Wu et al. BMC Bioinformatics (2016) 17:108
DOI 10.1186/s12859-016-0951-7

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Regularized logistic regression with network-based pairwise interaction for biomarker identification in breast cancer

Meng-Yun Wu[1,2†], Xiao-Fei Zhang[3†], Dao-Qing Dai[4*], Le Ou-Yang[5], Yuan Zhu[6] and Hong Yan[7]

Abstract

Backgroud: To facilitate advances in personalized medicine, it is important to detect predictive, stable and interpretable biomarkers related with different clinical characteristics. These clinical characteristics may be heterogeneous with respect to underlying interactions between genes. Usually, traditional methods just focus on detection of differentially expressed genes without taking the interactions between genes into account. Moreover, due to the typical low reproducibility of the selected biomarkers, it is difficult to give a clear biological interpretation for a specific disease. Therefore, it is necessary to design a robust biomarker identification method that can predict disease-associated interactions with high reproducibility.

Results: In this article, we propose a regularized logistic regression model. Different from previous methods which focus on individual genes or modules, our model takes gene pairs, which are connected in a protein-protein interaction network, into account. A line graph is constructed to represent the adjacencies between pairwise interactions. Based on this line graph, we incorporate the degree information in the model via an adaptive elastic net, which makes our model less dependent on the expression data. Experimental results on six publicly available breast cancer datasets show that our method can not only achieve competitive performance in classification, but also retain great stability in variable selection. Therefore, our model is able to identify the diagnostic and prognostic biomarkers in a more robust way. Moreover, most of the biomarkers discovered by our model have been verified in biochemical or biomedical researches.

Conclusions: The proposed method shows promise in the diagnosis of disease pathogenesis with different clinical characteristics. These advances lead to more accurate and stable biomarker discovery, which can monitor the functional changes that are perturbed by diseases. Based on these predictions, researchers may be able to provide suggestions for new therapeutic approaches.

Keywords: Protein-protein interaction network, Edge-biomarker discovery, Network-based pairwise interaction, Node degree, Adaptive elastic net

Background

Biomarker discovery for cancer based on multiple molecular data, such as gene or protein expression data, has become a major strategy in biomedical fields for personalized medicine. Diagnostic and prognostic biomarkers have the potential to provide deeper insights into disease pathogenesis [1]. Revealing the mechanisms of disease initiation and progression can be valuable for the selection of new therapeutic approaches and the prediction of later clinical benefit [2, 3].

With the increasingly accumulated "omics" (e.g. genomics, transcriptomics and proteomics) data generated from high-throughput technologies, extensive variable selection methods such as lasso [4] and elastic net [5] have been proposed to select relevant biomarkers for disease diagnosis or prognosis, where the genes or proteins are regarded as variables. These methods often

*Correspondence: stsddq@mail.sysu.edu.cn
†Equal contributors
[4]Intelligent Data Center and Department of Mathematics, Sun Yat-Sen University, Xingang West Road, 510275 Guangzhou, China
Full list of author information is available at the end of the article

focus on the variables that can discriminate patients in the training set with the outcome measured by categorical variables, such as normal and disease, and are able to predict well unseen patients [6]. It is a computational and statistical challenge to detect reliable and useful biomarkers for diseases with the relative small sample size and high dimensionality of most molecular data [7]. With a wealth of publicly molecular data for the same disease, it is common that the biomarker signatures from different studies have few overlaps [8]. This is partly because that many different biomarkers have similar discriminatory power. Another reason is the instability of the variable selection method towards the used samples. The low reproducibility of these signatures often result in the difficulty of achieving clear biological interpretation. On the other hand, the criterion that identifying differentially expressed genes often neglects non-differentially expressed genes which play a central role in the molecular mechanism of complex biological phenomena by interacting with other genes [9]. Complex diseases are multifactorial biological events manifested through changes in expressions of many individual genes and proteins, and mediated through complex interaction mechanisms [10–12]. This connectivity implies that the impact of a specific genetic abnormality is not restricted to the activity of the gene product that carries it, but can spread along the interactions and alter the activity of the connected gene products [13]. Therefore, changes of these interactions of which the involved genes or proteins are not differentially expressed may also result in the different states of a biological system.

In this situation, traditional variable selection methods (e.g. lasso and elastic net) which are based on the additive model are insufficient for predicting an outcome of interest. To overcome this limitation, a regression model with pairwise interactions between those variables is proposed by Bien et al. to select a subset of variables and interactions between variables that is predictive of the response [14]. However, their work is based on an assumption that an interaction is allowed into the model only if at least one of the corresponding variables is also in the model. This restriction makes it impossible to identify the interactions with the non-differentially expressed variables. Zhang et al. propose a new idea based on a new vector representation of an edge to identify edge-biomarkers which are the differentially correlated molecular pairs with optimal classification abilities [9]. In their work, the correlation of each molecular pair is depicted by two new coupled variables, which makes the dimension of the new space increase to $p(p-1)$, where p is the number of original variables. The computation is so time consuming that a preliminary variable screening is needed. In addition, the correlations used in these methods are only based on the expression data, which are not sufficient to reveal the physical interaction between two genes or proteins.

To deal with limitations in existing methods, a promising direction is to integrate expression profiles with rich biological knowledge in network datasets. A useful technique is to incorporate biological networks into variable selection process by a network-constrained regularization procedure, where the network is represented as a graph and its corresponding Laplacian matrix [15–17]. Another idea is to use these networks as Markov random field priors to guide the selection of relevant genes [18, 19]. The contribution of the network information used by these network-regularized approaches is to ensure smoothness of the coefficients on the network. They do not explicitly select the important interactions. Strategies to identify gene subnetwork or module have been proposed by integrating gene expression data and biological networks [3, 20–22]. Guo et al. treat the gene expressions within a functional module as an integrative data point based on the Gene Ontology (GO) enrichment analysis [22]. Some algorithms start from "seed" genes with highly differentially expressed in the network to generate the subnetwork-biomarkers, resulting in the neglect of the subnetworks with non-differentially expressed genes but different interactions [20, 21]. Moreover, Das et al. identify overlapping functional modules based on the topological properties of the protein interaction network by some clustering algorithms, and then use the elastic-net-based regression model to detect the differentially expressed functional modules [3]. The prognosis prediction results and the identified differentially modules may partly depend on the selected clustering algorithms. Another new perspective to reveal potential mechanisms altering the biological states is analyzing the comparison of biological networks across a set of conditions, and identifying the subnetwork-biomarkers which are differentially co-regulated [23, 24].

Motivated by the challenges posed by the instability and complex interactions in high-dimensional gene expression datasets, this study proposes a regularized logistic regression with network-based pairwise interaction via adaptive elastic net for biomarker identification (Fig. 1). The model embeds variable selection in the classifier construction process with the advantage that the differentially interactions can be used directly to predict the states of the new samples. Instead of identifying discriminative genes or functional modules with differential expression, where the modules often need to be determined in advance by some algorithms such as in [3, 22], we focus on the detection of gene pairs which exhibit different positive or negative interactions, thus the performance of the proposed method will not depend on the module detection algorithms. The results based on experimental characterization of mutant alleles in various

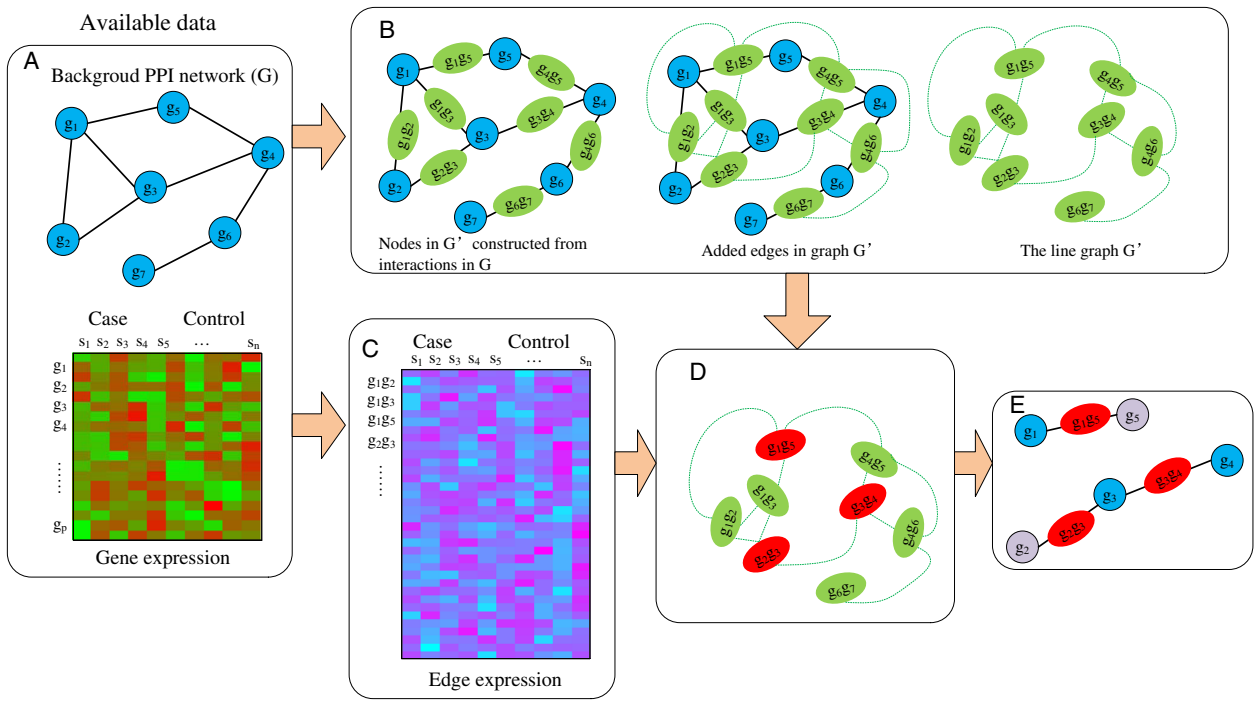

Fig. 1 Identification of an edge biomarker using the regularized logistic regression with network-based pairwise interaction via adaptive elastic net. **a** The available data, including expression data and PPI network data. **b** A line graph $G'(V', E', W')$ is constructed based on the similarities $S(e_{ik}, e_{jk})$ which are treated as the weights for the new edges of graph, where the elements of the new node set V' are the interactions in the biological network. **c** The expression of each gene is mapped to the edge of the network in different conditions by integrating both the network and expression information. **d** If the expression of the interaction change in the different states of a biological system, the edge is labeled in red, while green if not. **e** The informative edges identified by our model. The genes with differential expression is marked in purple, while blue if with non-differential expression

disorders show that the biochemical and physical interactions which are represented in models as edges are correlated with distinct structural properties of disease proteins and disease mechanisms [25, 26]. The model only considers the interactions belonging to a protein-protein interaction (PPI) network. The interactions based on both high-dimensional "omics" data and biological network can help to filter out the correlations between expression that have no underlying biological causality, which make the model have a low complexity [27]. The integrated information can lead to the improvement of both predictive accuracy and interpretability of the selection results [18]. In addition, different from the elastic-net-based regression model used by Das et al. in [3], our model includes an extension of the standard adaptive elastic net to consider the degree of proteins based on the line graph and the assumption that disease-genes tend to have high degrees in biological networks [28, 29]. This formulation leads to more stable biomarkers because the model is less reliant on the expression data. By applying the new model to six publicly available breast cancer datasets, we show that the algorithm is robust against the inclusion or exclusion of some patients on variable selection process at both gene and functional levels. Furthermore, our method can achieve competitive classification results with the state-of-the-art algorithms on detecting different responses to a certain survival time. The relevance of many identified biomarkers with breast cancer have been verified through biochemical or biomedical research. The Gene Ontology (GO) analysis further indicate the significant biological and functional correlations of the edge-biomarkers.

Methods

Regularized logistic regression with network-based pairwise interaction

Suppose that there are n independent p-dimensional observations, with binary response vector $y = (y_1, \cdots, y_n)^T$ and design matrix $X = (x_1, \cdots, x_n)^T$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $y_i \in \{0, 1\}$. Let $p(x_i)$ represents the class-conditional probability for observation i when $y_i = 1$ at particular parameters $\beta_0, \beta = (\beta_1, \cdots, \beta_p)^T$. In order to identify the biomarkers which have low discriminative power but play a central role in the molecular mechanism of complex biological phenomena by interacting with other genes, we define $p(x_i)$ through network-based pairwise interactions between variables as follows

$$p(x_i) = Pr(y_i = 1|x_i) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j\sim k} \beta_{jk} x_{ij} x_{ik})}}, \quad (1)$$

where $\sum_{j\sim k}$ denotes the sum over all unordered pairs $\{j, k\}$ for which j and k are adjacent on the certain undirected biological network. The pairs $\{j, k\}$ and $\{k, j\}$ are regarded as the same pair and will be treated in the model only once. The proposed regularized logistic regression model maximizes the penalized log-likelihood

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log p(x_i) + (1 - y_i)\log(1 - p(x_i))\right] - \lambda P_\alpha(\beta), \quad (2)$$

where $P_\alpha(\beta)$ is a penalty function which can shrink some components of β to zero for some appropriately chosen λ and α [30]. Compared with the two-way interaction model in [14] which includes $p + p(p-1)/2$ variables, the proposed model only takes account into the pairwise interactions or edges in the biological network, whose number is much less than $p(p-1)/2$. This form not only makes the model have an advantage in complexity, but also combines biological network information with genomics or proteomics datasets. The integrated information can reduce the impact of the noise from gene expression data on the process of estimating the biomarkers. Thus, the model will be less sensitive to the samples used in the phase of gene selection.

There are many penalty functions which are suitable for the regularized logistic regression model, such as lasso [4], adaptive lasso [31], and elastic net [5]. Zou and Zhang proposed the following adaptive elastic net as an improved version of the elastic net for analyzing high-dimensional data using a combination of the L_2 penalty and the adaptive L_1 penalty,

$$P_\alpha(\beta) = \sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha w_j|\beta_j|\right], \quad (3)$$

where $\{\omega_j\}_{j=1}^{p}$ are the adaptive data-driven weights [32]. The parameter α is typically fixed to select a trade-off between adaptive lasso penalization and ridge regression, while λ is varied to tune the model [33]. In order to overcome the problem that the correlated variables have different coefficients, the value of α should not be too close to one [34]. Existing studies show that this adaptive elastic net penalty not only have group effect which can select groups of correlated variables, but also can identify a number of representative biomarkers with clear biological meanings and achieve effective classification [32, 35, 36]. Instead of computing the weights by

$$w_j = \left(|\hat{\beta}_j^{EN}|\right)^{-r}, \quad (4)$$

where r is a positive constant and $\hat{\beta}_j^{EN}$ is the elastic net estimator, we calculate the weight for each gene according to the biological network information, resulting in the higher stability of the gene selection process. The specific method will be introduced in later section.

The intercept parameter β_0 and regression coefficient vector β can be estimated by the maximizer of objective function (2). The objective function can be written in the form of a concave function of the parameters as follows

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\beta_0 + \sum_{m=1}^{M}\tilde{\beta}_m\tilde{x}_{im}\right) - \log\left(1 + e^{\beta_0 + \sum_{m=1}^{M}\tilde{\beta}_m\tilde{x}_{im}}\right)\right] - \lambda P_\alpha(\beta), \quad (5)$$

where

$$\tilde{x}_{im} = x_{ij}x_{ik} \quad (6)$$

for the edge $j \sim k$ of the network, and M is the number of the edges that is much less than $p(p-1)/2$. The Newton algorithm can be used to approximate the objective function (5) by a second-order Taylor series expansion at current estimates [30]. Then the maximization of (5) is equivalent to a penalized weighted least-squares problem which can be solved by an efficient coordinate descent algorithm. Based on this algorithm, the matlab code "glmnet" can provide a path of solutions for a decreasing sequence of values for λ given a fixed value of α.

The weights for adaptive elastic net

For the regularized logistic regression with network-based pairwise interaction via adaptive elastic net (RLRNPI-AEN) (5), instead of computing the weights by Eq. (4), we calculate the weights based on the characteristics of networks topology.

Firstly, we present a biological network by a weighted graph $G(V, E, W)$, where V is the set of p nodes that correspond to the p variables, $E = \{e_{ij}\}$ is the set of edges between two nodes, and W is the set of weights of the edges. The edge weight can be used to measure uncertainly of the edge between vertices. For a network without explicit edge weights, the weight of every edge is set to one. A weighted adjacency matrix A can be used to represent the weighted edges, where $A_{ij} = w_{ij}$ if there exists a edge between nodes i and j and w_{ij} is the weight of the edge, and $A_{ij} = 0$ otherwise. For an undirected graph, the adjacency matrix A is symmetry. The degree of node i is defined as $d_i = \sum_{j=1}^{p} a_{ij}$.

Secondly, we construct a line graph for the interactions (edges) which describes the relationships between overlapping edges, where each node may belong to more than

one edge. The similarity between two edges is referred to the measure stated in [37]. The set that includes node i and its neighbors is denoted as $n_+(i)$. If two edges do not share a same node, their similarity will be set to zero. Otherwise, the similarity between edges e_{ik} and e_{jk} is defined as

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \bigcap n_+(j)|}{|n_+(i) \bigcup n_+(j)|}. \quad (7)$$

Since the shared node k provides no additional information, it does not present in the definition [37]. This definition is based on the assumption that edge pairs with a shared node are expected to be more similar than those unconnected pairs, and the similarity between two connected edges relies on the fraction of common neighbors that their unshared nodes have. Then, we can construct a line graph $G'(V', E', W')$, where V' consists of nodes which correspond to the interactions in the biological network (i.e. e_{ij}), and E' consists of the edges in this line graph. There is an edge between e_{ik} and e_{jk} since they share a common node and the weight of this edge is $S(e_{ik}, e_{jk})$ (Fig. 1). Since an interaction (node in line graph G') that includes high degree nodes (node in original graph G) will be more likely to have common nodes with other interactions, it will have high degree in the line graph G'. Studies have shown that disease-genes are always characterized by large degrees in biological networks, and are more likely to interact with other disease-genes [28, 29]. In addition, the variation of the high degree nodes will impose more influence on the action of the whole network, since they have more interacting partners. Therefore, we prefer to pick out the genes with higher degrees. The weights for adaptive elastic net (3) in objective function (5) are set to be inversely proportional to the degrees as follows,

$$w_j = \left(d'_j\right)^{-r} \quad (8)$$

where d'_j is the degree of the interaction j in the line graph G' and $r \geq 0$ is a parameter that controls the weights. This definition tends to select the interactions which include nodes with high degrees in the biological network. When $r = 0$, the adaptive elastic net penalty turns back to the elastic net. The larger value of r will promote the stability of the gene selection process because of the less dependence of the model on the samples. However, a too large r will generate a model that only focuses on the degrees of genes and neglects their discriminating abilities. We will discuss the specific method for the determination of r in the next section. The main steps of the proposed method are summarized in Algorithm 1.

Algorithm 1: RLRNPI-AEN

- **Input:** The gene expression of the training samples $X = (x_1, \cdots, x_n)^T$ and their corresponding labels $y = (y_1, \cdots, y_n)^T$, the biological network $G(V, E, W)$, the value of parameters α, λ, r, and the gene expression of the testing samples $X' = (x'_1, \cdots, x'_u)^T$.
- **Output:** The set of the selected genes SG, and the predicted posterior probability $p(x'_i) = Pr(y'_i = 1|x'_i)$ of the testing sample x'_i, $i = 1, \cdots, u$.

1. Constructing a line graph $G'(V', E', W')$ for the interactions (edges in graph $G(V, E, W)$) to describe the relationships between overlapping edges based on the similarity definition (7).
2. Computing the degree d'_j of each interaction in the line graph $G'(V', E', W')$.
3. The expression of each gene for both training and testing samples is mapped to the edge of the network $G(V, E, W)$ in different conditions according to (6).
4. Optimizing the objection function (5) by matlab code "glmnet" based on the training samples to obtain the estimators of $\tilde{\beta}_m, m = 1, \cdots, M$.
5. The interactions with the corresponding $\tilde{\beta}_m \neq 0$ are identified as informative edge-biomarkers, and the set of the selected genes SG includes the genes belonging to the edge-biomarkers.
6. The predicted posterior probability of testing sample $x'_i, i = 1, \cdots, u$ is

$$p(x'_i) = Pr\left(y'_i = 1|x'_i\right) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{m=1}^{M} \tilde{\beta}_m x'_{im}\right)}}.$$

Evaluation metrics

We use three metrics to assess the performance of various methods. Firstly, the classification performance of the model on the certain dataset is measured by the area under the ROC curve (AUC). Similar to [38] and [39], we perform ten times ten-fold cross-validation experiments for each dataset to minimize sampling noise. The repeated ten-fold cross-validation estimator is found to have better performance than the .632+ bootstrap estimator which suffers from a bias problem for large samples as well as for small samples [40]. Given a fixed α and r, another 10-fold validation is used for the selection of the parameter λ based on the training set (90 % of the original dataset), and the AUC is computed on the remaining testing set (10 % of the original dataset). Similar to that in [3], for each ten-fold cross-validation, the median AUC value is computed over the ten experiments.

Secondly, as mentioned in [41], the stability of gene selection process is compared on different samples

in three settings: the soft-perturbation, the hard-perturbation, and the between-datasets settings. For the regularized logistic regression with network-based pairwise interaction, the set of selected interactions will be transferred to gene set, where the genes which belong to different interactions will be considered only once. The soft-perturbation setting evaluates stability with respect to small perturbation of the training set, where a pair of training sets are randomly generated with 80 % overlap. Through the 10-fold cross-validation for λ, the following Jaccard coefficient between the two gene sets G_1 and G_2 is considered as the index for the stability of gene selection process,

$$JC = \frac{|G_1 \bigcap G_2|}{|G_1 \bigcup G_2|}. \quad (9)$$

The median is evaluated over the 20 repeats of the above random sampling. The hard-perturbation setting is referred to a procedure that randomly subsamples each dataset into pairs of subsets with no sample in common. A similar process is used to compute the median Jaccard coefficient. The between-datasets setting considers each dataset independently, using all samples on each dataset. In this setting, the genes are ranked by the number of times that the variables become selected when λ decreases. For each pair of the datasets, we compute the overlap of the two sets of the top k genes. The results are computed by taking the median stabilities of the total pairs of all the datasets.

Thirdly, the functional stability of a gene selection process is also assessed in the above three settings. We measure the functional similarity of two gene sets by the method stated in [42] which is based on the similarity of the GO terms of the genes. The three GO domains, biological process (BP), cellular component (CC) and molecular function (MF) are considered respectively.

Datasets

Breast cancer remains the most prevalent cancer among women in many countries. All the six datasets used in this study are measured on Affmetrix HGU133 microarrays, and each dataset includes 22283 transcripts. They are available through the Gene Expression Omnibus (GEO) database. A summary of these datasets are listed in Table 1. Except datasets GSE1456, GSE6532, GSE4922 which are downloaded as ready normalized, we preprocess the other three datasets by background correcting, quantile normalizing and log2 transforming using R package "preprocessCore" [43] (Bolstad, B: Probe level quantile normalization of high density oligonucleotide array data, unpublished). The probesets which do not have corresponding gene names are not considered and removed from our datasets, and the expression values for probesets that map to the same gene are averaged, resulting in 12754 genes. Both survival time information and the corresponding event indicator are used to divide samples into two classes according to whether patients develop a reported metastasis/relaspe/disease event within 5 years or are free of metastasis/relaspe/disease at least 7 years. A high-quality protein-protein interaction (PPI) network in H. sapiens from the High-quality INTeractomes (HINT) database (version: 06/03/2013) is used for constructing the pairwise interaction and calculating the weights in adaptive elastic net [44]. We only consider genes both in expression data and interaction network. In this PPI network, there are 18864 pairwise interactions including 6342 genes after self loops are removed. All datasets used in this study are obtained from papers and databases that have been already published and required no ethics approval.

Results and discussion

In the experiments, we apply seven algorithms, namely, regularized logistic regression via elastic net (RLR-EN) [30], regularized logistic regression via adaptive elastic net (RLR-AEN), regularized logistic regression with network-based pairwise interaction via elastic net (RLRNPI-EN), logistic regression with differentially expressed genes selected using the LIMMA package (limma) [45, 46], elastic-net-based prognosis prediction (ENCAPP) [3], SVM-based classification with average expression profile of pathways (SVM-AEP) [22] and RLRNPI-AEN on six human breast cancer datasets (The descriptions of RLR-EN and RLR-AEN are presented in Additional file 1). For limma, a 10-fold cross-validation is used to find the suitable number of differentially expressed genes. For ENCAPP, there are two parameters α and λ in the elastic net model, where α is fixed as described below and λ is chosen by a 10-fold cross-validation. SVM-AEP is implemented with R package "netClass" using default parameters [39].

Parameter settings

There are three parameters α, λ and r in the proposed model which require multi-parameter optimization based on cross-validation and grid search. However, when sample sizes are small, the performances of the proposed model with different parameters are often same or similar. So it is difficult to choose a proper parameter combination. In addition, a grid search for three parameters takes a long time. Therefore, it is useful to fix all but one parameter and select a model based on that parameter. We investigate the effect of α only on methods RLR-EN and RLRNPI-EN, where we do not need to consider the value of r. We run the RLR-EN and RLRNPI-EN models on the six datasets with $\alpha = 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2$. The median AUC based on 10-fold cross-validation for λ is shown in Fig. 2. For both RLR-EN and RLRNPI-EN

Table 1 Overview about employed breast datasets

Dataset	Publication	# patients	Classification	# patients of each class
GSE2034	[62]	242	time to relapse $\leq$ 5 y & relapse=True	95
			time to relapse $>$ 7 y & relapse=False	147
GSE1456	[63]	111	time to relapse $\leq$ 5 y & relapse=True	35
			time to relapse $>$ 7 y & relapse=False	76
GSE11121	[64]	125	t.dmfs $\leq$ 5 y & e.dmfs=True	28
			t.dmfs $>$ 7 y & e.dmfs=False	97
GSE6532	[65]	178	t.dmfs $\leq$ 5 y & e.dmfs=True	51
			t.dmfs $>$ 7 y & e.dmfs=False	127
GSE4922	[66]	204	DFS.time $\leq$ 5 y & DFS.status=True	70
			DFS time $>$ 7 y & DFS.status=False	134
GSE12093	[67]	79	DFS.time $\leq$ 5 y & DFS.status=True	12
			DFS.time $>$ 7 y & DFS.status=False	67

t.dmfs denotes the time for distant metastasis-free survival and e.dmfs is the corresponding event indicator. DFS.time denotes the time for disease-free survival and DFS.status is the corresponding event indicator

models, the results indicate that the classification processes are very similar with a small perturbation of α when α falls to $[0.2, 0.4]$. Thus, the value of α will be set to 0.3 for all the five models related with the elastic net.

The value of parameter r will result in a trade-off between stability of gene selection and accuracy of classification in the model with adaptive elastic net. We will consider the stability of RLR-AEN and RLRNPI-AEN in the between-dataset setting with different value of r ($r \in \{0.001, 0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8\}$). According to the definition of weight in (8), the larger value of r will enhance the influence of the degrees of the variables on the models which do not change with different samples. Figure 3 compares the median stability of the top 100 genes estimated by RLR-AEN and RLRNPI-AEN with different values of r using all samples on each dataset. With the increase of r, there is a clear growth trend for the median stability over total pairs of all the six datasets. This phenomenon accords with the theoretical analysis. However, on the other hand, a larger r may result in a poor classification process. A two-parameters grid search based on 10-fold cross-validation on the training set is used to find the proper value of r for each dataset. The chosen (λ, r) is the one giving the largest AUC over 10-fold cross-validation which may be not the same under different random partition for training and testing sets. If there are more than one combination (λ, r) that reach the biggest AUC, we will choose the one that has the largest r. Figure 4 presents the selected frequency for each r in 100 (10$\times$ 10-fold) experiments. In order to avoid a two-parameters grid search in the following experiments and the instability brought by different r selected based on

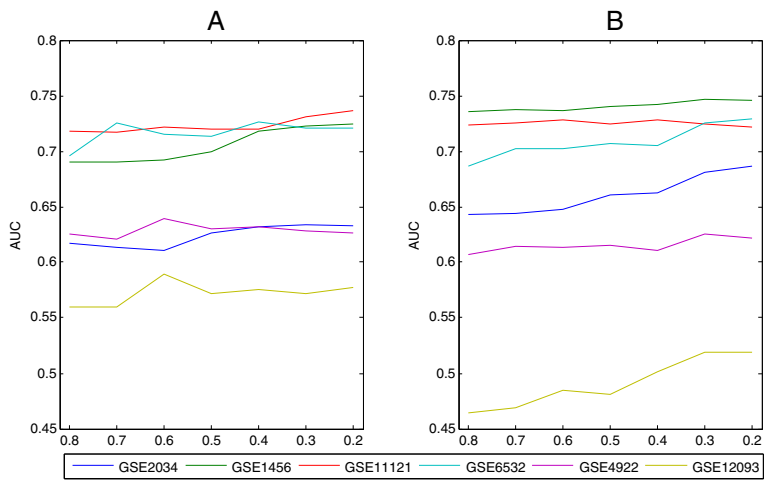

Fig. 2 Effect of parameter α on the resulting AUC of RLR-EN and RLRNPI-EN models. The median AUC are obtained over 100 experiments based on 10-fold cross validation for λ. **a** RLR-EN. **b** RLRNPI-EN

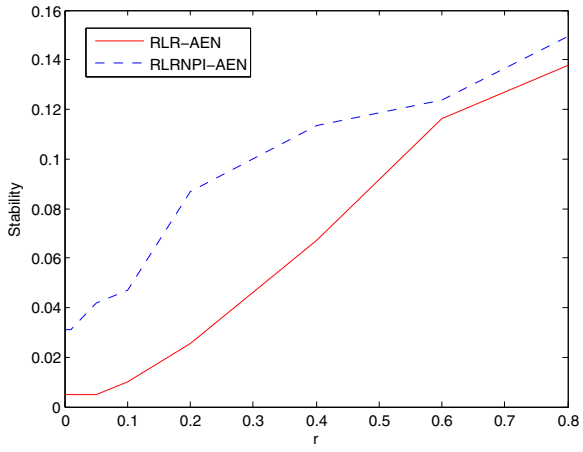

Fig. 3 Effect of parameter *r* on the stability of RLR-AEN and RLRNPI-AEN models in between-dataset setting. The median stability of the top 100 genes are computed using all samples on each dataset. The genes are ranked by the number of times that the variables become selected when λ decreases. The results are computed by taking the median stabilities of the total pairs of all the six datasets

different samples, we attempt to use a fixed *r* for each dataset, which is set to the most frequently occurring value in 100 experiments. According to Fig. 4, the value of *r* for each dataset is presented in Table 2. It is vary with the type of dataset for RLRNPI-AEN. The datasets considering the time for distant metastasis-free survival tend to choose the model with small *r*, while the other two types of datasets prefer the model with larger *r*.

Accuracy of the classification

Given the fixed α and *r*, we assess the predictive accuracy of the model based on a 10-fold cross-validation which is repeated ten times on each dataset. Table 3 shows the median AUC and the adjusted *p*-values for the seven models, where the *p*-values evaluate the significance of difference in classification performance between RLRNPI-AEN and the other six methods based on Mann-Whitney U test. For each dataset, the *p*-values are corrected using Holm-Bonferroni method for multiple testing. The Holm-Bonferroni method is more suitable for the multiple testing with small number of individual hypotheses and offers a simple test uniformly more powerful than the Bonferroni correction [47]. In general, classifiers using biological network information (RLR-AEN, RLRNPI-EN, ENCAPP, SVM-AEP and RLRNPI-AEN) have higher predictive power than those using gene expression dataset only (RLR-EN and limma). The results show that RLRNPI-AEN consistently outperforms limma with adjusted *p*-values < 0.05 for all the six datasets except GSE4922. Since the value of *r* used in the adaptive elastic net for RLRNPI-AEN is not large for datasets GSE2034, GSE1456 and GSE4922, the classification performances of RLRNPI-EN and RLRNPI-AEN do not have significant difference. RLR-AEN, RLRNPI-EN and SVM-AEP have similar performances which are significantly different from those of RLRNPI-AEN for datasets GSE11121, GSE6532 and GSE12093. ENCAPP generally achieve the second best classification performance. RLRNPI-AEN has the manifest superiority in the case of small sample size, especially

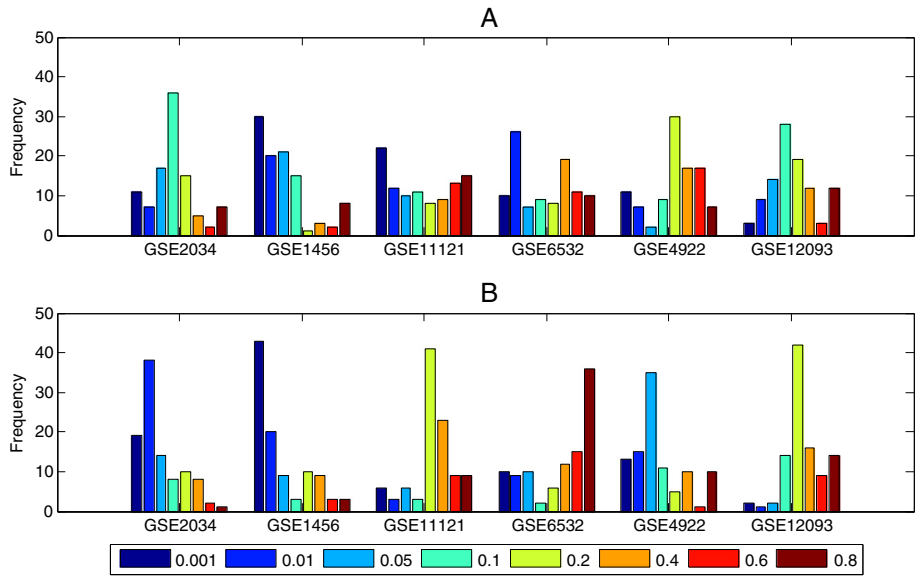

Fig. 4 The selected frequency of each *r* in 100 experiments for RLR-AEN and RLRNPI-AEN models. The chosen (λ, r) is the one giving the highest AUC over 10-fold cross validation. If there are more than one combination (λ, r) that reach the highest AUC, we will choose the one that has the largest *r*. **a** RLR-AEN. **b** RLRNPI-AEN

Table 2 The value of *r* used in RLR-AEN and RLRNPI-AEN for each dataset

Dataset	GSE2034	GSE1456	GSE11121	GSE6532	GSE4922	GSE12093
RLR-AEN	0.1	0.001	0.001	0.01	0.2	0.1
RLRNPI-AEN	0.01	0.001	0.2	0.8	0.05	0.2

for datasets GSE11121 and GSE12093 (adjusted *p*-values < 0.05 compared with all the other six methods). The network-based pairwise interaction may help the model away from the effects of noise and instability brought by the small sample size.

Stability of gene selection process

The stability of gene selection process is evaluated by Jaccard coefficient in three settings: the soft-perturbation, the hard-perturbation, and the between-datasets settings. Both soft-perturbation and hard-perturbation settings are based on the 10-fold cross-validation for the tuning parameter λ. Figure 5 shows the median stability of the selected genes estimated by seven models over 20 repeats of random sampling (The summary of median stability and adjusted *p*-values is presented in Additional file 2: Table S1). It appears very clearly that RLRNPI-AEN provides more stable gene selection process than the other six methods, especially in the hard-perturbation setting where the adjusted *p*-values are less than 10^{-3} in most cases. With a larger value of *r*, a significant stability is observable for datasets GSE11121, GSE6532 and GSE12093, using RLRNPI-AEN in soft-perturbation setting (adjusted *p*-values < 0.05 except that of ENCAPP for GSE11121). This indicates that the idea of setting the weights for adaptive elastic net to be inversely proportional to the degrees is very useful for improving the stability of the gene selection. Since ENCAPP and SVM-AEP also take into account the biological network information, they have higher median stabilities in some of the datasets GSE2034, GSE1456 and GSE4922 in the soft-perturbation setting, where the value of *r* for RLRNPI-AEN is small. However, in these cases, the *p*-values indicate that the differences are not significant. In the between-datasets setting, another way is used to present the ability of one gene selection method in identifying the similar gene set for the same cancer from different datasets. We use all samples of each dataset. For the five elastic net based methods, given a decreasing sequence of values for λ, the genes are ranked by the number of times that they are selected. For limma and SVM-AEP, the genes are sorted in increasing order by the *p*-values of the corresponding tests, i.e., the moderated t-statistic test for limma and the test related with GO category's enrichment analysis for SVM-AEP. Figure 6 gives the stabilities for the seven methods which are obtained by taking the median of pairwise stability measures. On one hand, for all the seven methods, the stability rise with the number of the top genes. The bigger gene sets will have more chance to share some genes. On the other hand, the four methods that consider biological network information yield much more stable biomarker selection than RLR-EN, RLR-AEN and limma. Because RLRNPI-AEN uses pairwise interactions based on both gene expression

Table 3 Prediction performance of RLRNPI-AEN in comparison to other methods in terms of area under ROC curve (AUC)

Dataset	RLR-EN	RLR-AEN	RLRNPI-EN	limma	ENCAPP	SVM-AEP	RLRNPI-AEN
GSE2034	0.638	0.663	0.657	0.627	**0.681**	0.647	**0.690**
	(0.0219)	(0.0516)	(0.0516)	**(0.0036)**	(0.5966)	(0.0516)	(–)
GSE1456	0.724	**0.734**	0.711	0.619	0.722	0.717	**0.736**
	(0.9768)	(0.9920)	(0.5600)	**(0.0114)**	(0.9920)	(0.6024)	(–)
GSE11121	0.736	0.725	0.739	0.542	**0.750**	0.695	**0.820**
	(0.0171)	**(0.0076)**	**(0.0250)**	**(0.0012)**	**(0.0250)**	**(0.0050)**	(–)
GSE6532	0.721	0.725	0.725	0.643	**0.730**	0.715	**0.747**
	(0.0451)	**(0.0424)**	**(0.0422)**	**(0.0012)**	(0.4725)	**(0.0219)**	(–)
GSE4922	**0.620**	0.611	0.611	0.606	0.593	**0.622**	0.614
	(1.0000)	(1.0000)	(1.0000)	(1.0000)	**(0.1032)**	(1.0000)	(–)
GSE12093	0.571	0.518	0.613	**0.685**	0.616	0.607	**0.845**
	(0.0012)	**(0.0012)**	**(0.0012)**	**(0.0208)**	**(0.0034)**	**(0.0012)**	(–)

The median AUC obtained for each method on the six datasets over ten times ten-fold cross validation. The adjusted *p*-values calculated using a Mann-Whitney U test are shown within parentheses, which evaluate the significance of difference in classification performance between RLRNPI-AEN and the other six methods. For each dataset, they are corrected using Holm-Bonferroni method for multiple testing. The best two median AUCs and the adjusted *p*-values that are less than 0.05 are shown in boldface

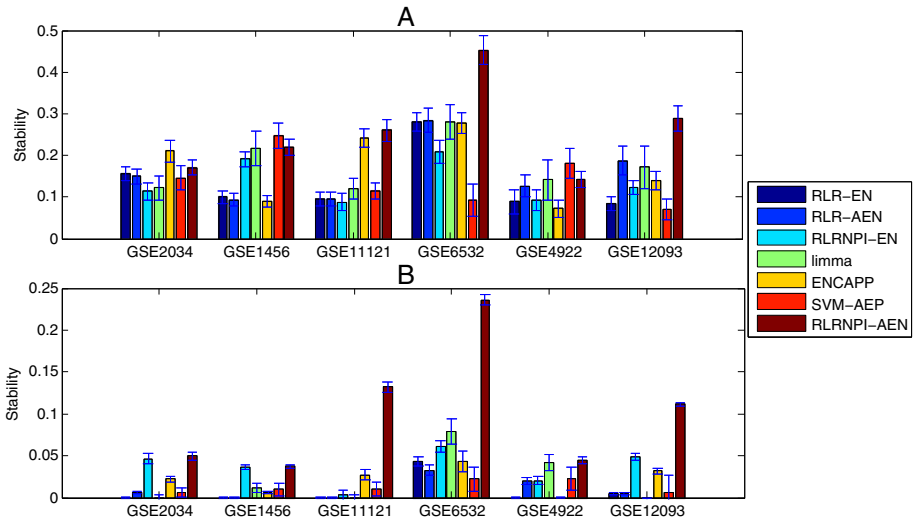

Fig. 5 Stability for the selected genes based on 10-fold cross-validation. Median and standard errors are obtained over the 20 times random sampling of subsets. **a** Soft-perturbation setting. **b** Hard-perturbation setting

dataset and biological network to identify marker genes, it can retrieve significantly more overlapped biomarkers. For example, among the top-100 genes estimated by RLRNPI-AEN from datasets GSE11121 and GSE6532 which both consider the time for distant metastasis-free survival, there are 20 common genes, while RLR-EN, RLR-AEN, RLRNPI-EN, limma, ENCAPP and SVM-AEP only identify 2, 1, 8, 0, 5 and 2 common genes, respectively (The number of common genes between each pair of datasets among the top-100 genes is presented in Additional file 3: Table S2). Our results indicate that RLRNPI-AEN is an effective way to produce a more stability gene set.

Functional stability

Since many gene selection methods are based on the samples, it is not easy to reach a high stability in term of genes with the change of samples. However, these biomarker sets with little common genes may exhibit the same biological function which also make sense for cancer diagnosis, treatment, and prognosis. Therefore, it is

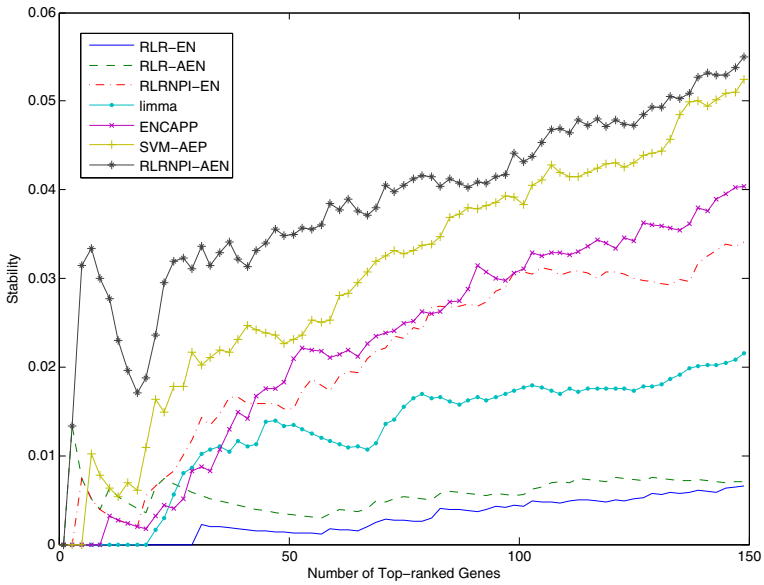

Fig. 6 Stability for the top-ranked genes in between-datasets setting based on all samples on each dataset. The stabilities are obtained by taking the median of the pairwise stability measures. The x-axis is the number of selected genes

important for one gene selection method to identify some specific biological function related with the cancer in a robust manner. The stability analysis of the model in terms of biological function is implemented by using R package "GOSemSim" [42]. We assess on Figs. 7 and 8 the functional stability of all methods in the soft-perturbation and hard-perturbation settings, respectively. The three GO domains are analyzed separately (The summary of the median GO stability and adjusted p-values for the three GO domains BP, CC and MF in soft-perturbation and hard-perturbation settings is presented in Additional file 4: Table S3). From Figs. 7, 8 and Additional file 4: Table S3 , we can find that the stability results at the functional level are very similar to the results at the gene level. Overall, RLRNPI-AEN is the most stable method. The advantages of the network-based methods in GO stability further show the strength of biological networks for achieving more clear biological interpretation. Since the sample difference is relatively small in soft-perturbation setting, the functional stabilities of the ENCAPP, SVM-AEP and RLRNPI-AEN models do not have large distinction for some datasets such as GES2034. However, RLRNPI-AEN exhibits significant functional stability benefits in hard-perturbation setting where there is no overlap in samples. In addition, the functional stability results with the change of the number of the top genes for the seven methods in between-datasets setting are presented in Fig. 9. In general, RLRNPI-EN and RLRNPI-AEN offers significant benefits in terms of stabilities for GO terms BP and MF. Another two methods ENCAPP and SVM-AEP which are integrated with biological network also perform better than RLR-EN, RLR-AEN and limma. There is no obvious difference between the functional stability of RLRNPI-EN and RLRNPI-AEN in this between-datasets setting. Since this pair of methods is based on similar ideas, they tend to identify genes with consistent functions. However, there are clear strengths of highlighting the heterogeneous underlying interactions.

Biomarker identification

The other five methods focus on the differentially expressed genes or modules, while RLRNPI-EN and RLRNPI-AEN identify the gene interactions of which the changing may result in different states of a biological system. Figure 10 presents the number of the different genes among the top-k genes selected by RLR-AEN and RLRNPI-AEN, using all the samples on each dataset. Although these numbers are vary with different datasets, all the results indicate that there are obvious difference between the biomarkers identified by RLR-AEN and RLRNPI-AEN. We will make a detailed analysis of the biomarkers for datasets GSE1456 and GSE11121.

For GSE1456, we first decide the number of biomarkers based on AUC by 10-fold cross-validation on the whole dataset. To reduce the variability, the cross-validation is conducted 10 times. Then we select the top-k variables which are sorted by the number of times that the variables become selected when λ decreases, where k is set

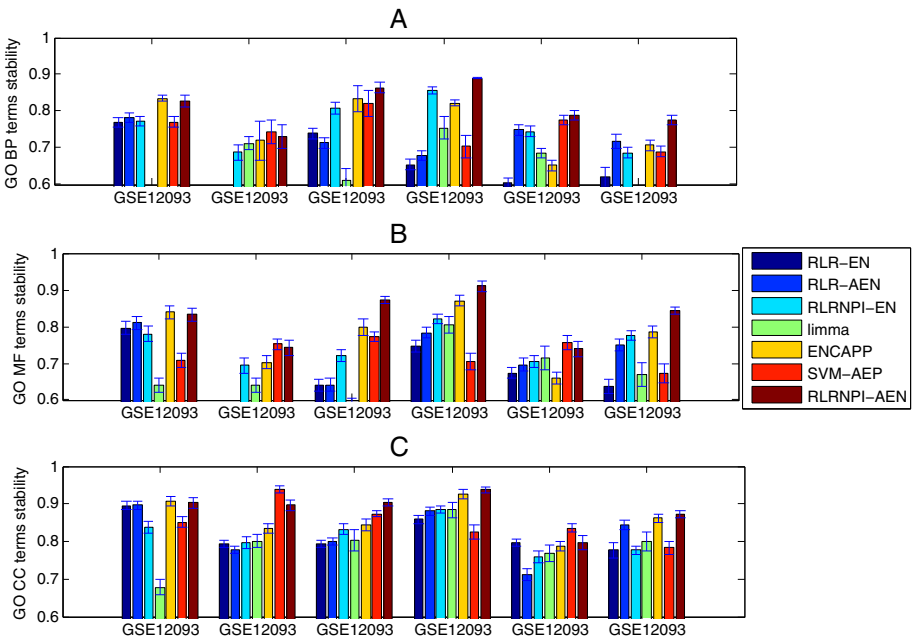

Fig. 7 GO stability for the selected genes in soft-perturbation setting based on 10-fold cross-validation. Median and standard errors are obtained over the 20 times random sampling of subsets. **a** BP term. **b** MF term. **c** CC term

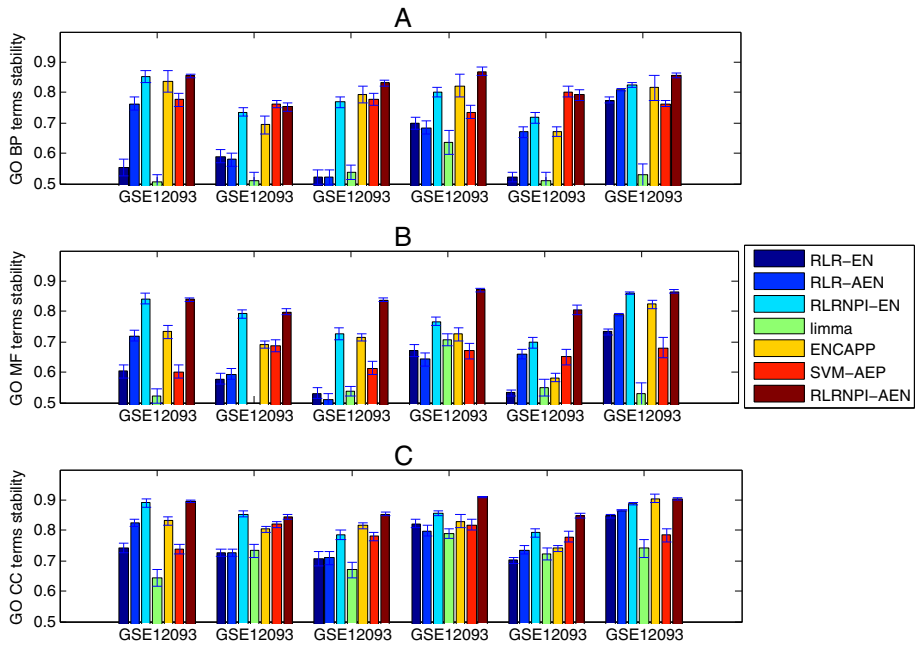

Fig. 8 GO stability for the selected genes in hard-perturbation setting based on 10-fold cross validation. Median and standard errors are obtained over the 20 times random sampling of subsets. **a** BP term. **b** MF term. **c** CC term

to the median biomarker number over the 10 repeated experiments. The subnetworks based on the 19 interactions identified by RLRNPI-AEN are presented in Fig. 11, where the color of each node indicates whether the gene belong to the top-19 genes estimated by RLR-AEN (red) or not (green). There are 34 genes selected as informative, including 5 expression-based discriminative genes that belong to the 19 genes selected by RLR-AEN (The 19 important interactions identified by RLRNPI-AEN for datasets GSE1456 are presented in Additional file 5: Table S4). RLR-AEN tends to select a part of genes of the subnetwork which are differentially expressed and neglect the genes who interact with some discriminative genes to form a collective biological function or present differential correlation under many kinds of biological states. The functional and biological relationships of the selected genes of each subnetwork are analyzed based on the GO annotation, which is implemented by using R

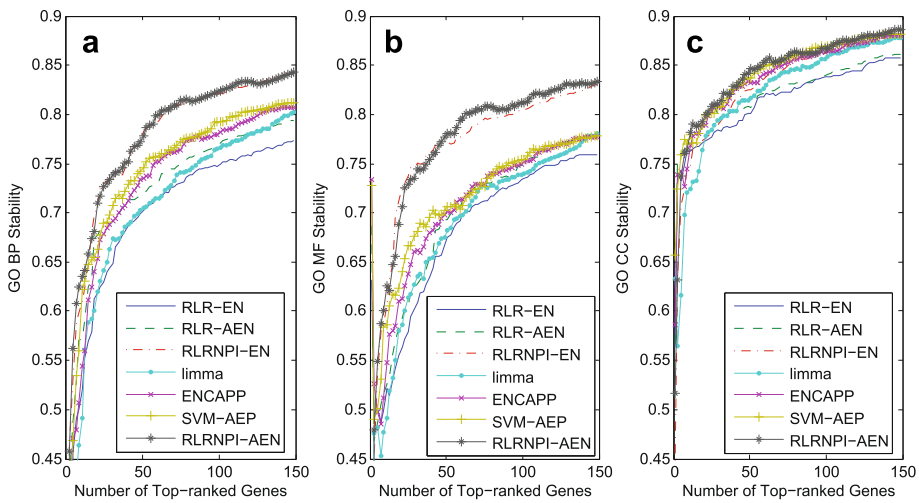

Fig. 9 GO stability for the top-ranked genes in between-datasets setting based on all samples on each dataset. The GO stabilities are obtained by taking the median of the pairwise stability measures. The x-axis is the number of selected genes. **a** BP term. **b** MF term. **c** CC term

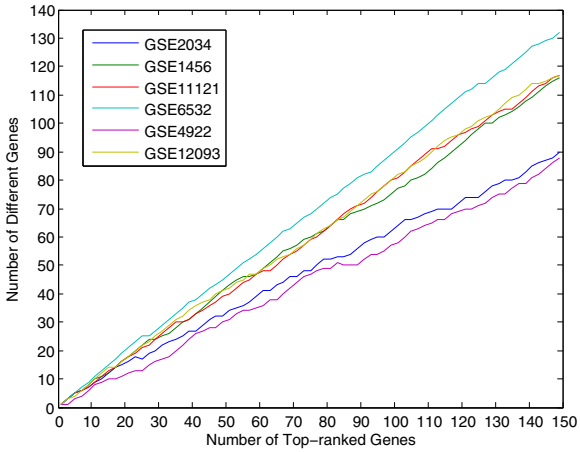

Fig. 10 The number of the different genes among the top-k genes selected by RLR-AEN and RLRNPI-AEN. The x-axis is the number of selected marker genes ranked by the number of times that the variables become selected when regularization parameter decreases. The y-axis is the number of the different genes that do not selected by RLR-AEN and RLRNPI-AEN simultaneously

package "clusterProfiler" [48]. A p-value of a GO terms set transferred from a gene set is calculated using the hypergeometric distribution and then is adjusted using the FDR correction for multiple testing. Table 4 lists the GO terms with the smallest adjusted p-value for some subnetworks shown in Fig. 11 at 5 % FDR. The small p-value shows that the genes in each subnetwork have significant biological and functional correlation, and the common GO functions they share are often related to the relapse time of breast cancer.

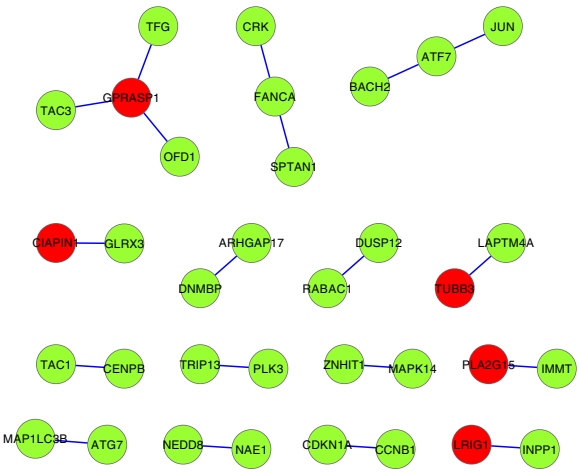

Fig. 11 The subnetworks for dataset GSE1456 estimated by RLRNPI-AEN. Nodes represent human genes, and they are connected by a link if they belong to the PPI network. Each gene is labeled by its gene symbol. The color of each node indicates whether the gene are identified by RLR-AEN (*red*) or not (*green*)

Specifically, the smallest adjusted p-value is 2.61×10^{-4} corresponding to GO:0000977 which is related to RNA polymerase II regulatory region sequence-specific DNA binding. All the genes in the third subnetwork share this common GO function, which are not identified by RLR-AEN. It is a molecular function related with RNA polymerase II which plays an important role in breast [49, 50].

Figure 12 shows the heatmap of the expression profile of the involved 34 genes and the 19 interactions. The distinction between two classes are more significant in edge data than that in the gene expression, which shows that our proposed method does select reasonable biomarkers for classification. Most of identified biomarkers are considered to have diagnostic values for breast. For example, the suppression of LRIG1 gene of the top-1 edge is identified as a common feature of breast tumors, and contributes to poor patient prognosis and therapeutic resistance [51]. TAC1 has been implicated in the development of breast which can lead to the production of cytokines with growth promoting functions [52]. In addition, CIAPIN1 is reported to participate in breast cancer multi-drug resistance, changing cell cycle and enhancing the anti-apoptotic capability of cells [53].

Unlike conventional regularized logistic regression, our model with network-based pairwise interaction can implicate disease-related genes with low discriminative potential, such as ATF2 and OFD1. The activation of ATF2 has been detected to play a direct role in malignant phenotypic changes of human breast epithelial cells [54]. Furthermore, OFD1 is reported to be possible to reverse the cilia-defective phenotype of a transformed breast cancer cell line [55].

Next, the top-26 interactions for dataset GSE11121 are also analyzed, including 31 informative genes of which 3 are shared with RLR-AEN (The 26 important interactions identified by RLRNPI-AEN for datasets GSE11121 are presented in Additional file 5: Table S4). Figure 13 presents the network of the top 26 interactions. Different from the subnetworks with a few nodes for GSE1456, the biomarkers tend to form a larger subnetwork for GSE11121. The larger value of r for GSE11121 may make RLRNPI-AEN select the genes with higher degree. It indicates that there may exist different modes between breast metastasis and relapse. The GO results for some subnetworks are shown in Table 5. It can be seen from this table that genes in each subnetwork share some common GO functions with high statistical significance, indicating high biological and functional correlation of the genes in this subnetwork. Therefore, the genes without differentially expressed remain essential for maintaining the integrity of the subnetwork. For the first subnetwork, GO:0002433 with adjusted p-value 2.66×10^{-10} is a biological process shared by seven genes among 16 genes, which is related

Table 4 The Gene Ontology results of the subnetwork identified by RLRNPI-AEN for dataset GSE1456

Subnetwork	GO number	Ontology description	Adjusted *p*-value
TFG GPRASP1 TAC3 OFD1	GO:0043015	gamma-tubulin binding	3.07×10^{-2}
	GO:0043014	alpha-tubulin binding	3.07×10^{-2}
SPTAN1 CRK FANCA	GO:0045309	protein phosphorylated amino acid binding	3.02×10^{-2}
	GO:0030507	spectrin binding	3.02×10^{-2}
	GO:0046875	ephrin receptor binding	3.02×10^{-2}
	GO:0042169	SH2 domain binding	3.02×10^{-2}
	GO:0051219	phosphoprotein binding	3.48×10^{-2}
BACH2 ATF7 JUN	GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding	2.61×10^{-4}
	GO:0001012	RNA polymerase II regulatory region DNA binding	2.61×10^{-4}
	GO:0000976	transcription regulatory region sequence-specific DNA binding	2.61×10^{-4}
	GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity	2.61×10^{-4}
	GO:0000980	RNA polymerase II distal enhancer sequence-specific DNA binding	2.61×10^{-4}

The first column (Subnetwork) presents the elements of subnetwork of which the functional and biological relationship are analyzed based on the GO annotation

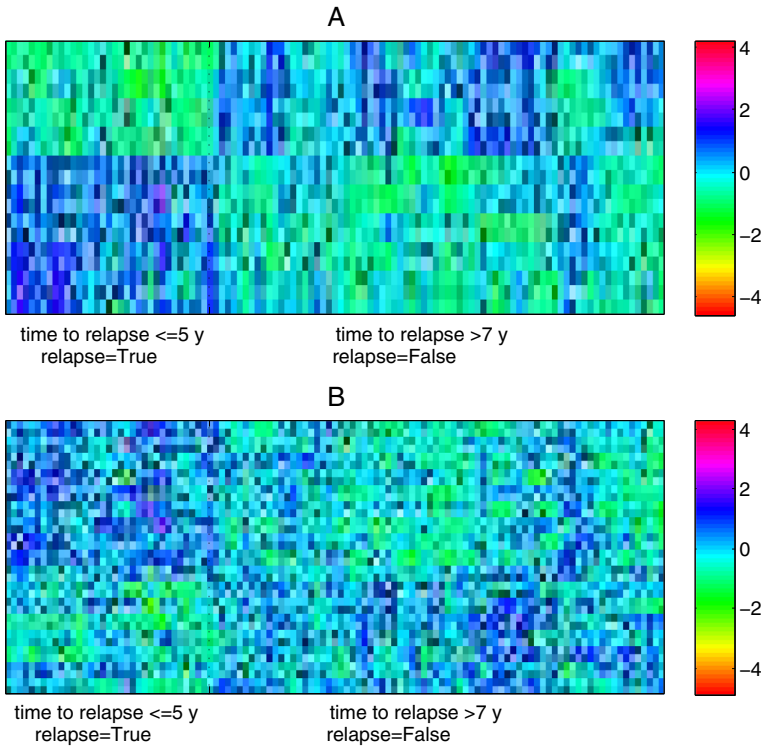

Fig. 12 The heatmap of the expression profile of the 19 interactions identified by RLRNPI-AEN and their corresponding genes for dataset GSE1456. **a** Interactions (**b**) Genes

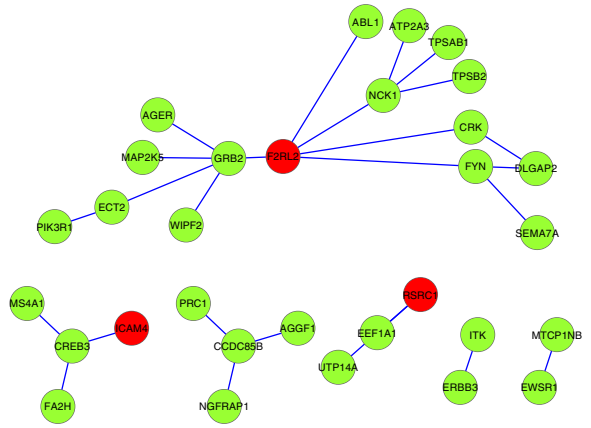

Fig. 13 The subnetworks for dataset GSE11121 estimated by RLRNPI-AEN. Nodes represent human genes, and they are connected by a link if they belong to the PPI network. Each gene is labeled by its gene symbol. The color of each node indicates whether the gene are identified by RLR-AEN (*red*) or not (*green*)

to immune response-regulating cell surface receptor signaling pathway involved in phagocytosis. There are results demonstrate that phagocytosis of extracellular matrix is an inherent feature of breast tumor cells that correlates with and may even directly contribute to their invasive capacity [56].

The heatmap of the interactions and the corresponding genes are presented in Fig. 14. It is consistent with the results presented in Fig. 12 that the opposite patterns between the two classes are clear in interaction data but not in the gene expression data. Some of the genes are well known in the literatures. Although one node of top-1 edge ICAM4 is not yet indicated to play an important role in breast cancer susceptibility, its role in cell adhesion and cell signaling together with its low level expression in cancer-relevant tissues leave the possibility that its dysregulation or dysfunction may increase cancer risk [57]. The other node CREB3 is not selected by RLR-AEN of which the over-expression is shown to substantially increase the migration of MDA-MB-231 metastatic breast cancer cells [58]. FYN which belongs to the top-2 edge is implicated in diverse biological functions such as neuronal development, T-cell receptor signaling, and

Table 5 The Gene Ontology results of the subnetwork identified by RLRNPI-AEN for dataset GSE11121

Subnetwork	GO number	Ontology description	Adjusted *p*-value
AGER ATP2A3 CRK CSF2RA DLGAP2 ECT2 F2RL2 FYN GRB2 MAP2K5 NCK1 PIK3R1 SEMA7A TPSAB1 TPSB2 WIPF2	GO:0002433	immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	2.66×10^{-10}
	GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis	2.66×10^{-10}
	GO:0038094	Fc-gamma receptor signaling pathway	2.66×10^{-10}
	GO:0002431	Fc receptor mediated stimulatory signaling pathway	2.70×10^{-10}
	GO:0006909	phagocytosis	5.53×10^{-8}
CREB3 MS4A1 ICAM4 FA2H	GO:0031726	CCR1 chemokine receptor binding	2.25×10^{-2}
	GO:0008140	cAMP response element binding protein binding	2.25×10^{-2}
	GO:0044877	macromolecular complex binding	2.25×10^{-2}
	GO:0035497	cAMP response element binding	2.76×10^{-2}
	GO:0005102	receptor binding	2.76×10^{-2}
AGGF1 CCDC85B PRC1 NGFRAP1	GO:0005123	death receptor binding	4.17×10^{-2}
	GO:0008656	cysteine-type endopeptidase activator activity involved in apoptotic process	4.17×10^{-2}
	GO:0016505	peptidase activator activity involved in apoptotic process	4.17×10^{-2}
	GO:0019894	kinesin binding	4.17×10^{-2}
	GO:0016504	peptidase activator activity	4.17×10^{-2}

The first column (Subnetwork) presents the elements of subnetwork of which the functional and biological relationship are analyzed based on the GO annotation

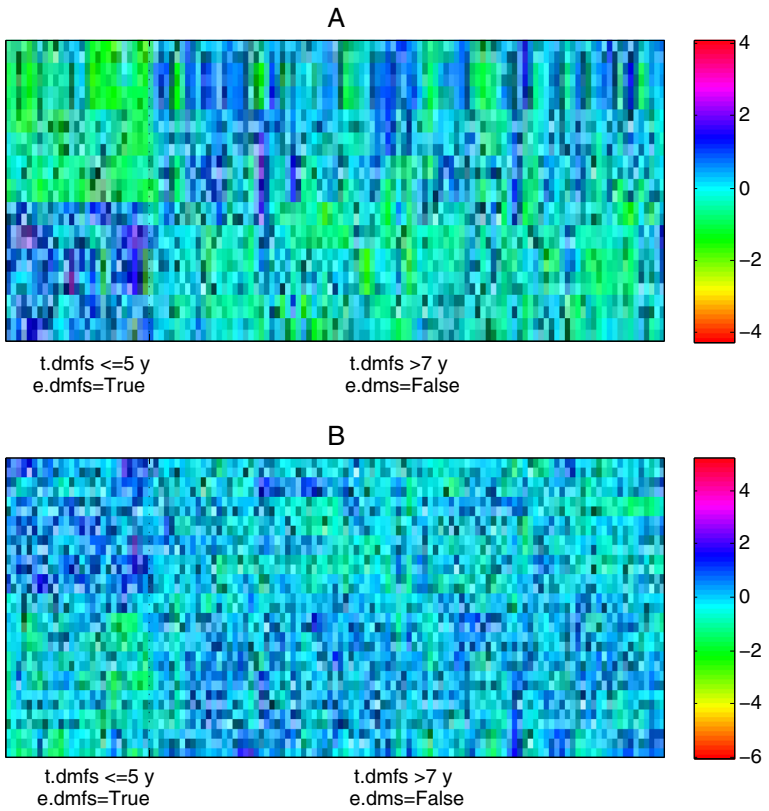

Fig. 14 The heatmap of the expression profile of the 26 interactions identified by RLRNPI-AEN and their corresponding genes for dataset GSE11121. **a** Interactions (**b**) Genes

is reported to be linked to increased breast cancer risk, especially in women with expression of ER and PR in their breast tumors [59]. GRB2 of the top-3 edge plays an important role in the first subnetwork. Over expression of GRB2 might modulate the growth factor sensitivity of human breast cancer cells and has influence on tumor progression [60].

Conclusions

In this paper, we present an effective biomarker discovery and cancer classification algorithm: a regularized logistic regression with network-based pairwise interaction via adaptive elastic net. Different from the algorithm based on functional modules proposed by Das et. al in [3], where the modules often need to be determined in advance by some clustering methods and thus the algorithm performance may depend on the module detection, we focus on the gene pairs which exhibit different positive or negative interactions. The discriminative modules treat the genes and the interactions as a whole and do not consider the diversity of the interactions between a variety of diseases. Since the mutation of buried residues of the proteins leads to loss or gain of specific interactions, the interactions identified by the proposed method can be great helpful for the analysis of exposed residues in turn of which the mutation has been shown to be a higher fraction of mutations associated with autosomal-dominant diseases [25, 61]. In addition, by considering the changes of interactions towards different biological states, we can identify the non-differentially expressed genes which play central roles in functional process within cells. Our algorithm combines gene expression profiles with PPI networks, which can reduce the influence of noise brought from the correlation between expression that in fact have no underlying biological causality. The degree information based on the PPI network is introduced to make the model less sensitive to the training samples and predict biomarkers with higher reproducibility.

Since we only take account of the interactions between two genes, it may miss some informative biomarkers which do not participate in any interactions. Therefore, in the future work, the subnetworks or the pathways consisted of both nodes and edges are needed for the accuracy of diagnostic and prognostic biomarker identification. Although the network information introduced in our model facilitate the discovery of more reproductivity biomarkers, the results may be dependent on the network structure. With the availability of a variety of biological

networks such as KEGG pathways, we can incorporate all these sources as prior information to build variable selection methods to decline the sensitive of the model towards both gene expression data and networks.

Availability of data and materials

The datasets supporting the results of this article are included within its additional files.

Additional files

Additional file 1: The introductions for RLR-EN and RLR-AEN. This section provides the brief introductions for the two compared methods: RLR-EN and RLR-AEN. (PDF 128 kb)

Additional file 2: Table S1. Summary of the median stability and the adjusted *p*-values for the different datasets in soft-perturbation and hard-perturbation settings. (XLSX 12.9 kb)

Additional file 3: Table S2. Summary of the number of common genes between each pair of datasets among the top-100 genes. (XLSX 10.4 kb)

Additional file 4: Table S3. Summary of the median GO stability and the adjusted *p*-values for the three GO domains (BP, CC, MF) in soft-perturbation and hard-perturbation settings. (XLSX 18.7 kb)

Additional file 5: Table S4. The important interactions identified by RLRNPI-AEN for datasets GSE1456 and GSE11121. (XLSX 10.3 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MYW, XFZ, DQD, LOY, YZ and HY designed the method and conceived the study. MYW and XFZ implemented the method and performed the experiments. MYW, XFZ, DQD, LOY, YZ and HY wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the Associate Editor and reviewers for their thoughtful and useful comments. This work is partially supported by the National Science Foundation of China [61402276, 61402190, 11401110, 61375033, 11171354, 61532008], the Program for Changjiang Scholars and Innovative Research Team in SUFE [IRT13077], the State Key Program in the Major Research Plan of National Natural Science Foundation of China [91546202], the Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE [CCNU15A05039, CCNU15ZD011], the Ministry of Education of China [20120171110016], the Natural Science Foundation of Guangdong Province [2013KJCX0086], the Foundation of China University of Geosciences (Wuhan) [AU2015CJ008], and the international Program Fund of 985 Project, Sun Yat-sen University.

Author details

[1]School of Statistics and Management, Shanghai University of Finance and Economics, Guoding Road, 200433 Shanghai, China. [2]Key Laboratory of Mathematical Economics SUFE, Ministry of Education, Guoding Road, 200433 Shanghai, China. [3]School of Mathematics and Statistics & Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Luoyu Road, 430079 Wuhan, China. [4]Intelligent Data Center and Department of Mathematics, Sun Yat-Sen University, Xingang West Road, 510275 Guangzhou, China. [5]College of Information Engineering, Shenzhen University, Nanhai Avenue, 518060 Shenzhen, China. [6]School of Automation, China University of Geosciences, Lumo Road, 430074 Wuhan, China. [7]Department of Electronic and Engineering, City University of Hong Kong, Tat Chee Avenue, 999077 Hong Kong, China.

Received: 6 August 2015 Accepted: 28 January 2016
Published online: 27 February 2016

References

1. Mitra K, Carvunis AR, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013;14(10):719–32.
2. Hathout Y, Brody E, Clemens PR, Cripe L, DeLisle RK, Furlong P, Gordish-Dressman H, Hache L, Henricson E, Hoffman EP, et al. Large-scale serum protein biomarker discovery in Duchenne muscular dystrophy. Proc Natl Acad Sci. 2015;112(23):7153–8.
3. Das J, Gayvert KM, Bunea F, Wegkamp MH, Yu H. ENCAPP: elastic-net-based prognosis prediction and biomarker discovery for human cancers. BMC Genomics. 2015;16(1):263.
4. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.
5. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Methodol. 2005;67(2):301–20.
6. Cun Y, Fröhlich H. Network and data integration for biomarker signature discovery via network smoothed t-statistics. PloS ONE. 2013;8(9):73074.
7. Qin G, Zhao XM. A survey on computational approaches to identifying disease biomarkers based on molecular networks. J Theor Biol. 2014;362:9–16.
8. Fröhlich H. Network based consensus gene signatures for biomarker discovery in breast cancer. PloS ONE. 2011;6(10):25364.
9. Zhang W, Zeng T, Chen L. EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers. J Theor Biol. 2014;362:35–43.
10. Michailidis G. Statistical challenges in biological networks. J Comput Graph Stat. 2012;21(4):840–55.
11. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. 2012;30(2):159–64.
12. Das J, Hao RL, Adithya S, Robert F, Liang J, Wei X, Wang X, Mort M, Stenson PD, Cooper DN. Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks. Hum Mutat. 2014;35(5):585–93.
13. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
14. Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. Ann Stat. 2013;41(3):1111–41.
15. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008;24(9):1175–82.
16. Kim S, Pan W, Shen X. Network-based penalized regression with application to genomic data. Biometrics. 2013;69(3):582–93.
17. Zhang W, Ota T, Shridhar V, Chien J, Wu B, Kuang R. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. PLoS Comput Biol. 2013;9(3):1002975.
18. Zhe S, Naqvi SA, Yang Y, Qi Y. Joint network and node selection for path way-based genomic data analysis. Bioinformatics. 2013;29(16):1987–96.
19. Wang Z, Xu W, San Lucas FA, Liu Y. Incorporating prior knowledge into gene network study. Bioinformatics. 2013;29(20):2633–640.
20. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3(1):140.
21. Zhang X, Gao L, Liu ZP, Chen L. Identifying module biomarker in type 2 diabetes mellitus by discriminative area of functional activity. BMC Bioinforma. 2015;16(1):92.
22. Zheng G, Zhang T, Xia L, Qi W, Xu J, Hui Y, Jing Z, Wang H, Wang C, Topol EJ. Towards precise classification of cancers based on robust gene functional expression profiles. BMC Bioinforma. 2005;6(11):1–12.
23. Gambardella G, Moretti MN, de Cegli R, Cardone L, Peron A, di Bernardo D. Differential network analysis for the identification of condition-specific pathway activity and regulation. Bioinformatics. 2013;29(14):1776–85.
24. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012;8(1):565.
25. Quan Z, Nicolas S, Li Q, Benoit C, Fabien H, Niels K, Stanley T, Yu H, Kavitha V, Mou D. Edgetic perturbation models of human inherited disorders. Mol Syst Biol. 2009;5(1):321.
26. Das J, Fragoza R, Lee HR, Cordero NA, Guo Y, Meyer MJ, Vo TV, Wang X, Yu H. Exploring mechanisms of human disease through structurally resolved protein interactome networks. Mol BioSyst. 2013;10(1):9–17.
27. Winter C, Kristiansen G, Kersting S, Roy J, Aust D, Knösel T, Rümmele P, Jahnke B, Hentrich V, Rückert F, et al. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. PLoS Comput Biol. 2012;8(5):1002511.

28. Xu J, Li Y. Discovering disease-genes by topological features in human protein–protein interaction network. Bioinformatics. 2006;22(22):2800–5.
29. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009;27(2):199–204.
30. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
31. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–29.
32. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Ann Stat. 2009;37(4):1733–51.
33. Günther OP, Chen V, Freue GC, Balshaw RF, Tebbutt SJ, Hollander Z, Takhar M, McMaster WR, McManus BM, Keown PA, et al. A computational pipeline for the development of multi-marker bio-signature panels and ensemble classifiers. BMC Bioinforma. 2012;13(1):326.
34. Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. Bioinformatics. 2012;28(10): 1368–75.
35. Hughey JJ, Butte AJ. Robust meta-analysis of gene expression using the elastic net. Nucleic Acids Res. 2015;43(12):e79.
36. Falgreen S, Dybkær K, Young KH, Xu-Monette ZY, El-Galaly TC, Laursen MB, Bødker JS, Kjeldsen MK, Schmitz A, Nyegaard M, et al. Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. BMC Cancer. 2015;15(1):235.
37. Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature. 2010;466(7307):761–4.
38. Hamp T, Rost B. More challenges for machine learning protein interactions. Bioinformatics. 2015;31(10):1521–5.
39. Cun Y, Fröhlich H. netClass: An R-package for network based, integrative biomarker signature discovery. Bioinformatics. 2014;30(9):1325–6.
40. Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal. 2009;53(11): 3735–45.
41. Haury AC, Gestraud P, Vert JP. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS ONE. 2011;6(12):28210.
42. Wang JZ, Du Z, Payattakool R, Philip SY, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81.
43. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.
44. Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol. 2012;6(1):92.
45. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
46. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):1–28.
47. Hommel G. A stagewise rejective multiple test procedure on a modified boneferroni test. Biometrika. 1988;75(2):383–6.
48. Yu G, Wang LG, Han Y, He QY. Clusterprofiler: an R package for comparing biological themes among gene clusters. OMICS J Integr Biol. 2012;16(5):284–7.
49. Scully R, Anderson SF, Chao DM, Wei W, Ye L, Young RA, Livingston DM, Parvin JD. BRCA1 is a component of the RNA polymerase II holoenzyme. Proc Natl Acad Sci. 1997;94(11):5605–610.
50. Zhang D, Wang G, Wang Y. Transcriptional regulation prediction of antiestrogen resistance in breast cancer based on RNA polymerase II binding data. BMC Bioinforma. 2014;15(Suppl 2):10.
51. Miller JK, Shattuck DL, Ingalla EQ, Yen L, Borowsky AD, Young LJ, Cardiff RD, Carraway KL, Sweeney C. Suppression of the negative regulator LRIG1 contributes to ErbB2 overexpression in breast cancer. Cancer Res. 2008;68(20):8286–94.
52. Patel HJ, Ramkissoon SH, Patel PS, Rameshwar P. Transformation of breast cells by truncated neurokinin-1 receptor is secondary to activation by preprotachykinin-A peptides. Proc Natl Acad Sci. 2005;102(48): 17436–41.
53. Lu D, Xiao Z, Wang W, Xu Y, Gao S, Deng L, He W, Yang Y, Guo X, Wang X. Down regulation of CIAPIN1 reverses multidrug resistance in human breast cancer cells by inhibiting MDR1. Molecules. 2012;17(6):7595–611.
54. Song H, Ki SH, Kim SG, Moon A. Activating transcription factor 2 mediates matrix metalloproteinase-2 transcriptional activation induced by p38 in breast epithelial cells. Cancer Sci. 2006;66(21):10487–96.
55. Tang Z, Lin MG, Stowe TR, Chen S, Zhu M, Stearns T, Franco B, Zhong Q. Autophagy promotes primary ciliogenesis by removing OFD1 from centriolar satellites. Nature. 2013;502(7470):254–7.
56. Coopman PJ, Do M, Thompson EW, Mueller SC. Phagocytosis of cross-linked gelatin matrix by human breast carcinoma cells correlates with their invasive capacity. Clin Cancer Res. 1998;4(2):507–15.
57. Kammerer S, Roth RB, Reneland R, Marnellos G, Hoyal CR, Markward NJ, Ebner F, Kiechle M, Schwarz-Boeger U, Griffiths LR, et al. Large-scale association study identifies ICAM gene region as breast and prostate cancer susceptibility locus. Cancer Res. 2004;64(24):8906–10.
58. Kim HC, Choi KC, Choi HK, Kang HB, Kim MJ, Lee YH, Lee OH, Lee J, Kim YJ, Jun W, et al. HDAC3 selectively represses CREB3-mediated transcription and migration of metastatic breast cancer cells. Cell Mol Life Sci. 2010;67(20):3499–510.
59. Wang X, Fredericksen ZS, Vierkant RA, Kosel ML, Pankratz VS, Cerhan JR, Justenhoven C, Brauch H, Olson JE, Couch FJ, et al. Association of genetic variation in mitotic kinases with breast cancer risk. Breast Cancer Res Treat. 2010;119(2):453–62.
60. Daly RJ, Binder MD, Sutherland RL. Overexpression of the Grb2 gene in human breast cancer cell lines. Oncogene. 1994;9(9):2723–7.
61. Yu G, Wei X, Das J, Grimson A, Lipkin S, Clark A, Yu H. Dissecting disease inheritance modes in a three-dimensional protein network challenges the guilt-by-association principle. Am J Hum Genet. 2013;93(1):78–89.
62. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005;365(9460):671–9.
63. Pawitan Y, Bjöhle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res. 2005;7(6):953–64.
64. Schmidt M, Böhm D, von Törne C, Steiner E, Puhl A, Pilch H, Lehr HA, Hengstler JG, Kölbl H, Gehrmann M. The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res. 2008;68(13):5405–13.
65. Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Ellis P, Harris A, Bergh J, Foekens JA, et al. Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. J Clin Oncol. 2007;25(10):1239–46.
66. Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, Lindahl T, Pawitan Y, Hall P, Nordgren H, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. 2006;66(21):10292–301.
67. Zhang Y, Sieuwerts AM, McGreevy M, Casey G, Cufer T, Paradiso A, Harbeck N, Span PN, Hicks DG, Crowe J, et al. The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. Breast Cancer Res Treat. 2009;116(2):303–9.