.org/web/packages/BootMRMR.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceived and designed the study: SD. Developed the methodology: SD. Performed the experiments: SD. Analyzed the data: SD. Contributed materials: SD, DCM. Drafted the manuscript: SD. Corrected the manuscript: SD, DCM, AR, SNR. Developed the R package (BootMRMR): SD.

Acknowledgment

Research reported in this publication was partially supported by National Institute of General Medical Sciences of NIH under Award Number P20GM113226. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH. Authors are also thankful to Indian Council of Agricultural Research (ICAR), New Delhi for providing Netaji Subhas-ICAR International Fellowship to Mr. Das. The help obtained from ICAR-Indian

Agricultural Statistics Research Institute, New Delhi was also duly acknowledged. Also Dr. S.N. Rai was partially supported by Wendell Cherry Chair in Clinical Trial Research. Extremely thankful to Swarnaprabha Chhuria, Ph.D. research scholar, OUAT, Bhubaneswar for her support and help.

References

Ahn, B.S., 2011. Compatible weighting method with rank order centroid: maximum entropy ordered weighted averaging approach. Eur. J. Oper. Res. 212, 552–559.

Al-Shahrour, F., Diaz-Uriarte, R., Dopazo, J., 2005. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics 21, 2988–2993.

Ash, R.B., 2000. Probability and Measure Theory. Harcourt and Academic Inc., New York.

Barry, W.T., Nobel, A.B., Wright, F.A., 2005. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 21, 1943–1949.

Benjamini, Y., Hochberg, Y., 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J. Stat. Plan. Inference 82 (1–2), 171–196.

Bland, M., 2013. Do Baseline p-values follow a uniform distribution in randomised trials? PLoS One 8 (10), e76010. http://dx.doi.org/10.1371/journal.pone.0076010.

Chen, M.F., Tzeng, G.H., 2004. Combining gray relation and TOPSIS concepts for selecting an expatriate host country. Math. Comput. Model. 40, 1473–1490.

Cheng, T., Wang, Y., Bryant, S.H., 2012. FSelector: a ruby gem for feature selection. Bioinformatics 28 (21), 2851–2852. http://dx.doi.org/10.1093/bioinformatics/bts528.

Cui, X., Churchill, G., 2003. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 4 (4), 210.

Das, S., Meher, P.K., Rai, A., Bhar, L.M., Mandal, B.N., 2017a. Statistical approaches for gene selection, hub gene identification and module interaction in gene co-expression network analysis: an application to aluminum stress in soybean (*Glycine max* L.). PLoS One 12 (1), e0169605. http://dx.doi.org/10.1371/journal.pone.0169605.

Das, S., Meher, P.K., Pradhan, U.K., Paul, A.K., 2017b. Inferring gene regulatory networks using Kendall's tau correlation coefficient and identification of salinity stress responsive genes in rice. Curr. Sci. 112 (6), 1257–1262.

Diaz-Uriarte, R., 2007. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics 8, 328. http://dx.doi.org/10.1186/1471-2105-8-328.

Díaz-Uriarte, R., de Andrés, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3. http://dx.doi.org/10.1186/1471-2105-7-3.

Ding, C., Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. J. Bioinforma. Comput. Biol. 3 (2), 185–205.

Du, L., Cai, C., Wu, S., Zhang, F., Hou, S., Guo, W., 2016. Evaluation and exploration of favorable QTL alleles for salt stress related traits in cotton cultivars (*G. hirsutum* L.). PLoS One 11 (3), e0151076. http://dx.doi.org/10.1371/journal.pone.0151076.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, London.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3, 1289–1305.

Goeman, J.J., Buhlmann, P., 2007. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23 (8), 980–987. http://dx.doi.org/10.1093/bioinformatics/btm051.

Goeman, J.J., van de Geer, S.A., de Kort, F., van Houwelingen, H.C., 2004. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20, 93–99.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Guoyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.

Hossain, A., Willan, A.R., Beyene, J., 2013. An improved method on Wilcoxon rank sum test for gene selection from microarray experiments. Commun. Stat. Simul. Comput. 42 (7), 1563–1577.

Hwang, C.L., Yoon, K., 1981. Multiple Attribute Decision Making. Springer-Verlag, Berlin.

Inza, I., Larranaga, P., Blanco, R., Cerrolaza, A., 2004. Filter *versus* wrapper gene selection approaches in DNA microarray domains. Artif. Intell. Med. 31, 91–103.

Khezrian, M., Jahan, A., Wan-Kadir, W.M.N., Ibrahim, S., 2014. An approach for web service selection based on confidence level of decision maker. PLoS One 9 (6), e97831. http://dx.doi.org/10.1371/journal.pone.0097831.

Kursa, M.Z., 2014. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics 15, 8. http://dx.doi.org/10.1186/1471-2105-15-8.

Kwong, C.K., Tam, S.M., 2002. Case-based reasoning approach to concurrent design of low power transformers. J. Mater. Process. Technol. 128, 136–141.

Lai, C., Reinders, M.J.T., van't Veer, L.J., Wessels, L.F.A., 2006. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics 7, 235. http://dx.doi.org/10.1186/1471-2105-7-235.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., et al., 2012. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (4), 1106–1119.

Liang, Y., Zhang, F., Wang, J., Joshi, T., Wang, Y., Xu, D., 2011. Prediction of drought-resistant genes in *Arabidopsis thaliana* using SVM-RFE. PLoS One 6 (7), e21750. http://dx.doi.org/10.1371/journal.pone.0021750.

Liu, X., Krishnan, A., Mondry, A., 2005. An entropy-based gene selection method for cancer classification using microarray data. BMC Bioinformatics 6, 76. http://dx.doi.org/10.1186/1471-2105-6-76.

Lord, P., Stevens, R., Brass, A., Goble, A., 2003. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics 19, 1275–1283.

Mao, K., Zhao, P., Tan, P.H., 2006. Supervised learning based cell image segmentation for p53 immunohistochemistry. IEEE Trans. Biomed. Eng. 53 (6), 1153–1163.

Mazandu, G.K., Mulder, N.J., 2014. Information content-based gene ontology functional similarity measures: which one to use for a given biological data type? PLoS One 9 (12), e113859. http://dx.doi.org/10.1371/journal.pone.0113859.

Milani, A.S., Shanian, A., Madoliat, R., Nemes, J.A., 2005. The effect of normalization norms in multiple attribute decision making models: a case study in gear material selection. Struct. Multidiscip. Optim. 29 (4), 312–318.

Mundra, P.A., Rajapakse, J.C., 2010. SVM-RFE with MRMR filter for gene selection. IEEE Trans. Nanobiosci. 9 (1), 31–37. http://dx.doi.org/10.1109/TNB.2009.2035284.

Ni, J., Pujar, A., Youens-Clark, K., Yap, I., Jaiswal, P., et al., 2009. Gramene QTL Database: Development, Content and Applications. Database (Oxford) 2009, bap005. http://dx.doi.org/10.1093/database/bap005.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., et al., 2007. The TIGR rice genome annotation resource: improvements and new features. Nucleic Acids Res. 35, D883–D887. http://dx.doi.org/10.1093/nar/gkl976.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27 (8), 1226–1238.

Rohatgi, V.K., Ehsanes Saleh, A.K.M., 2011. An Introduction to Probability and Statistics. John Wiley and Sons Inc., New Jersey.

Saeys, Y., Inza, I., Larran, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23 (19), 2507–2517. http://dx.doi.org/10.1093/bioinformatics/btm344.

Srdjevic, B., Medeiros, Y.D.P., Faria, A.S., 2004. An objective multi-criteria evaluation of water management scenarios. Water Resour. Manag. 18, 35–54.

The Gene Ontology Consortium, 2015. Gene ontology consortium: going forward. Nucleic Acids Res. 43 (Database issue), D1049–D1056. http://dx.doi.org/10.1093/nar/gku1179.

Tiwari, S., SL, K., Kumar, V., Singh, B., Rao, A.R., et al., 2016. Mapping QTLs for salt tolerance in Rice (*Oryza sativa* L.) by bulked segregant analysis of recombinant inbred lines using 50K SNP chip. PLoS One 11 (4), e0153610. http://dx.doi.org/10.1371/journal.pone.0153610.

Trevino, V., Falciani, F., Barrera-Saldana, H.A., 2007. DNA microarrays: a powerful genomic tool for biomedical and clinical research. Mol. Med. 13 (9–10), 527–541. http://dx.doi.org/10.2119/2006-00107.

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F., 2007. A new method to measure the semantic similarity of GO terms. Bioinformatics 23 (10), 1274–1281. http://dx.doi.org/10.1093/bioinformatics/btm087.

Wang, J., Chen, L., Wang, Y., Zhang, J., Liang, Y., Xu, D., 2013. A computational systems biology study for understanding salt tolerance mechanism in Rice. PLoS One 8 (6), e64929. http://dx.doi.org/10.1371/journal.pone.0064929.

Wang, Y., Xi, C., Zhang, S., Zhang, W., Yu, D., 2015. Combined approach for government e-tendering using GA and TOPSIS with intuitionistic fuzzy information. PLoS One 10 (7), e0130767. http://dx.doi.org/10.1371/journal.pone.0130767.

Wei, Q., Dunbrack Jr., R.L., 2013. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One 8 (7), e67863. http://dx.doi.org/10.1371/journal.pone.0067863.

Yang, T., Chou, P., 2005. Solving a multi-response simulation–optimization problem with discrete variables using a multi-attribute decision-making method. Math. Comput. Simul. 68, 9–21.