doi:10.20944/preprints202009.0699.v1

Statistical Approach for Biologically Relevant Gene Selection from High-Throughput Gene Expression Data

Samarendra Das[1-4], Shesh N Rai[3-8,*]

[1]Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India

[2]Netaji Subhas-ICAR International Fellow, Indian Council of Agricultural Research, Krishi Bhawan, New Delhi 110001, India

[3]Biostatistics and Bioinformatics Facility, JG Brown Cancer Center, University of Louisville, Louisville 40202, KY, USA

[4]School of Interdisciplinary and Graduate Studies, University of Louisville, Louisville 40292, KY, USA

[5]University of Louisville Alcohol Research Center, University of Louisville, Louisville 40202, KY, USA

[6]Department of Hepatobiology and Toxicology, University of Louisville, Louisville 40292, KY, USA

[7]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville 40202, KY, USA

[8]Wendell Cherry Chair in Clinical Trial Research, University of Louisville, Louisville, KY 40202, USA

Authors' email addresses:

SD: samarendra.das@louisville.edu

SNR: shesh.rai@louisville.edu

***To whom correspondence should be addressed-** email: shesh.rai@louisville.edu; Telephone: +1 502-426-0016

Abstract

Selection of biologically relevant genes from high dimensional expression data is a key research problem in gene expression genomics. Most of the available gene selection methods are either based on relevancy or redundancy measure, which are usually adjudged through post selection classification accuracy. Through these methods the ranking of genes was done on a single high-dimensional expression data, which leads to the selection of spuriously associated and redundant genes. Hence, we developed a statistical approach through combining Support Vector Machine with Maximum Relevance and Minimum Redundancy under a sound statistical setup for the selection of biologically relevant genes. Here, the genes are selected through statistical significance values computed using a non-parametric test statistic under a bootstrap based subject sampling model. Further, a systematic and rigorous evaluation of the proposed approach with nine existing competitive methods was carried on six different real crop gene expression datasets. This performance analysis was carried out under three comparison settings, *i.e.* subject classification, biological relevant criteria based on quantitative trait loci and gene ontology. Our analytical results showed that the proposed approach selects genes which are more biologically relevant as compared to the existing methods. Moreover, the proposed approach was also found to be better with respect to the competitive existing methods. The proposed statistical approach provides a framework for combining filter, and wrapper methods of gene selection.

Availability: BSM R software Package is available at https://github.com/sam-uofl/BSM.

Keywords: SVM, MRMR, Bootstrap, Genes, Gene Expression, Biological Relevance, Subject Classification

1. Background

Emergence of high-throughput sequencing technologies exponentially increase the size of output data in genome sciences with respect to number of features [1]. For example, Gene Expression (GE) studies generate the expression measurements of several thousand(s) of genes for tissue samples over two contrasting conditions in a single study [2,3]. These huge amount of expression data being generated for complex traits, and are deposited in public domain databases, such as NCBI GEO, ArrayExpress, *etc*., over the years by researchers across the globe [4,5]. Further, these publicly available high-throughput data need to be analyzed for gaining valid biological insights. One such aspect of research is to select genes, which are highly relevant to the phenotype/trait under study, out of several thousands of genes in the data. This is called feature selection in machine learning in general and gene selection in genomics [5–7]. Gene selection has been the focused area of functional genomics research, and thus several statistical and machine learning approaches have been developed for this purpose [8,9]. The main aim of gene selection is to reduce the curse of high-dimensionality in the expression data [5,6,10,11]. Further, these selected genes are used as predictors for further predictive analysis, *i.e.* subjects classification [7,8,11], gene regulation modeling [12], gene network analysis [5,6], *etc*.

Gene selection methods can be grouped into: (*i*) filter; and (*ii*) wrapper methods [9,13]. Filter methods select individual genes or evaluate a gene subset based on a performance measure computed from the data with respect to class variables regardless of the predictive modeling algorithm [14]. Further, these methods include univariate approaches such as t-test [15,16], Fold change [16], F-score [17,18], Volcano plot [15], Wilcoxon's statistic (Wilcox) [19,20], Information Gain (IG) [21,22], Gain Ratio (GR) [21,22], symmetric uncertainty [19], *etc*. These methods select genes by only considering their relevance within a level of the experimental condition/trait. However, these approaches may not sufficient to discover some complex relationships among genes (*i.e.,* gene-gene interactions) for certain conditions/traits, under which the data is generated [10]. Therefore multivariate filter approaches, such as Pearson's Correlation (PCR), Spearman's rank correlation [21,22], Maximum Relevance and Minimum Redundancy (MRMR) [17,23,24], *etc*. have been developed to select genes from GE data [9,13].

doi:10.20944/preprints202009.0699.v1

Wrapper methods select gene subsets through assessing the performance of the predictive modelling algorithm [25]. In other words, this class of gene selection methods are embedded in the classification process. For instance, a wrapper method evaluates the gene subsets based on the classifiers' performance on GE data and selects the most relevant gene subset. However, the Wrapper methods have better performance over filter methods [13,22], but are more complex and computationally expensive [25]. This class includes Support Vector Machine-Recursive Feature Elimination (SVM-RFE) [8,26], Multiple SVM-RFE (MSVM-RFE) [27] and Random Forest (RF) [11], to name a few. Further, hybrids of filter and wrapper methods are also reported in literature (known as embedded methods [9]) such as combination of SVM-RFE with MRMR weights (SVM-MRMR) [28], SVM with F-score and other methods [18] to select relevant genes from GE data. Besides, hybrid gene selection methods through combining ReliefF with ant colony optimization [29], and particle swarm optimization algorithms [30] are also developed to select cancer responsible genes from GE data. Moreover, the existing methods select genes through the weights (*i.e.* gene ranking criteria) computed from single high-dimensional GE data, which lead to selection of spuriously associated and redundant genes (*i.e.* genes may not informative but are correlated with other relevant genes) [5,6]. Therefore, the permutation procedures are used to compute statistical significance values for genes [6]. However, it has some serious limitations, such as highly sensitive to small permutation of experimental conditions (*i.e.* class labels) [5,6], computationally slow [31,32], cannot possibly give any significant *p-values* after multiple testing adjustments [32,33], large permutations are required to get a significant *p-value* [32]. To address such issues, bootstrap procedures are used in gene selection which ably remove the spurious associations of genes with the classes and other genes [6,34,35].

Gene selection methods are mostly used to select cancer responsible genes from GE datasets, and subsequently used for patient classification (*e.g.* with and without cancer) [6,7,35–37,8,28–34]. There are limited studies available in literature to systematically explore the performance of gene selection methods on crop GE datasets as there are typically limited experimental data available. Further, the performance of the existing methods were usually assessed through computation of post selection Classification Accuracy (CA) on cancer GE datasets [7,8,28,35–37]. In other words, these techniques are adjudged based on their ability to discriminate the GE samples between case and control groups. Here, it is worthy to note, this traditional criterion is statistically sound but may not be biologically relevant for performance evaluation of gene selection methods [35,38]. For instance, a gene selection technique identified a set of genes which accurately predicts the class of GE samples for a salinity *vs.* control GE study in rice, but it fails to tell whether these selected genes are biologically relevant or not to the salinity stress. Hence, it is pertinent to evaluate the gene selection methods with respect to biology-based criteria. For this purpose, data related to traits, such as Quantitative Trait Locus (QTLs) and Gene Ontology (GO) for model crop plants may be used, which are hugely available in public domains.

We, therefore, propose an improved statistical approach (BSM=Bootstrap-SVM-MRMR) that combines MRMR filter with SVM wrapper method to minimise the redundancy among genes and improve the relevancy of genes with the traits/phenotype under a sound statistical setup. Through this, relevant genes are selected from a high-dimensional GE data through the statistical significance values computed using a Non-Parametric (NP) test statistic under a bootstrap based subject sampling model. Further, the comparative performance analysis of the proposed BSM approach is carried out with nine existing competitive methods (*i.e.* IG [21,22], GR) [21,22] , t-test [15,16], F-score [17,18], MRMR [17,39] , SVM-RFE [8,26], SVM-MRMR [28], PCR [21,22] and Wilcox [19,20]). The comparative performance measures include, CA along with its standard error computed through varying sliding windows size technique, and two biological criteria based on QTL [40], and GO [41] terms. We demonstrate these procedures on six publicly available, independent crop GE datasets, and find that the BSM approach outperforms in terms of classification and biological relevance criteria compared to the existing methods. We have also developed BSM R package for public use.

2. Materials and Methods

2.1 Motivation

The genes expression datasets, from various experiments conducted to understand the behavior of biological mechanisms, are hugely available in public domain databases. For example, GE datasets generated for 125,376 experiments over 19,893 Microarray platforms consists data on 3,406,218 samples are available in NCBI GEO database till date [4]. Usually, researchers used data from single experiment to test their methodology or select genes for further study. For instance, Wang *et al.* (2013) used the salinity stress GE samples from GSE14403 to test their methodology, and select salinity responsive genes to understand salinity tolerance mechanism in rice [6]. Such study is important but may not be enough to test the hypothesis of salinity tolerance in rice due to limited sample size. Hence, the real challenge is to integrate or combine the GE datasets generated for same or cross platforms over different experimental conditions and test the methodology(s) on the meta-data. Moreover, meta-analysis of data generated by GE experiments for the same or related stress(s) is essential to enhance the sensitivity of the hypothesis under consideration for drawing valid biological conclusions. Therefore, we performed meta-analysis on GE datasets correspond to different stresses from multiple experiments and tested the performance of methods on these metadata, as shown in Table 1. The outlines of meta-analysis are given in Figure 1A.

Table 1. Rice gene expression datasets used in the study.

Sl. No	Descriptions	#Series	Series Id	#Genes	#Samples	Stress type
1	Salinity stress	3	GSE14403, GSE16108, GSE6901	6637	45 (23, 22)	Abiotic
2	Cold stress	4	GSE31077, GSE33204 GSE37940, GSE6901	8840	28 (15, 13)	Abiotic
3	Drought stress	5	GSE6901, GSE26280 GSE21651, GSE23211 GSE24048	9078	70 (35, 35)	Abiotic
4	Bacterial (xanthomonas) stress	3	GSE19239, GSE36093 GSE36272	8356	74 (37, 37)	Biotic
5	Fungal (blast) stress	2	GSE41798, GSE7256	7072	26 (13, 13)	Biotic
6	Insect (brown plant hopper) stress	1	GSE29967	7241	18 (12, 6)	Biotic

#Series: Number of GEO series for each dataset; #Genes: Number of genes; #Samples: Number of GEO samples; (x, y): number samples for case and control respectively; #Class: Number of classes (*e.g.* 2 in control *vs.* stress genomic study)

2.2 Data source

Rice GE experimental datasets were collected from Gene Expression Omnibus database of NCBI for platforms GPL2025 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2025) [4]. Here, we used the rice data, as it is a model crop plant, huge amount GE and other related biological (QTL and GO) datasets are available publicly, and its genome is well annotated. These GE datasets were generated under biotic (bacterial (*Xanthomonas*), fungal (Blast), and insect (Brown plant hopper)), and abiotic (salinity, cold and drought) stresses in rice. The summary and details of these datasets are given in Table 1 and Supplementary Table S1, respectively. Further, the detail descriptions of data collection, pre-processing, meta-analysis, and preliminary gene selection of these datasets are given in Supplementary Document S1. The QTL datasets for the stresses in rice, *viz.* salinity, drought, cold, insect, fungal and bacterial, were collected from the Gramene QTL database (http://www.gramene.org/qtl/) [42]. The lists of the respective stress responsive QTLs along with their mapped positions on the genome are given in Supplementary Document S2. The GO annotations data of the rice genome used in this study were collected from AGRIGO database [43].

2.3 Methods

2.3.1 Notations

Let, $X_{N \times M} = [x_{i,m}]$ be the GE data matrix, where $x_{i,m}$ represents the expression of i^{th} (i=1, 2, …, N) gene in m^{th} (m = 1, 2, …, M) sample/subject; $\boldsymbol{x}_m$ be the N-dimensional vector of expression values of genes for m^{th} sample; y_m be the outcome variable for target class label of m^{th} sample and takes values, {+1, -1} for case and control conditions respectively; M_1 and M_2 be the number of GE samples in case and control classes respectively ($M_1 + M_2 = M$); $(\bar{x}_{i,1}, S^2_{i,1})$ and $(\bar{x}_{i,2}, S^2_{i,2})$ be the mean and variance of i^{th} gene for case and control classes respectively; $\bar{x}_i$ be the mean of i^{th} gene across all M samples; $S_{i,j}$ be the co-variance between i^{th} and j^{th} genes.

2.3.2 *Maximum Relevance and Minimum Redundancy (MRMR) Filter*

MRMR method aims at selecting maximally relevant and minimally redundant set of genes for discriminating the tissue samples (*e.g.* case *vs*. control). This method is extensively used for selection of cancer responsible genes from high-dimensional GE data for patient classification (*i.e.* with and without cancer) [17,24,39]. For continuous GE data (*e.g.* Microarrays), the relevance measure for i^{th} gene over the given classes (*i.e.* case and control), is computed through F-statistic [39] and is expressed as:

$$F(i) = \frac{M_1(\bar{x}_{i,1}-\bar{x}_i)^2+M_2(\bar{x}_{i,2}-\bar{x}_i)^2}{\{(M_1-1)S_{i,1}^2+(M_2-1)S_{i,2}^2\}/(M-2)} \tag{1}$$

Further, the redundancy measure in MRMR method is computed through Pearson's correlation (ignoring the class information) for continuous GE data [39] and is given as:

$$R(i,j) = Corr(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{S_{i,j}}{S_i S_j} = \frac{\sum_{m=1}^{M}(x_{im}-\bar{x}_\iota)(x_{jm}-\overline{x_J})}{\sqrt{\sum_{m=1}^{M}(x_{im}-\bar{x}_\iota)^2}\sqrt{\sum_{m=1}^{M}(x_{jm}-\overline{x_J})^2}} \tag{2}$$

In MRMR method, genes are ranked by the combination of relevance, and redundancy measures under F-score with Correlation Quotient scheme for continuous GE data [17,24,39]. The weights computed through MRMR method for gene ranking can be expressed in terms of Eq. 1 and 2, and is given as:

$$w_i = F(i)/\{\frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} |R(i,j)|\} \qquad \forall \quad i = 1, 2, \dots, N \tag{3}$$

where, w_i (≥ 0) is the weight associated with i^{th} gene. The functions $F(i)$ and $R(i, j)$ in Eq.3 represent F-statistic for i^{th} gene and Pearson's correlation co-efficient between i^{th} and j^{th} genes. In other words, the i^{th} gene weight is F-statistic adjusted with average absolute correlation of i^{th} gene with the remaining genes.

2.3.3 Support Vector Machine (SVM)

SVM method is used for selection of genes (in a two group case) from high dimensional GE data [26]. Further, $\{\boldsymbol{x}_m, y_m\} \epsilon R^Nx\{-1, 1\}$ is given as input to SVM. Here, we wish to find out a hyperplane that divides the GE samples/subjects for case ($y_m = 1$) from that of control class ($y_m = -1$) in such a way that the distance between the hyperplane and the point, $\boldsymbol{x}_m$, is maximum. Then the hyperplane can be written as:

$$\sum_{i=1}^{N} k_i \boldsymbol{x}_{i,m} + b = 0 \qquad \forall\, m = 1, 2, \dots, M \tag{4}$$

where, k_i and b are the weight of i^{th} gene and bias respectively. Here, we assume that the GE samples for two classes are linearly separable. In other words, we can select two parallel hyperplanes that separate the case, and control classes in such a way that the distance between them is maximum.

For case class, the hyperplane becomes:

$$\sum_{i=1}^{N} k_i x_{i,p} + b = 1 \qquad \text{for any } p = 1, 2, \dots, M_1 \tag{5}$$

For control class, the hyperplane becomes:

$$\sum_{i=1}^{N} k_i x_{i,q} + b = -1 \qquad \text{for any } q = 1, 2, \dots, M_2 \tag{6}$$

The expressions in Eq. 5, and 6 can be combined as:

$$y_m\left(\sum_{i=1}^{N} k_i \boldsymbol{x}_{i,m} + b\right) = 1 \qquad \forall\, m = 1, 2, \dots, M \tag{7}$$

Here, we wish to maximize the distance between the case, and control hyperplanes in Eq. 5 and 6 respectively under the constraint that there will be no GE samples between these two hyperplanes given in Eq. 7. Mathematically, it can be written as:

$$\sum_{i=1}^{N} \frac{k_i}{(\sum k_i)^2} \left|x_{i,p} - x_{i,q}\right| = \frac{\mathbf{2}}{(\sum k_i)^2} \tag{8}$$

So, to maximize the distance between the planes in Eq. 8, we need to minimize $\frac{(\sum_i k_i)^2}{2}$ under the constraint of Eq. 7. Mathematically, it can be written as:

$$L_p = \min_{k_i} \frac{(\sum_i k_i)^2}{2} + \sum_{m=1}^{M} \varphi_m \left\{1 - y_m\left(\sum_{i=1}^{N} k_i x_{i,m} + b\right)\right\} \quad \forall\, m = 1, 2, \dots, M \tag{9}$$

where, φ_m ($\geq$ 0): Lagrange multiplier. Here, k_i's are obtained by minimizing the objective function in Eq. 9. Through the principle of maxima-minima, we have:

$$\frac{\partial L_p}{\partial k_i} = \sum_i k_i - \sum_i\left(\sum_{m=1}^{M} \varphi_m y_m x_{i,m}\right) = 0 \;\text{ and }\; \frac{\partial L_p}{\partial b} = \sum_{m=1}^{M} \varphi_m\, y_m = 0 \tag{10}$$

Moreover, from Eq. 10, the k_i can be expressed as:

$$k_i = \sum_{m=1}^{M} \varphi_m y_m x_{i,m} \quad \text{with } \sum_{m=1}^{M} \varphi_m\, y_m = 0 \text{ and } \varphi_m \geq 0 \tag{11}$$

Here, $|k_i|$ ($\geq$ 0) in Eq. 11 is used as a metric for ranking of genes in the GE data [26]. Alternatively, k_i^2 as a gene ranking metric can also be derived by using Taylor series approximation [44], which is given in Supplementary Document S3.

2.3.4 Proposed hybrid approach of gene selection

MRMR method may not yield optimal CA because it performs independent of the classifier and only involved in selection of genes [28]. On the contrary, SVM method of gene selection does not consider the redundancy among genes (*i.e.* gene-gene correlations) while selecting genes [28]. Hence, Mundra and

Rajapakse (2010) have developed a gene selection method by taking linear combination of weights computed through MRMR, and SVM methods [28], and is given as:

$$SL_i = \delta w_i + (1-\delta)|k_i| \tag{12}$$

where, parameter $\delta \in [0, 1]$ decides the tradeoff between SVM and MRMR weights. The SL_i in Eq. 12 is highly dependent on the value of δ. In other words, the choice of δ may alter the order of genes by MRMR (w_i) or by SVM (k_i); especially when w_i and k_i are negatively correlated. Hence, we propose a statistical approach by combining SVM and MRMR weights under sound statistical framework, where genes are selected through *p-values* computed using the NP test statistic, which is described as follows.

First, we normalized the w_i, and k_i's through minimax normalization. Then ranked w_i and k_i based on the ascending order of their magnitudes and assign ranks $\gamma_i^{MR}$and γ_i^{SV} for i^{th} gene, respectively. Then, we developed a technique, *i.e.* quadratic integration, for integrating the gene scores based on ranks, which automatically assigned more weights to higher value of w_i and k_i. Now, the quadratic integration score can be expressed as:

$$SD_i = \frac{\beta\gamma_i^{MR}{w_i}^{norm} + (1-\beta)\gamma_i^{SV}|k_i|^{norm}}{\beta\gamma_i^{MR} + (1-\beta)\gamma_i^{SV}} \tag{13}$$

where, ${w_i}^{norm}$ and $|k_i|^{norm}$ are the normalized values, expressed in Eq. 14 and 15, respectively.

$${w_i}^{norm} = \Big(w_i - \min_i w_i)/(\max_i w_i - \min_i w_i\Big) \tag{14}$$

$$|k_i|^{norm} = (|k_i| - \min_i |k_i|)/(\max_i |k_i| - \min_i |k_i|) \tag{15}$$

Further, $\beta\big(\in (0, 1)\big)$ in Eq. 13 is determined empirically from the data through five-fold cross validation technique. The detail procedure for determining the optimum value of β is given in Supplementary Document S4. If SD_i in Eq. 13 is alone used for ranking of genes, it will become a filter approach and lead to selection of spuriously associated genes. Hence, we used bootstrap procedure under a subject sampling model setup to obtain the empirical distribution of SD_i for computation of statistical significance value for i^{th} (i=1, 2, …, N) gene. The bootstrap procedure is described as:

The *M* samples (as columns) in the GE data matrix either belong to case or control, can be considered as subjects/units in a population model, as shown in Eq. 16.

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m), \ldots, (\boldsymbol{x}_{M-1}, y_{M-1}), (\boldsymbol{x}_M, y_M) \tag{16}$$

Here, we assume that the subjects are independent and identically distributed (iid), but the genes within each subject may be correlated. In the bootstrap procedure, *M* units are randomly drawn from *M* population units in Eq. 16 with replacement to constitute a bootstrap GE data matrix, *i.e.* $X_{NXM}^{(b)}$ (*M* units serve as *M* columns of *X*). This process is repeated *B* times to get *B* bootstrap GE data matrices, *i.e.* $X_{NXM}^{(1)}, X_{NXM}^{(2)}, \ldots, X_{NXM}^{(b)}, \ldots, X_{NXM}^{(B)}$. Here, *B* (*i.e.* number of bootstrap samples) must be sufficiently large. So, we set *B*=200 as several empirical studies showed that the number of bootstrap samples required for an estimation procedure is ~ 200 [6,45].

Now, the *B* bootstrap GE data matrices are given as input to Eq. 3, 11, and 13 to compute the *SD* scores, and subsequently gene ranking is performed on each of the *B* bootstrap GE data matrices.

Let, P_{ib}, be a random variable (*rv*) shows position of i^{th} gene in b^{th} bootstrap GE matrix. Then, another *rv* can be defined based on P_{ib} (without loss of generality), given as:

$$R_{ib} = \frac{N+1-P_{ib}}{N} \,;\, 0 \leq R_{ib} \leq 1 \tag{17}$$

where, R_{ib} in Eq. 17 is the rank score of i^{th} (*i*=1, 2, …, *N*) gene in b^{th} (*b*=1, 2, …, *B*) bootstrap GE matrix. Here, it may be noted that the distribution of the rank scores of genes, computed from a bootstrap GE data matrix, is symmetric around the median value (as rank scores are function of ranks). The values of median and the third quartile (Q_3) are given as 0.5 and 0.75, respectively.

To decide, whether i^{th} gene is biologically relevant or not to the condition/trait under study, the following null hypothesis can be tested.

$$H_0: R_i \leq Q_3 \ (i-th \text{ gene is not so relevant to the trait})$$

$$H_1: R_i > Q_3 \ (i-th \text{ gene is relevant to the trait})$$

where, R_i is the rank score for i^{th} gene over all possible bootstrap samples.

To obtain the distribution of test statistic under H_0, we define another *rv* Z_{ib}, as:

$$Z_{ib} = \begin{cases} 1 & |R_{ib} - Q_3| > 0 \\ 0 & |R_{ib} - Q_3| < 0 \end{cases} \tag{18}$$

Let, r_{ib} be the rank assigned to $(R_{ib} - Q_3)$ (after arranging in ascending order of their magnitudes). To test *H₀ vs. H₁* the test statistic for i^{th} gene, W_i, is developed, and is given as:

$$W_i = \sum_{b=1}^{B} Z_{ib} r_{ib} = \sum_{b=1}^{B} U_{ib} \ (say) \tag{19}$$

In other words, W_i in Eq. 19 is the sum of the ranks of positive signed scores for i^{th} gene over B bootstrap samples. Further, U_{ib} in Eq. 19 is a Bernoulli *rv,* and its probability mass function can be given as:

$$P[U_{ib} = u_{ib}] = \begin{cases} \frac{3}{4} & if\ u_{ib} = 0 \\ \frac{1}{4} & if\ u_{ib} = 1 \end{cases} \tag{20}$$

Here, the expected value and variance of W_i in Eq. 19 under H_0 can be obtained as:

$$E(W_i) = \sum_{b=1}^{B} E(U_{ib}) = \sum_{b=1}^{B} (0.\tfrac{3}{4} + b.\tfrac{1}{4}) = \tfrac{1}{4} \sum_{b=1}^{B} b = \frac{B(B+1)}{8} \tag{21}$$

The variance becomes:

$$V(W_i) = E\left(W_i^2\right) - [E(W_i)]^2$$

$$= \sum_{b=1}^{B} E\left(U_{ib}^2\right) - \sum_{b=1}^{B} E(U_{ib})^2$$

$$= \sum_{b=1}^{B} \left(\frac{b^2}{4} - \frac{b}{16}^2\right) = \frac{B(B+1)(2B+1)}{32} \tag{22}$$

As B is sufficiently large, then under central limit theorem, the distribution of W_i in Eq. 19 becomes:

$$Z_i = \frac{W_i - E(W_i)}{\sqrt{V(W_{i)}}} \xrightarrow{d} N(0,1) \tag{23}$$

Through the Eq. 23, the *p-value* for i^{th} (i=1, 2, …, N) gene is computed and similarly this testing procedure is repeated for the remaining N-1 genes. Let, $p_1, p_2, \ldots, p_N$ be the corresponding *p-values* for all the genes in GE data, and α be the level of significance. Here, we assume that all genes in the GE data are equally important for the trait development, hence, we employed Hochberg procedure [46] for correcting the multiple testing, and to compute the adjusted (*adj.*) *p-values* for genes. It is worthy to note that Hochberg's procedure is computationally simple, quite popular in genomic data analysis [47] and more powerful than Holm's procedure [48]. The algorithm for Hochberg's procedure [46] is given as:

Step 1. If $p_{(l)} > \alpha$, then retain corresponding null hypothesis ($H_{(l)}$) and go to the next step. Else reject it and stop.

Step $i = 2, 3, \dots, N-1$. If $p_{(N-i+1)} > \alpha/i$, then retain $H_{(N-i+1)}$ and go to the next step. Else reject all remaining hypotheses and stop.

Step *N*. If $p_{(1)} > \alpha/N$, then retain ($H_{(1)}$). Else reject it.

Now, the *adj. p-values* are given recursively beginning with the largest *p-value* [46]:

$$\widetilde{p_{(i)}} = \begin{cases} p_{(i)} \quad if\ i{=}N \\ \min\left(\widetilde{p}\ (i+1), (N-i+1)p_{(i+1)}\right. \quad if\ i = N-1, \dots, 1 \end{cases} \tag{24}$$

Further, based on the computed *adj. p-values,* the relevant genes are selected from the GE data. In other words, lesser value of *adj. p-value* indicates more relevance of the gene for the target trait, and *vice-versa*. The outlines and key analytical steps of the proposed approach are shown in Figure 1.

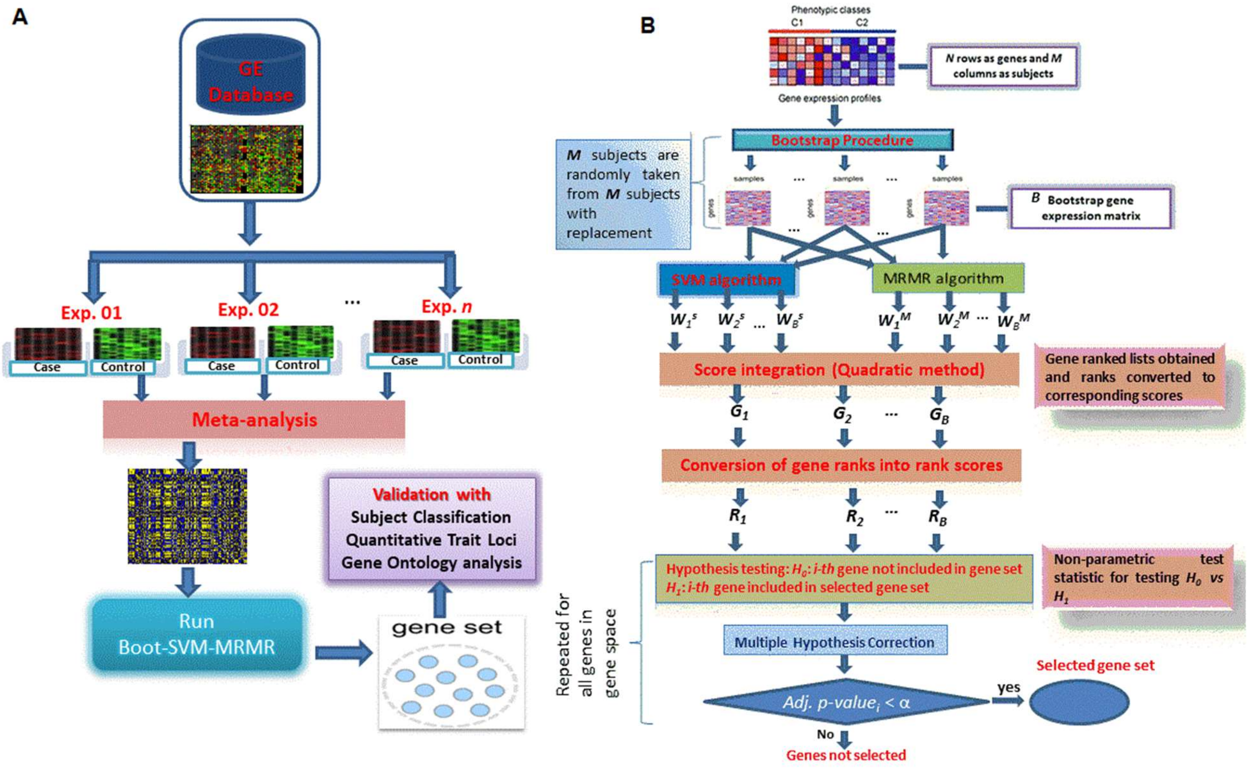
A
GE Database
Exp. 01
Exp. 02
...
Exp. *n*
Case
Control
Case
Control
Case
Control
Meta-analysis
Validation with
Subject Classification
Quantitative Trait Loci
Gene Ontology analysis
Run
Boot-SVM-MRMR
gene set
B
Phenotypic classes
C1
C2
N rows as genes and *M* columns as subjects
Gene expression profiles
Bootstrap Procedure
M subjects are randomly taken from *M* subjects with replacement
samples
genes
B Bootstrap gene expression matrix
SVM algorithm
MRMR algorithm
W_1^S W_2^S ... W_B^S
W_1^M W_2^M ... W_B^M
Score integration (Quadratic method)
Gene ranked lists obtained and ranks converted to corresponding scores
G_1 G_2 ... G_B
Conversion of gene ranks into rank scores
R_1 R_2 ... R_B
Hypothesis testing: H_0: *i-th* gene not included in gene set
H_1: *i-th* gene included in selected gene set
Non-parametric test statistic for testing H_0 vs H_1
Repeated for all genes in gene space
Multiple Hypothesis Correction
Adj. p-value$_i$ < α
yes
Selected gene set
No
Genes not selected

Figure 1. Operational procedure of proposed BSM gene selection approach. (A) Outline of the proposed study; (B) Flowchart depicting the implemented algorithm of BSM approach. $W_i^{(S)}$'s and $W_i^{(M)}$'s are the *N*-dimensional vectors of weights computed through SVM and MRMR approach respectively. G_i's and R_i's are

the *N*-dimensional vectors of gene lists and corresponding gene rank scores. SVM, MRMR stand for Maximum Relevance and Minimum Redundancy and Support Vector Machine algorithms. *p_i-value* is statistical significance value for *i-th* gene. α is the desired level of statistical significance.

2.4 Comparative performance analysis of the proposed approach

The comparative performance analysis of the proposed BSM approach with respect to 9 competitive gene selection methods (Supplementary Document S5) was carried out on 6 different rice GE datasets. For this purpose, different gene sets (***G***) of various sizes given in Supplementary Table S10 were selected through the 10 gene selection methods including proposed BSM approach. Then, these gene sets were validated with respect to subject classification, QTL testing, and GO analysis.

2.4.1 Performance analysis with subject classification

Under this comparison setting, the performance of the gene selection methods (Supplementary Document S5) including the proposed approach were assessed in terms of subject classification using mean CA, and Standard Error (SE) in CA computed through a varying window size technique. In other words, genes in ***G*** were validated with their ability to discriminate the class labels of subjects/samples between case (+1), and control (-1). Further, the gene set selected through a method which provides maximum discrimination between the subjects of two groups (*i.e.* case *vs*. control) through CA will be considered as highly relevant gene sets. The expressions for mean CA, and SE in CA computed through varying window size technique are given in Eq. 25 and 26.

Let, *n* be the size of ***G***, *S* be the size of the windows (*i.e.,* size refers to number of ranked genes), and *L* be the sliding length. Then, the total number of windows becomes, $K = (n - S)/L$. The genes in ***G***, arranged in different windows along with their expression values, were then used in SVM classifiers with four basis-functions, *i.e.,* linear (SVM-LBF), radial (SVM-RBF), polynomial (SVM-PBF), and Sigmoidal (SVM-SBF) to compute CA over a five-fold cross validation. Let, CA_1, CA_2, …, CA_K be the CA's for each sliding windows, then the mean CA and SE in CA can be defined as:

$$\mu_{CA}^{G} = {(\sum_{k=1}^{K} CA_k)}\Big/_{K} \tag{25}$$

doi:10.20944/preprints202009.0699.v1

$$SE_{CA}^{G} = \sqrt{\sum_{k=1}^{K}(CA_k - \mu_{CA}^{G})^2 \Big/ K} \qquad (26)$$

Here, we took different combinations of *n*, *S,* and *L*, as given in Supplementary Table S10, to compute μ_{CA}^{G} and SE_{CA}^{G} for the comparative performance analysis of the gene selection methods (Supplementary Document S5). The higher value of μ_{CA}^{G}, and a lower value of SE_{CA}^{G} indicates the better performance of the gene selection method, and *vice-versa.*

2.4.2 Performance analysis with QTL testing

The comparative criteria based on subject classification are popularly used for assessing the performance of gene selection methods [7,8,28,35–37,39]. However, these criteria fails to tell the biological relevancy of the genes selected through the gene selection methods [38]. Hence, under this comparative setting we assessed the performance of the proposed and existing methods through their ability to select genes which are associated with QTL regions. For this purpose, the criteria given in Eq. 27 and 28 are developed.

$$Qstat = \sum_{t=1}^{|Q|} \sum_{i=1}^{n} I_{q_t}(\,g_i) \qquad (27)$$

where, ***G***: gene set selected by a method, *Qstat: rv* whose values represent the number of genes covered by QTLs, *Q*: set of associated QTLs, and the indicator function present in Eq. 27 is represented in Eq. 28.

$$I_{q_t}(\,g_i) = \begin{cases} 1 & if\ g_i^c[a,] \geq q_t^c[d,]\ and\ g_i^c[,b\,] \leq q_t^c[,e\,] \\ 0 & else \end{cases} \qquad (28)$$

where, $g_i^c\,[a, b] \in$ ***G*** (*a* and *b* represents start and stop positions in terms of bp of the gene g_i on chromosome *c*) and $q_t^c[d, e] \in Q$ (*d* and *e* represents the start and stop positions of the QTL q_t on chromosome c).

Here, the $Qstat$ *rv* follows a hyper-geometric distribution and its distribution function is given in Eq. 29.

$$\mathrm{P}[Qstat = v] = 1 - \binom{M}{v}\binom{N-V}{n-v}\Big/\binom{N}{n} \qquad (29)$$

where, *V*: total number of genes covered by the QTLs in the whole GE data and *v*: number of genes in ***G*** that are covered by QTLs. Further, using the Eq. 29 the statistical significance value (*p-value*) associated with the ***G*** can be computed. In other words, this *p-value* reveals the enrichment significance of ***G*** with stress specific QTLs. Here, the higher values of $Qstat$ and $-log_{10}(p - value)$ indicates the better performance of the gene selection method, and *vice-versa*.

2.4.3 Performance analysis with GO enrichment

GO analysis involves with annotation of gene functions under three taxonomic categories, *i.e.* Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) [41]. This analysis helps in evaluating the functional similarities among the genes in ***G*** [49], as there exists a direct relationship between semantic similarity of gene pairs with their structural (sequence) similarity [50,51]. Under this comparison setting, we assessed the performance of 10 gene selection methods including the proposed method using GO based biologically relevant criterion. In other words, first different gene sets are selected through these methods, then GO based criterion is computed for each selected gene set. For this purpose, we developed a GO based semantic distance measure to assess the GO based biologically relevancy of ***G*** selected thorough the proposed and existing gene selection methods. Then the GO based semantic distance measure (d_{ij}) between i^{th} and j^{th} genes can be expressed in Eq. 30, as:

$$d_{ij\,(i\neq j)}^{GO} = \frac{|GO_i \Delta GO_j|}{|GO_i \cup GO_j|} \qquad \forall\, i,j = 1, 2, \dots, n \tag{30}$$

where, $GO_i = \{go_{i1}, go_{i2}, \dots, go_{iI}\}$ and $GO_j = \{go_{j1}, go_{j2}, \dots, go_{jJ}\}$ be the two sets of GO terms that annotate i^{th} and j^{th} genes in ***G*** respectively. Further, the GO based average biologically relevant score for ***G*** (for a gene selection method) can be developed based on Eq. 30 and is shown in Eq. 31.

$$D_G^{avg} = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} d_{ij}^{GO} \tag{31}$$

where, D_G^{avg} in Eq. 31 represents is the average biologically relevant score for ***G*** based on GO annotations. Using Eq. 31, the D_G^{avg} scores under MF, BP, and CC taxonomies were computed for each of the gene sets selected through different methods. A lower value of D_G^{avg} indicates better performance of the gene selection method and *vice-versa*.

3. Results

3.1 Computation of gene selection criteria through proposed approach

The distributions of weights computed from SVM-MRMR method [28] and *adj. p-values* for genes computed from the proposed BSM approach for abiotic and biotic stresses in rice are shown in Figures 2 and S3 respectively. The distributions of SVM-MRMR weights of genes for salinity stress data contained values, which are not so clearly separated (*i.e.,* higher values from lower values) (Figure 2A). In other words, the genes relevant to salinity stress condition was not well visualized from Figure 2A. However, from the distribution of *adj. p-values* computed through the proposed approach, it was observed that the relevant genes are found to be well separated from the irrelevant genes, and a small number of genes found to be statistically significant (*i.e.,* relevant to salinity stress) (Figure 2A1). In other words, for salinity stress data, the separation between relevant, and irrelevant genes can be well visualized from Figure 2A1 as compared to Figure 2A. Similar interpretations can be observed for other stress datasets, *viz*. cold, drought, bacterial, fungal, and insect (Figures 2, S3). Hence, the comparative graphical analysis showed a clear distinction between relevant, and irrelevant genes through the proposed BSM approach as compared to the existing SVM-MRMR approach. Further, the proposed approach offered a more statistically meaningful, and easily biologically interpretable values, *i.e., adj. p-values*, to measure gene relevancy, and subsequently selecting the genes in high-dimensional GE data compared to rank-based methods. The relevant genes selection using *adj. p-values* computed through the NP test statistic is more statistically sound as it is independent from the distribution of GE data, and corrected over multiple hypothesis testing. This comparative graphical analysis showed the improvement of BSM approach over SVM-MRMR method (Figures 2, S2), at least in terms of visualization. Moreover, such claim was validated through the performance analysis of the proposed BSM approach with respect to 9 existing methods based on statistically necessary criteria, *i.e.* subject classification, and biologically relevant criteria based on QTL and GO terms on multiple rice GE datasets.

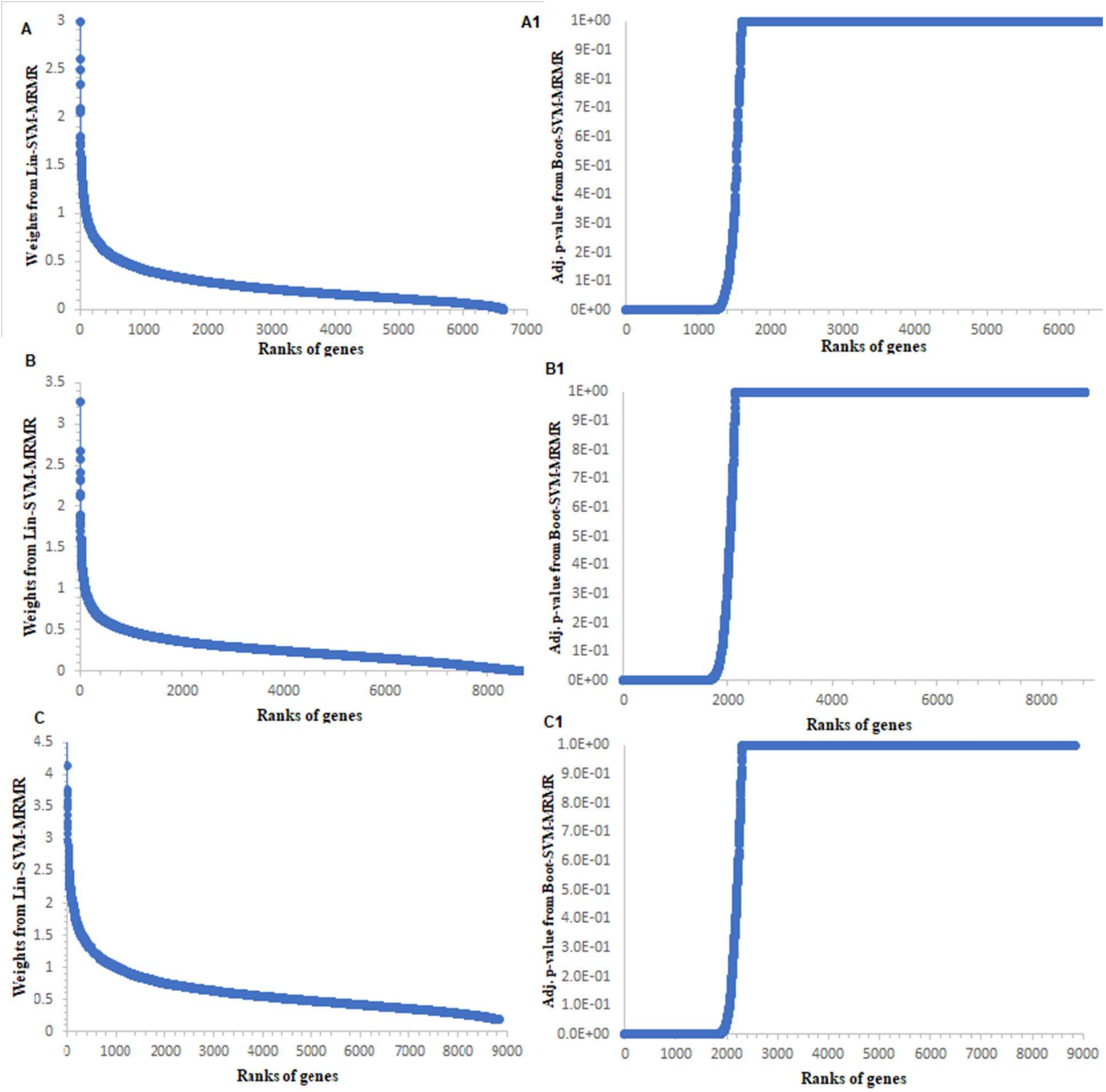

Figure 2. Graphical analysis of the proposed BSM approach with SVM-MRMR approach. (A) Distribution of gene weights computed from SVM-MRMR approach for the abiotic stresses. The distributions are shown for (A1) Salinity; (A2) Cold; and (A3) Drought stress datasets in rice. (B) Distribution of adj. *p-values* computed from proposed BSM approach for the abiotic stresses. The distributions are shown for (B1) Salinity; (B2) Cold; and (B3) Drought stress datasets in rice.

3.2 Comparative performance analysis based on subject classification

In this study, we used μ_{CA}^{G} and SE_{CA}^{G} computed through the varying sliding window size technique as statistically necessary criteria for performance analysis of the proposed BSM approach on 6 different GE datasets. Here, these measures were computed over five-fold cross validations through training the SVM-LBF, SVM-PBF, SVM-RBF, and SVM-SBF classifiers. The results are shown in Figures 3 and S4. The values of CA and SE in CA are also given in Supplementary Document S6. For cold stress data in rice, the μ_{CA}^{G} computed through SVM-LBF classifier for the proposed BSM approach was observed to be higher than other gene selection methods followed by SVM-RFE and SVM-MRMR over all selected gene sets (Figure 3, Document S6). This indicated the better performance of the BSM approach in terms of its ability to classify the subjects/samples through selecting relevant genes from cold stress GE data. Further, the values of SE_{CA}^{G} from SVM-LBF classifier for the BSM approach was found to be much lesser (over all the gene sets) than that of 9 existing gene selection methods considered in this study (Supplementary Document S6). This shows that the genes selected through the proposed BSM approach is much more relevant (informative), and robust than other methods. For cold stress data, similar interpretations can be made for SVM-PBF, SVM-RBF, and SVM-SBF classifiers.

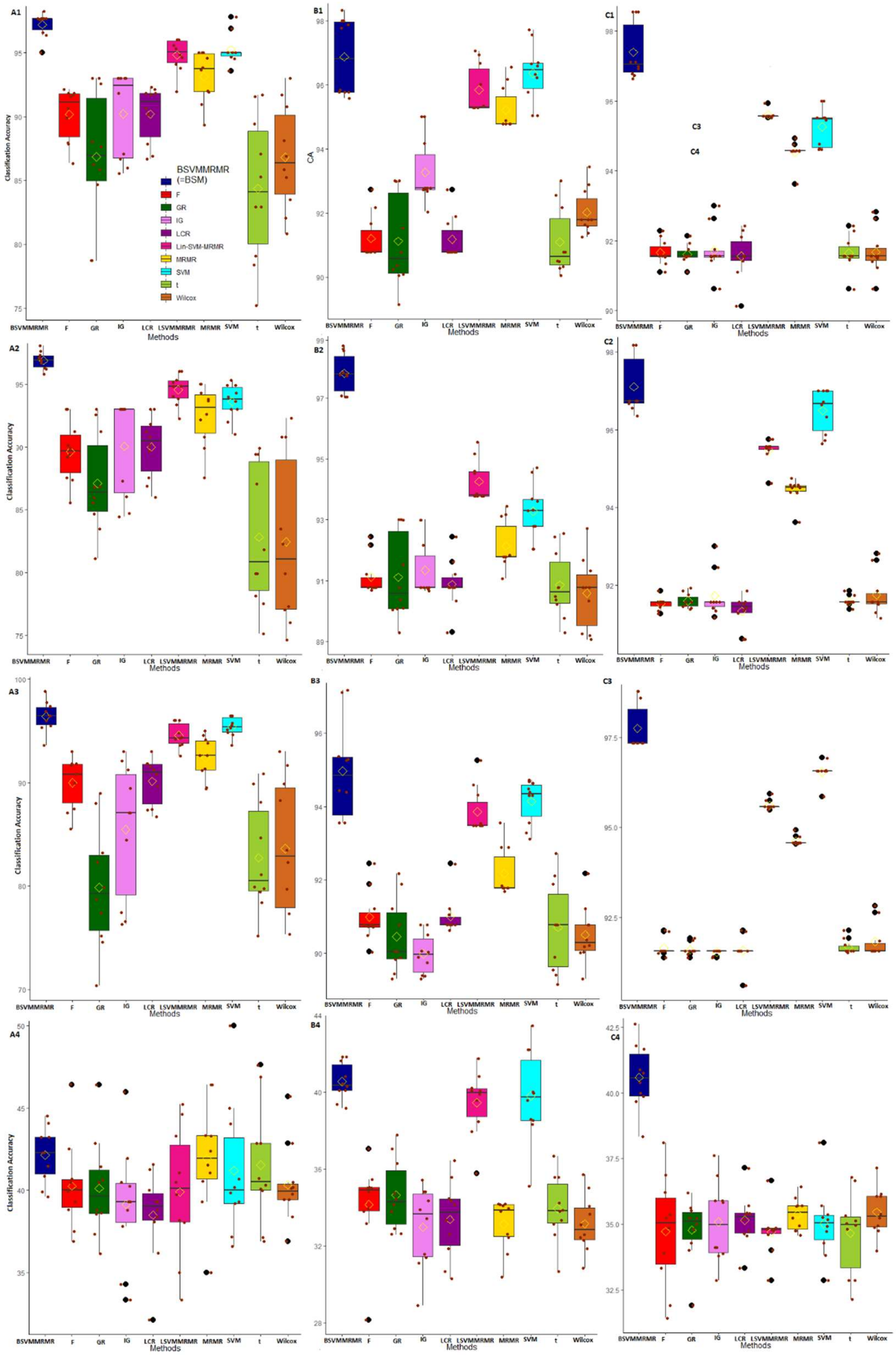

Figure 3. Classification based comparative performance analysis of gene selection methods for abiotic stresses in rice. The horizontal axis represents the gene selection methods. The vertical axis represents post selection

classification accuracy obtained by using varying sliding window size technique. The classification accuracies over the window sizes are presented as boxes. The distributions of classification accuracy are shown for (A1-A4) Cold stress with SVM-LBF (A1), SVM-PBF (A2) SVM-RBF (A3), and SVM-SBF (A4) classifiers; (B1-B4) Salt stress with SVM-LBF (B1), SVM-PBF (B2) SVM-RBF (B3), and SVM-SBF (B4) classifiers; (C1-C4) Drought stress with SVM-LBF (C1), SVM-PBF (C2), SVM-RBF (C3), and SVM-SBF (C4) classifiers.

For salinity stress data, the μ_{CA}^{G} (except gene sets of sizes 500, 1000 and 1500) computed for the proposed BSM approach through SVM-LBF classifier were found to be higher than other methods followed by SVM-RFE, and SVM-MRMR (Figure 3, Document S6). This indicated the proposed approach was quite better, and competitive with popular methods, *i.e.* SVM-RFE, SVM-MRMR. However, for SVM-PBF classifier, the μ_{CA}^{G} over all gene sets for BSM approach was higher than all other considered gene selection methods followed by SVM-RFE (Figure 3, Document S6). Furthermore, the SE_{CA}^{G} computed through SVM-LBF, and SVM-PBF classifiers for the BSM approach was found to be least followed by SVM-RFE (Document S6), indicating its better performance in terms of selection of robust, and relevant gene sets. Similar interpretation can be made for other classifiers, *i.e.,* SVM-RBF and SVM-SBF. It was observed that the μ_{CA}^{G} from SVM-SBF classifier was found to be least (with high SE_{CA}^{G}) among the four classifiers for all the datasets (Figures 3, S4, Document S6). Here, it is pertinent to note that the sigmoid basis function may not be recommended to use in SVM training for real crop GE datasets. Furthermore, similar interpretations can be made for other abiotic (*i.e.* drought) and biotic (*i.e.* bacterial, fungal and insect) stress GE datasets (Figures 3, S4 and Document 6).

3.3 Comparative performance analysis based on QTL testing

We used publicly available QTL data to statistically measure the biological relevancy of the genes selected through the proposed and existing gene selection method(s). The main rationale behind such analysis is that the genes selected for a stress condition (through a gene selection method) are expected to be overlapped with the that stress specific QTL regions. Therefore, the QTLs and genes selected through these 10 gene selection methods including the proposed BSM are mapped to the whole rice genome using MSU rice genome browser [52]. Further, the list of mapped QTLs for different abiotic (*i.e.* salinity, cold and drought), and biotic (*i.e.* bacterial, fungal and insect) stresses in rice along with their chromosomal positions in the genome are given in Supplementary Document S2.

The biological relevance of the selected genes through both proposed and existing gene selection methods were measured through two criteria, *i.e. Qstat* and -*log*10(*p-value*). The distributions of *Qstat* and -*log*10(*p-value*) over the selected genes for the 6 different datasets in rice are given in Figures 4 and 5, respectively. For salinity stress data, the values of *Qstat* over all the gene sets of sizes (< 1000) selected through the proposed BSM approach were found to be higher than that of SVM-MRMR, SVM-RFE, MRMR, IG, F, Wilcox, and PCR (Figure 4A). Further, it may be noted that the proposed approach was equally competitive with the univariate gene selection method *like* t-test, while they are assessed in terms of *Qstat* (Figure 4A). For higher gene set sizes (> 1000), the value of *Qstat* for Wilcox method was found to be higher than that of proposed, and existing approaches (Figure 4A) in the same data. This may be attributed to that the Wilcox method is non-parametric and does not dependent on the distribution of GE data. For cold stress data, the values of *Qsta*t statistic for all the selected gene sets through the BSM approach were higher than that of other existing methods (Figure 4B). This indicates that the performance of the proposed BSM approach is better in terms selecting cold stress related biologically relevant genes that are mostly overlapped with cold stress QTL regions in rice. Similar interpretations can be made for other abiotic (drought), and biotic (bacterial, fungal and insect) stress datasets in rice (Figure 4). Here, it is worthy to note that the *Qstat* is a linear function of the number of genes selected (through a gene selection approach), number of QTLs reported for that stress, and length of the QTL regions (Figure 4).

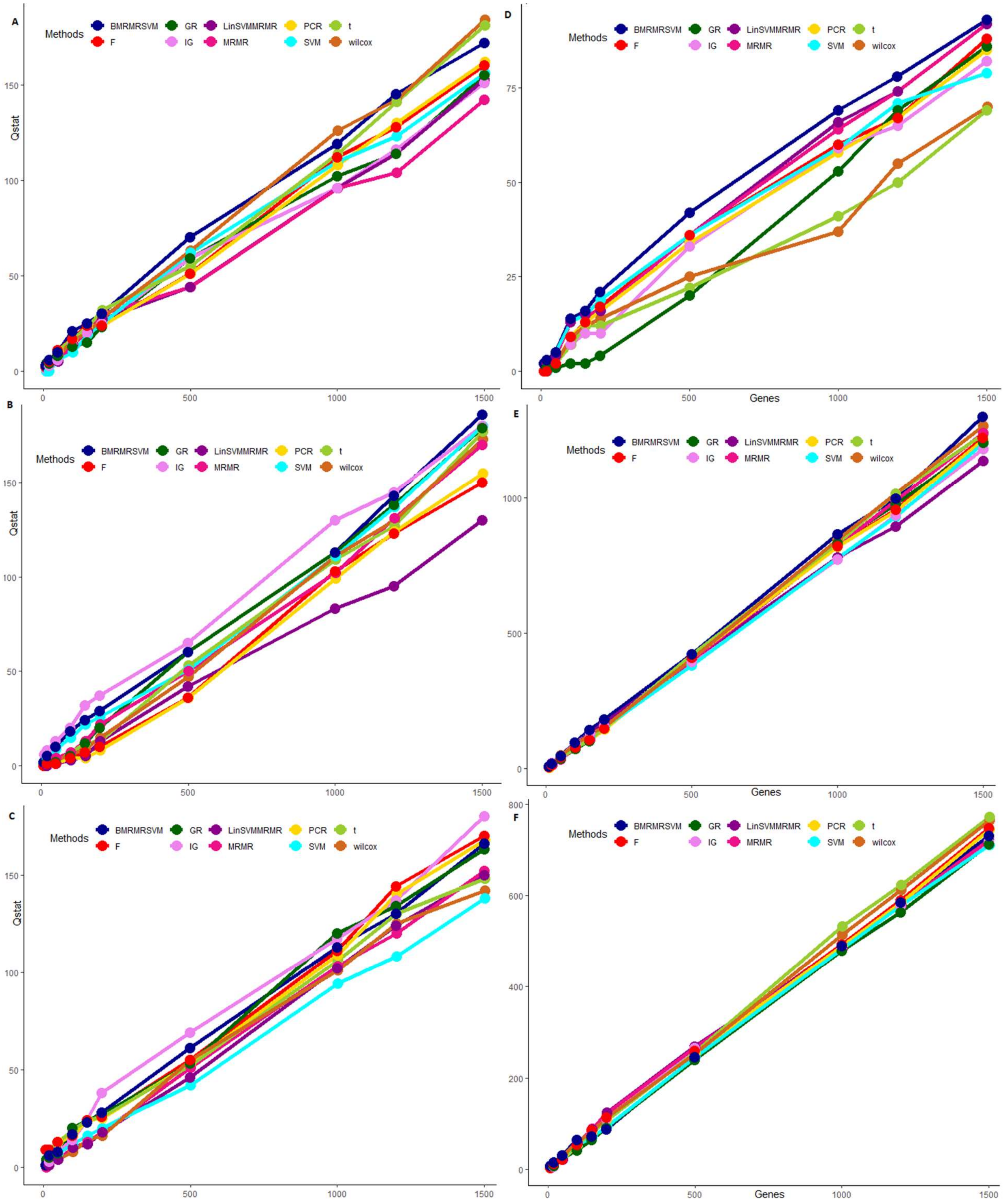

Figure 4. Comparative performance analysis of gene selection methods through distribution of *Qstat* statistic. The horizontal axis represents the gene sets obtained by each of the ten gene selection methods. The vertical

axis represents the value of *Qstat* statistic. The distribution of *Qstat* statistic are shown for (A) Salinity; (B) Cold; (C) Drought; (D) Bacterial; (E) Fungal, and (F) Insect stress datasets in rice.

Further, for salinity stress data, the -log10(*p-value*) from hypergeometric test over all the selected gene sets for the proposed BSM approach was observed to be higher than other existing gene selection methods (except t and GR) (Figure 5). In other words, genes selected by the BSM approach were more enriched with the underlying salinity responsive QTLs as compared to other existing methods. Similar interpretations can be made for other abiotic (*i.e.,* cold and drought), and biotic (*i.e.,* bacterial, fungal and insect) stress datasets in rice (Figure 5). Moreover, it is interesting to note that the values of *Qstat* and -*log*10(*p-value*) for (univariate) methods, such as t, F, PCR, Wilcox, IG, and GR were found to be higher than that of the existing (multivariate) methods, such as MRMR, SVM-MRMR (Figures 4, 5). This indicates the better, and equally competitive performance of univariate over multivariate methods of gene selection, when evaluated through QTL based biological relevancy criteria. Such observations are not expected in statistics, but well established in biology through previous studies [53].

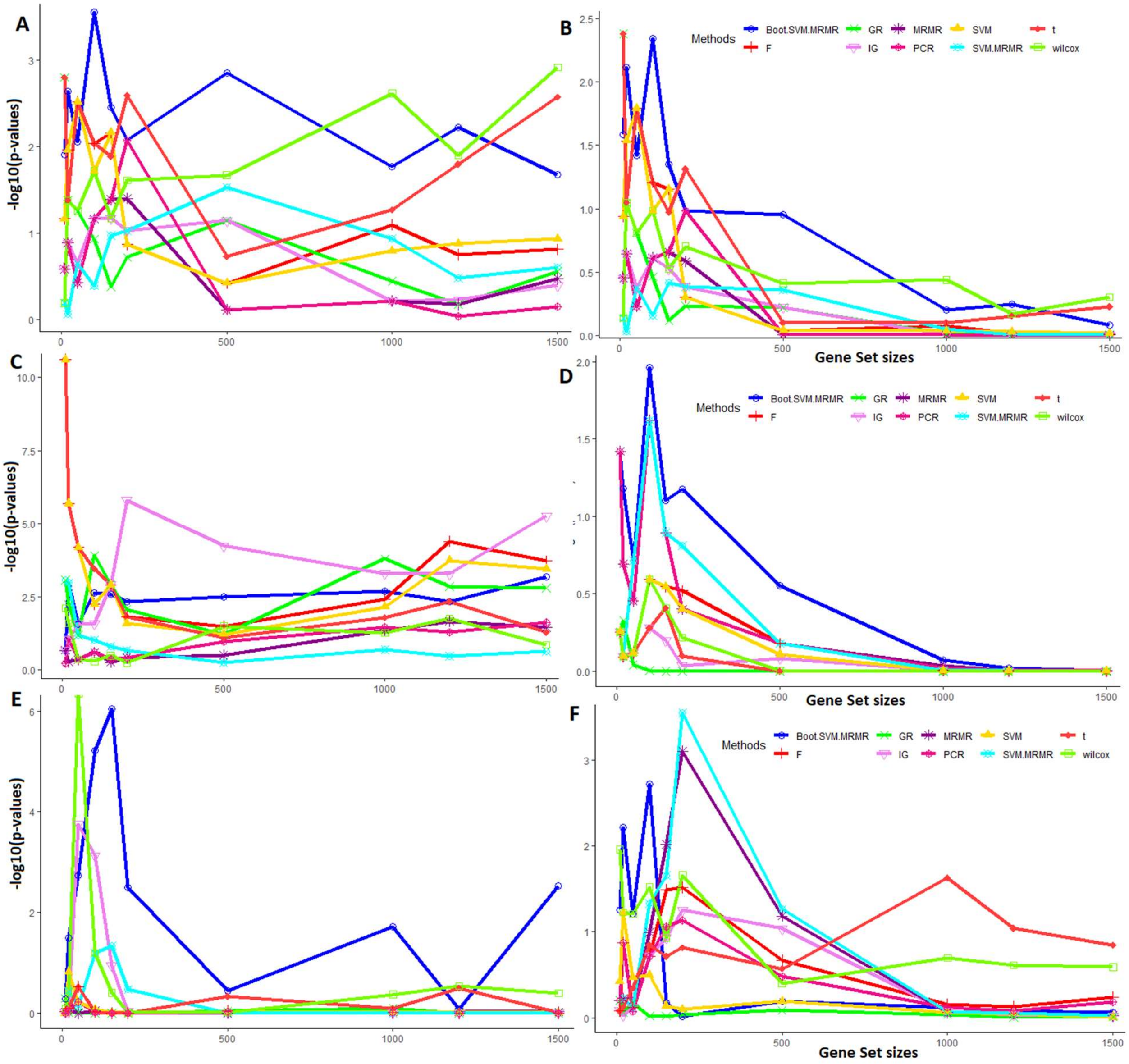

Figure 5. Comparative performance analysis of gene selection methods through distribution of *p-values* from QTL-hypergeometric test. The horizontal axis represents the gene sets obtained by each of the ten gene selection methods. The vertical axis represents the value of -log10(*p-value*) from QTL-hypergeometric test. The distribution of -log10(*p-value*) are shown for (A) Salinity; (B) Cold; (C) Drought; (D) Bacterial; (E) Fungal, and (F) Insect stress datasets in rice.

3.4 Comparative performance analysis based on GO analysis

The comparative performance analysis of the proposed BSM approach with 9 competitive gene selection methods (Document S5) was carried out through GO based distance analysis on 6 different rice GE datasets. Here, we set *n* (*i.e.* number of selected genes) as 10, 20, 50, 100, 150, 200 and 500. Then, the selected genes, through the 10 gene selection methods including the proposed BSM, were annotated with the GO terms under MF, BP, and CC categories using *AgriGo* database [43]. The results from this analysis for abiotic stresses under MF, BP, and CC GO categories are given in Tables 2-4 respectively and for biotic stresses in Supplementary Document S7. For salinity stress data, under MF category, the values of GO based average distance scores for the proposed BSM approach were found to be least than that of 9 existing methods over all the selected gene sets (Table 2). This indicated that the proposed approach selected more (molecular) functionally similar genes which are responsible salinity tolerance in rice. Similar results can be found for BP, and CC GO based distance analysis for the same stress data (Table 2). In other words, the proposed BSM approach selects more biologically relevant genes attributed to each GO category for salinity stress as compared to other 9 competitive methods (Table 2). For bacterial stress, the values of GO based average distance score under MF, BP, and CC GO categories for the proposed BSM approach were found to be least among other gene selection methods (Supplementary Document S7). Similar interpretations can be made for other abiotic (*i.e.,* cold and drought) and biotic (*i.e.,* fungal and insect) datasets in rice (Tables 2-4, Document S7). Through this analysis, it was found that the proposed BSM approach performed better in terms of selecting more functionally relevant genes, which conferred biotic and abiotic stresses tolerance in rice.

Table 2. Comparative Performance analysis of gene selection methods based on GO_MF based dissimilarity score for abiotic stresses in rice.

Methods	**MRMR**	**SVM**	**SVM-MRMR**	**IG**	**GR**	**Wilcox**	**t**	**PCR**	**F**	**BSM**
Salt stress in rice										
10	0.98	0.95	0.97	0.92	0.89	0.93	0.93	0.96	0.96	**0.88**
20	0.97	0.89	0.93	0.92	0.86	0.89	0.89	0.91	0.91	**0.86**
50	0.92	0.91	0.92	0.90	0.90	0.87	0.87	0.92	0.92	**0.85**
100	0.92	0.90	0.89	0.90	0.88	0.87	0.88	0.92	0.91	**0.83**
150	0.90	0.89	0.90	0.89	0.88	0.87	0.87	0.90	0.91	**0.83**
200	0.90	0.89	0.88	0.89	0.87	0.88	0.88	0.90	0.90	**0.84**
500	0.90	0.90	0.89	0.90	0.90	0.89	0.90	0.89	0.89	**0.83**
Cold stress in rice										
10	0.82	0.84	0.82	0.92	0.99	0.92	0.87	0.77	0.77	**0.75**
20	0.93	0.88	0.93	0.95	0.93	0.88	0.90	0.91	0.88	**0.71**
50	0.91	0.88	0.91	0.93	0.90	0.91	0.91	0.92	0.92	**0.73**
100	0.91	0.90	0.91	0.90	0.88	0.91	0.91	0.91	0.91	**0.74**
150	0.90	0.89	0.90	0.89	0.89	0.89	0.90	0.91	0.91	**0.72**
200	0.90	0.89	0.90	0.89	0.88	0.89	0.90	0.90	0.90	**0.73**
500	0.90	0.88	0.90	0.90	0.89	0.88	0.89	0.88	0.89	**0.73**
Drought stress in rice										
10	0.82	0.86	0.81	0.90	0.93	0.65	0.76	0.76	0.76	**0.71**
20	0.79	0.86	0.78	0.91	0.90	0.80	0.81	0.81	0.81	**0.75**
50	0.88	0.84	0.87	0.88	0.90	0.84	0.88	0.89	0.89	**0.75**
100	0.89	0.89	0.88	0.89	0.89	0.88	0.88	0.88	0.88	**0.76**
150	0.88	0.88	0.87	0.89	0.88	0.88	0.88	0.88	0.88	**0.76**
200	0.88	0.88	0.87	0.88	0.89	0.89	0.88	0.88	0.88	**0.74**
500	0.88	0.88	0.87	0.88	0.88	0.89	0.88	0.87	0.87	**0.73**

Values marked as bolds represent dissimilarity scores obtained from proposed BSM approach

Table 3. Comparative Performance analysis of gene selection methods based on GO_BP based dissimilarity score for abiotic stresses in rice.

Methods	**MRMR**	**SVM**	**SVM-MRMR**	**IG**	**GR**	**Wilcox**	**t**	**PCR**	**F**	**BSM**
Salt stress in rice										
10	0.86	0.94	0.86	0.92	0.97	0.90	0.90	0.88	0.88	**0.83**
20	0.90	0.91	0.90	0.89	0.91	0.92	0.92	0.84	0.85	**0.84**
50	0.89	0.90	0.88	0.88	0.90	0.88	0.89	0.88	0.88	**0.82**
100	0.88	0.89	0.86	0.89	0.89	0.85	0.86	0.89	0.87	**0.82**
150	0.87	0.89	0.90	0.88	0.89	0.85	0.85	0.89	0.89	**0.83**
200	0.87	0.89	0.86	0.88	0.89	0.84	0.85	0.89	0.88	**0.82**
500	0.87	0.89	0.87		0.89	0.86	0.86	0.86	0.86	**0.82**
Cold stress in rice										
10	0.79	0.82	0.79	0.86	0.94	0.91	0.90	0.79	0.79	**0.79**
20	0.93	0.89	0.93	0.90	0.88	0.86	0.88	0.90	0.86	**0.82**
50	0.88	0.89	0.88	0.90	0.88	0.88	0.87	0.89	0.90	**0.71**
100	0.88	0.89	0.88	0.89	0.87	0.90	0.88	0.89	0.89	**0.74**

150	0.89	0.88	0.89	0.88	0.88	0.88	0.87	0.88	0.88	**0.73**
200	0.89	0.87	0.89	0.87	0.87	0.87	0.87	0.88	0.84	**0.73**
500	0.88	0.86	0.88	0.86	0.86	0.84	0.86	0.87	0.83	**0.71**
Drought stress in rice										
10	0.86	0.79	0.85	0.81	0.89	0.62	0.83	0.83	0.83	**0.73**
20	0.84	0.79	0.83	0.89	0.90	0.80	0.84	0.84	0.84	**0.72**
50	0.88	0.81	0.87	0.88	0.88	0.81	0.88	0.88	0.88	**0.72**
100	0.87	0.84	0.86	0.88	0.88	0.84	0.86	0.87	0.87	**0.72**
150	0.86	0.84	0.85	0.88	0.88	0.84	0.87	0.87	0.87	**0.71**
200	0.86	0.84	0.85	0.87	0.87	0.85	0.86	0.86	0.86	**0.72**
500	0.87	0.85	0.86	0.86	0.87	0.87	0.86	0.85	0.83	**0.72**

Values marked as bolds represent dissimilarity scores obtained from proposed BSM approach

Table 4. Comparative Performance analysis of gene selection methods based on GO_CC based dissimilarity score for abiotic stresses in rice.

	MRMR	**SVM**	**SVM-MRMR**	**IG**	**GR**	**Wilcox**	**t**	**PCR**	**F**	**BSM**
Salt stress in rice										
10	0.77	0.71	0.70	0.94	0.97	0.93	0.93	0.95	0.95	**0.78**
20	0.88	0.87	0.85	0.92	0.90	0.91	0.91	0.88	0.88	**0.81**
50	0.88	0.89	0.86	0.92	0.92	0.90	0.90	0.89	0.89	**0.84**
100	0.88	0.90	0.8	0.91	0.89	0.86	0.86	0.88	0.88	**0.83**
150	0.87	0.90	0.87	0.90	0.89	0.86	0.87	0.88	0.88	**0.83**
200	0.87	0.89	0.85	0.90	0.90	0.88	0.89	0.88	0.88	**0.83**
500	0.88	0.90	0.88	0.89	0.90	0.88	0.89	0.87	0.87	**0.82**
Cold stress in rice										
10	0.78	0.80	0.78	0.96	0.81	0.87	0.86	0.70	0.70	**0.70**
20	0.88	0.86	0.88	0.96	0.87	0.87	0.89	0.81	0.83	**0.71**
50	0.86	0.89	0.86	0.90	0.85	0.84	0.85	0.89	0.90	**0.73**
100	0.88	0.90	0.88	0.90	0.81	0.83	0.84	0.87	0.87	**0.74**
150	0.88	0.89	0.88	0.90	0.82	0.82	0.86	0.87	0.88	**0.74**
200	0.87	0.90	0.87	0.90	0.84	0.85	0.86	0.87	0.85	**0.73**
500	0.88	0.89	0.88	0.89	0.86	0.97	0.86	0.88	0.87	**0.73**
Drought stress in rice										
10	0.82	0.86	0.81	0.91	0.89	0.83	0.87	0.87	0.87	**0.74**
20	0.89	0.85	0.88	0.93	0.90	0.87	0.89	0.89	0.89	**0.74**
50	0.86	0.88	0.85	0.91	0.87	0.87	0.88	0.88	0.88	**0.73**
100	0.87	0.87	0.86	0.89	0.86	0.87	0.88	0.88	0.88	**0.74**
150	0.87	0.87	0.86	0.90	0.85	0.85	0.87	0.87	0.87	**0.74**
200	0.87	0.87	0.86	0.89	0.86	0.86	0.87	0.87	0.87	**0.73**
500	0.87	0.86	0.86	0.89	0.87	0.88	0.87	0.86	0.85	**0.72**

Values marked as bolds represent dissimilarity scores obtained from proposed BSM approach

4. Developed R Software package

To popularize the use of the proposed gene selection approach among the users, we developed an R software package which includes BSM R package, and accompanying documentation with examples. This package

doi:10.20944/preprints202009.0699.v1

is supplied with the manuscript as supplementary information and also available in https://github.com/sam-uofl/BSM. The guidelines for the use of BSM R package is given in Supplementary Document S8. This software is capable of computing weights for gene selection through MRMR and SVM, SVM-MRMR, and also provide functions for computing *p-values,* and adjusted *p-values* through BSM approach for different parameter options. Further, it also allowed different functions for selecting relevant gene sets through existing MRMR, SVM, SVM-MRMR, and proposed BSM gene selection approaches.

5. Discussion

In GE data analysis, the main aim is to select relevant genes or gene sets which are highly informative for particular condition/trait, which can be further used as predictors for diagnosing a disease [7,8,28,39] or to understand the stress/disease response mechanism in plants [6,10]. Further, the selected genes can also be used as predictors to develop statistical/classification models to handle high-dimensionality in GE data analysis, which enhances the stability, power and feasibility of the developed models [37]. Therefore, we proposed an improved statistical approach, *i.e.* BSM, for selection of genes from GE data, which is both effective in reducing redundancy among the genes and improves biological relevancy of genes with the target trait. In this approach, genes are selected based on the assessment of the statistical significance of the undertaken self-contained null hypothesis under a sound computational framework, which is more appealing as compared to the gene ranking methods, *e.g.* SVM-MRMR, MRMR, SVM-RFE. Moreover, thousand(s) of null hypotheses (*i.e.* equal to the number of genes) are usually tested simultaneously in GE data analysis which increased the chance of selection of false positives. Hence, we adjusted for multiple testing to compute the adjusted statistical significance values in the proposed BSM approach. In other words, an adjusted *p-value* was assigned to each gene through NP test statistic with multiple adjustments, and relevant genes were selected based on the adjusted *p-values*. These metrics (values between 0 and 1) represent more scientifically well defined, and statistically calculated biologically interpretable values to genome researchers, and experimental biologists as compared to gene ranks/weights. In BSM approach, a significant *p-value* gives confidence that the given gene is relevant for the target condition/trait. The BSM approach computes the statistical significance values through NP statistic(s) which does not depend on the distribution of the GE data unlike t-test. Further, the bootstrap procedure in the proposed approach can minimize the redundancy among genes as well as reduce the spurious association of genes with traits during gene selection.

The proposed BSM approach more clearly separates the relevant genes from the irrelevant genes as compared to the existing gene selection methods. Further, the mean CA and SE in CA computed using varying sliding window size technique through training the classifiers provided a proper platform for comparative performance analysis of the proposed approach. These performance metrics can be considered as statistically necessary to check whether the selected genes are informative, and robust. Through such analysis, it was found that the BSM approach performed better in terms of selecting informative, and robust genes from the high-dimensional GE data as compared to other competitive methods such as SVM-RFE, MRMR, SVM-MRMR and the information theoretic measures. The reason may be attributed to the inclusion of bootstrap based subject sampling model along with the self-contained statistical tests, which reduces the spurious association of genes with the target trait as well as with other genes. Further, the performance of BSM approach, in terms of the ability to classify the GE samples, found to be better as compared to multivariate approaches, *i.e.* MRMR, SVM-MRMR, univariate approaches, *i.e.* t-test, F-score, Wilcox, and informative theoretic measures, *i.e.* IG and GR. Here, it is worthy to note that the multivariate approaches performed better as compared to the univariate approaches when assessed under classification-based criteria, as the former considers gene-gene associations.

Adjudging the performance of gene selection methods based on only classification, might lead to the selection of biologically irrelevant genes. In other words, the comparative performance analysis of gene selection methods through classification-based metrics may be statistically necessary but may not be biologically sufficient. Therefore, we used criteria based on genetically rich information such as QTLs and GO to test the biological relevancy of the selected genes through proposed, and existing gene selection methods. Further, through this performance analysis, it was found that BSM approach selects more biological relevant genes measured in terms of overlapping of the selected genes with given QTL regions as compared to multivariate approaches, *i.e.* MRMR, SVM-MRMR and machine learning approach *like*

SVM-RFE. Moreover, the proposed BSM approach was equally competitive (and better) with univariate approaches, *i.e.* t-test, F-score, Wilcox, and PCR, and information theoretic measures, *i.e.* IG and GR, when QTL based criteria are considered. Through the QTLs-hypergeometric test analysis, it was evident that genes selected through the proposed BSM approach were more enriched with the QTL regions. In other words, genes selected through the BSM are more statistically enriched with the QTLs as compared to other competitive methods. The GO based distance analysis showed that higher functional similarities (which may have biological functions important to stress tolerance) exist among the genes selected by the BSM, as compared to methods, such as MRMR, SVM-RFE, SVM-MRMR, information theoretic measures. The performance of the BSM was found to be better and equally competitive with the univariate approaches, *viz.* t-score, F-score, Wilcox, and correlation-based approach in terms of selecting genes which are biologically relevant (in terms of GO annotations) for the target trait/condition.

Here, it is worthy to note that, when statistical necessary criteria were considered, the multivariate approaches performed better than univariate approaches, which is obvious from statistical standpoint as the former considers inter-variable relationships. When the biological relevant criteria are considered, the univariate approaches of gene selection performed equal, and even better as compared to their multivariate counterparts. Such observation was well established in GE genomics [53], which shows the real biological complexity for assessing the performance of gene selection approaches on real data. One way to handle such situation is to perform comparative performance analysis of gene selection methods under a framework of statistical necessary, and biological relevant criteria. Hence, we used such framework to comprehensively evaluate the performance of the BSM approach. The comparative performance analysis of the proposed approach with 9 existing competitive methods on multiple rice GE datasets revealed that the BSM approach is better under classification-based criteria. Further, the proposed approach is better as well as competitive with univariate approaches under QTL and GO based criteria. In other words, the proposed approach has the features of an ideal technique of gene selection, as it performed better under both statistically necessary, and biologically relevant criteria.

Here, we have undertaken a systematic and rigorous study to evaluate the performance of gene selection methods under statistically necessary and biologically relevant criteria on multiple real crop GE datasets. This study may provide a framework to comparatively study all the available gene selection methods under multi-criteria setup, which will guide genome researchers and experimental biologists to select the best method(s) objectively. The proposed approach will also provide a statistical template for combing other filter and wrapper gene selection methods under a sound, and effective computational environment.

Supplementary information

R software Package. BSVMRMR R package with real data example and required documentations (.gz).

Figure S1. Correlation plot of the gene expression experimental samples under salinity stress for rice.

Figure S2. Choosing the optimal value of beta for quadratic integration of SVM and MRMR scores. The horizontal axis represents the value of beta. The vertical axis represents the classification accuracy computed over five-fold cross validation through Support Vector Machine-Linear Basis Function classifier. The line plots are shown for selected gene sets with size (A) 10, (B) 20, (C) 50, (D) 100, (E) 150 and (F) 200.

Figure S3. Graphical analysis of the proposed BSM approach with SVM-MRMR approach. (A) Distribution of gene weights computed from SVM-MRMR approach for the biotic stresses. The distributions are shown for (A1): Bacterial; (A2): Fungal; and (A3): Insect stress datasets in rice. (B) Distribution of gene *p-values* computed from BSM approach for the biotic stresses. The distributions are shown for (B1): Bacterial; (A2): Fungal; and (A3): Insect stress datasets in rice.

Figure S4. Classification based comparative performance analysis of gene selection methods in biotic stresses in rice. (A) The horizontal axis represents the gene selection methods. The vertical axis represents post selection classification accuracy obtained by using varying sliding window size technique. The classification accuracies over the window sizes are presented as boxes. The distributions of classification accuracy are shown for (A1-A4) Bacterial stress with SVM-LBF (A1), SVM-PBF (A2) SVM-RBF (A3), and SVM-SBF (A4) classifiers; (B1-B4) Fungal stress with SVM-LBF (B1), SVM-PBF (B2) SVM-RBF (B3), and SVM-SBF (B4) classifiers; (C1-C4) Insect stress with SVM-LBF (C1), SVM-PBF (C2), SVM-RBF (C3), and SVM-SBF (C4) classifiers.

doi:10.20944/preprints202009.0699.v1

Document S1. Data collection, pre-processing, meta-analysis and preliminary gene selection to prepare the data for performance analysis of gene selection methods.

Document S2: Stress(s) specific Quantitative Trait Loci information for rice (*Oryza sativa L.*).

Document S3: Objective function of Support Vector Machine.

Document S4: Determination of 'beta' for quadratic integration.

Document S5. List of gene selection methods for performance analysis.

Document S6. Mean accuracy and standard error in classification computed through varying sliding windows size technique.

Document S7. Comparative performance analysis of gene selection methods based on GO based dissimilarity scores for biotic stresses in rice.

Table S1. Description about the gene expression studies used in this study (.xlxs).

Table S10. Different parameters (gene set size, window size and sliding lengths) combination for computation of classification accuracies.

Conflict of interest

The authors declare no conflicts of interest.

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 and R codes are freely available at http://github/sam-uofl/BSM.

Acknowledgement

This study was supported in part by National Institutes of Health (NIH), USA grants, P20GM113226, and 5P50AA024337 (SNR). Further, it was also supported by Netaji Subhas-ICAR International Fellowship, OM No. 18(02)/2016-EQR/Edn. (SD) of Indian Council of Agricultural Research (ICAR), New Delhi, India and partly by Wendell Cherry Chair (SNR) in Clinical Trial Research, University of Louisville, USA. The content is solely the responsibility of the authors and does not necessarily represent the official views of

NIH or ICAR. Authors duly acknowledge the support obtained from ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India.

Author Contributions

Conceived and designed the study: SD. Developed the methodologies: SD. Developed the R-code and R package: SD. Contributed materials: SD, SNR. Drafted the manuscript: SD. Corrected the manuscript: SD, SNR. Funding Acquisition: SNR.

References

1. Reuter JA, Spacek D V., Snyder MP. High-Throughput Sequencing Technologies. Mol Cell. 2015;58: 586–597. doi:10.1016/j.molcel.2015.05.004
2. Trevino V, Falciani F, Barrera-Saldaña HA. DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. Mol Med. 2007;13: 527–541. doi:10.2119/2006-00107.trevino
3. Charpe AM. DNA Microarray. Advances in Biotechnology. New Delhi: Springer India; 2014. pp. 71–104. doi:10.1007/978-81-322-1554-7_6
4. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41: D991–D995. doi:10.1093/nar/gks1193
5. Das S, Meher PK, Rai A, Bhar LM, Mandal BN. 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. 2017;12. doi:10.1371/journal.pone.0169605
6. Wang J, Chen L, Wang Y, Zhang J, Liang Y, Xu D. A Computational Systems Biology Study for Understanding Salt Tolerance Mechanism in Rice. Xu Y, editor. PLoS One. 2013;8: e64929. doi:10.1371/journal.pone.0064929
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science (80-). 1999;286: 531–537. doi:10.1126/science.286.5439.531
8. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector

machines. Mach Learn. 2002. doi:10.1023/A:1012487302797

9. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23: 2507–2517. doi:10.1093/bioinformatics/btm344

10. Liang Y, Zhang F, Wang J, Joshi T, Wang Y, Xu D. Prediction of Drought-Resistant Genes in Arabidopsis thaliana Using SVM-RFE. Zhu D, editor. PLoS One. 2011;6: e21750. doi:10.1371/journal.pone.0021750

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

12. Das S, Pandey P, Rai A, Mohapatra C. A computational system biology approach to construct gene regulatory networks for salinity response in rice (Oryza sativa). Indian J Agric Sci. 2015;85.

13. Inza I, Larrañaga P, Blanco R, Cerrolaza AJ. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med. 2004. doi:10.1016/j.artmed.2004.01.007

14. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinforma. 2012. doi:10.1109/TCBB.2012.33

15. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biology. 2003. doi:10.1186/gb-2003-4-4-210

16. Das S, Meher PK, Pradhan UK, Paul AK. Inferring gene regulatory networks using Kendall's tau correlation coefficient and identification of salinity stress responsive genes in rice. Curr Sci. 2017;112. doi:10.18520/cs/v112/i06/1257-1262

17. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Computational Systems Bioinformatics CSB2003 Proceedings of the 2003 IEEE Bioinformatics Conference CSB2003. IEEE Comput. Soc; 2003. pp. 523–528. doi:10.1109/CSB.2003.1227396

18. Chen YW, Lin CJ. Combining SVMs with various feature selection strategies. Stud Fuzziness Soft Comput. 2006. doi:10.1007/978-3-540-35488-8_13

19. Hossain A, Willan AR, Beyene J. An improved method on wilcoxon rank sum test for gene selection

from microarray experiments. Commun Stat Simul Comput. 2013. doi:10.1080/03610918.2012.667479

20. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics. 2002. doi:10.1093/bioinformatics/18.11.1454
21. Cheng T, Wang Y, Bryant SH. FSelector: a Ruby gem for feature selection. Bioinformatics. 2012;28: 2851–2852. doi:10.1093/bioinformatics/bts528
22. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007. doi:10.1093/bioinformatics/btm344
23. Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics. 2017;18: 9. doi:10.1186/s12859-016-1423-9
24. DING C, PENG H. MINIMUM REDUNDANCY FEATURE SELECTION FROM MICROARRAY GENE EXPRESSION DATA. J Bioinform Comput Biol. 2005;03: 185–205. doi:10.1142/S0219720005001004
25. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997. doi:10.1016/s0004-3702(97)00043-x
26. Guyon I. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn. 1998. doi:10.1109/5254.708428
27. Duan KB, Rajapakse JC, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience. 2005. doi:10.1109/TNB.2005.853657
28. Mundra PA, Rajapakse JC. SVM-RFE With MRMR Filter for Gene Selection. IEEE Trans Nanobioscience. 2010;9: 31–37. doi:10.1109/TNB.2009.2035284
29. Sun L, Kong X, Xu J, Xue Z, Zhai R, Zhang S. A Hybrid Gene Selection Method Based on ReliefF and Ant Colony Optimization Algorithm for Tumor Classification. Sci Rep. 2019.

doi:10.1038/s41598-019-45223-x

30. Mahi M, Baykan ÖK, Kodaz H. A new hybrid method based on Particle Swarm Optimization, Ant Colony Optimization and 3-Opt algorithms for Traveling Salesman Problem. Appl Soft Comput. 2015;30: 484–490. doi:10.1016/j.asoc.2015.01.068

31. Sohn I, Owzar K, George SL, Kim S, Jung SH. A permutation-based multiple testing method for time-course microarray experiments. BMC Bioinformatics. 2009. doi:10.1186/1471-2105-10-336

32. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47–e47. doi:10.1093/nar/gkv007

33. Knijnenburg TA, Wessels LFA, Reinders MJT, Shmulevich I. Fewer permutations, more accurate P-values. Bioinformatics. 2009. doi:10.1093/bioinformatics/btp211

34. Das S, Meher PK, Rai A, Bhar LM, Mandal BN. 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.). Tian Z, editor. PLoS One. 2017;12: e0169605. doi:10.1371/journal.pone.0169605

35. Das S, Rai A, Mishra DC, Rai SN. Statistical approach for selection of biologically informative genes. Gene. 2018;655. doi:10.1016/j.gene.2018.02.044

36. Lai C, Reinders MJT, van't Veer LJ, Wessels LFA. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics. 2006. doi:10.1186/1471-2105-7-235

37. Kursa MB. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics. 2014. doi:10.1186/1471-2105-15-8

38. Das S, Rai A, Mishra DC, Rai SN. Statistical Approach for Gene Set Analysis with Trait Specific Quantitative Trait Loci. Sci Rep. 2018;8: 2391. doi:10.1038/s41598-018-19736-w

39. Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans Pattern Anal Mach Intell. 2005.

doi:10.1109/TPAMI.2005.159

40. Tiwari S, SL K, Kumar V, Singh B, Rao A, Mithra SV A, et al. Mapping QTLs for Salt Tolerance in Rice (Oryza sativa L.) by Bulked Segregant Analysis of Recombinant Inbred Lines Using 50K SNP Chip. Yadav RS, editor. PLoS One. 2016;11: e0153610. doi:10.1371/journal.pone.0153610
41. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004. doi:10.1093/nar/gkh036
42. Ware D. Gramene: a resource for comparative grass genomics. Nucleic Acids Res. 2002. doi:10.1093/nar/30.1.103
43. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, et al. AgriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017. doi:10.1093/nar/gkx382
44. Sahani M, Linden J. Advances in neural information processing systems. … Processing Systems: Proceedings from the 2002 …. 2003. doi:http://dx.doi.org/10.1016/j.oraloncology.2016.02.011
45. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. An Introduction to the Bootstrap. Boston, MA: Springer US; 1993. doi:10.1007/978-1-4899-4541-9
46. Benjamini Y, Hochberg Y. Multiple Hypotheses Testing with Weights. Scand J Stat. 1997;24: 407–418. doi:10.1111/1467-9469.00072
47. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. Ann Appl Stat. 2011;5: 1752–1779. doi:10.1214/11-AOAS466
48. Chen S-Y, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. J Thorac Dis. 2017;9: 1725–1729. doi:10.21037/jtd.2017.05.34
49. Mazandu GK, Mulder NJ. Information content-based gene ontology functional similarity measures: Which one to use for a given biological data type? PLoS One. 2014. doi:10.1371/journal.pone.0113859
50. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. Bioinformatics. 2003. doi:10.1093/bioinformatics/btg153

51. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007. doi:10.1093/bioinformatics/btm087

52. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: Improvements and new features. Nucleic Acids Res. 2007. doi:10.1093/nar/gkl976

53. Glazko G V., Emmert-Streib F. Unite and conquer: Univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics. 2009. doi:10.1093/bioinformatics/btp406