Genet. Res., Camb. (2012), **94**, *pp.* 205–221.
doi:10.1017/S0016672312000419

Gene network-based cancer prognosis analysis with sparse boosting

SHUANGGE MA[1]*, YUAN HUANG[2], JIAN HUANG[3] AND KUANGNAN FANG[4]
[1] *School of Public Health, Yale University, New Haven, CT 06520, USA*
[2] *Department of Statistics, Penn State University, University Park, PA 16802, USA*
[3] *Departments of Statistics and Actuarial Science and Biostatistics, University of Iowa, Iowa City, IA 52242, USA*
[4] *Department of Statistics, Xiamen University, Xiamen, China*

(*Received 22 April 2012; revised 10 June 2012; accepted 19 June 2012*)

Summary

High-throughput gene profiling studies have been extensively conducted, searching for markers associated with cancer development and progression. In this study, we analyse cancer prognosis studies with right censored survival responses. With gene expression data, we adopt the weighted gene co-expression network analysis (WGCNA) to describe the interplay among genes. In network analysis, nodes represent genes. There are subsets of nodes, called modules, which are tightly connected to each other. Genes within the same modules tend to have co-regulated biological functions. For cancer prognosis data with gene expression measurements, our goal is to identify cancer markers, while properly accounting for the network module structure. A two-step sparse boosting approach, called Network Sparse Boosting (NSBoost), is proposed for marker selection. In the first step, for each module separately, we use a sparse boosting approach for within-module marker selection and construct module-level 'super markers'. In the second step, we use the super markers to represent the effects of all genes within the same modules and conduct module-level selection using a sparse boosting approach. Simulation study shows that NSBoost can more accurately identify cancer-associated genes and modules than alternatives. In the analysis of breast cancer and lymphoma prognosis studies, NSBoost identifies genes with important biological implications. It outperforms alternatives including the boosting and penalization approaches by identifying a smaller number of genes/modules and/or having better prediction performance.

1. Introduction

High-throughput gene expression profiling studies have been extensively conducted, searching for markers associated with the development and progression of cancer. In this study, we analyse cancer prognosis studies, where the outcome variables are progression-free, overall, or other types of survival. In many cancer gene expression studies especially the early ones, it has been assumed that genes have interchangeable effects (Knudsen, 2006). Biomedical studies have shown that there exists inherent coordination among genes and, essentially, all biological functions of living cells are carried out through the coordinated effects of multiple genes. There are multiple ways of describing the interplay among genes. The most popular ways are gene pathways and networks (Casci, 2010). Compared with pathway analysis, network analysis sometimes can be more informative as it describes not only whether two genes are connected but also the strength of connection. In addition, some network analysis methods can analyse all genes, whereas many pathway analysis methods focus on the annotated genes only. On the negative side, unlike with pathways, research linking specific network structures with biological functions remains scarce. In the literature, there is no definitive evidence on the relative performance of pathway and network analysis methods. Here, we focus on developing a network analysis method and refer to other studies for discussions and comparisons of pathway and network analyses.

In network analysis, nodes represent genes. Nodes are connected if the corresponding genes have co-regulated biological functions or correlated expressions. There are multiple ways of constructing gene networks. For example, directed, biological networks can be constructed based on the results

* Corresponding author: School of Public Health, Yale University, New Haven, CT 06520, USA. Tel: 203-785-3119. Fax: 203-785-6912. E-mail: shuangge.ma@yale.edu

of knockout experiments. The weighted gene co-expression network analysis (WGCNA: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/), which is adopted in this study, is based only on gene expression data and does not demand additional experiments. There are multiple model-based approaches, using the Akaike information criterion (AIC), multi-model inference (MMI), Bayesian model selection and averaging or minimum description length (MDL) as the network construction criteria. Friedman *et al.* (2008) proposed network construction using a penalization approach. More recently, Maathuis *et al.* (2010) investigated directed networks when the biological information is partially available. A sparse singular value decomposition-based method has also been proposed for network construction (Jornsten *et al.*, 2011). In addition, multiple approaches have been proposed to compute the connectedness measure between pairs of genes. See for example Langfelder & Horvath (2007), Saris *et al.* (2009) and references therein. Published studies have suggested that the network connectedness measure may have important implications. For example, hub genes, which are genes 'well connected' with a large number of genes, tend to have important biological functions. There are subsets of nodes, called modules, which are tightly connected to each other. Genes within the same modules tend to have coordinated biological functions, whereas genes in different modules tend to have different, unrelated biological functions.

Statistical methods that can accommodate the high dimensionality of cancer gene expression data can be roughly classified as dimension reduction and variable selection methods. Both families of methods have been employed in network analysis. Dimension reduction methods search for linear combinations of all genes or genes within the same modules as cancer markers. In Ma *et al.* (2011), principal component analysis is used for network-based dimension reduction. Such methods may have satisfactory prediction performance but often suffer a lack of interpretability. In addition, they contradict the fact that not all genes are involved in cancer development and progression. Variable selection methods search for a subset of genes as markers and may be more interpretable. A network thresholding regularization method is proposed in Ma *et al.* (2010*b*). Huang *et al.* (2011*a*) proposed a penalization method for network variable selection (see references therein for more penalization network analyses). In this article, we focus on the development of a network variable selection method and refer to other publications for comprehensive discussions and comparisons of dimension reduction and variable selection methods.

The rest of the article is organized as follows. In Section 2, we first describe the WGCNA. We describe prognosis using an accelerated failure time (AFT) model and adopt a weighted least squares estimation approach. We then develop the NSBoost approach for gene selection. Simulation study in Section 3 demonstrates satisfactory performance of the proposed approach. Four cancer prognosis studies are analysed in Section 4. The article concludes with discussion in Section 5. Some additional analysis results are provided in Appendices.

2. Methods

(i) *Network construction*

As described in Section 1, there are multiple ways of building gene networks. They can be roughly classified as biological and statistical constructions. Different statistical construction methods rely on different, usually unverifiable data and model assumptions. To the best of our knowledge, in the literature there is still a lack of definitive evidence on the relative performance of different network construction methods. The WGCNA approach is built on the understanding that the coordinated co-expressions of genes encode interacting proteins with closely related biological functions and cellular processes. Detailed investigations of WGCNA have been conducted by Dr Steve Horvath and his group at UCLA. Published studies suggest that modules in the weighted co-expression network have important biological implications. Genes with a higher connectivity are more likely to be involved in important molecular processes. In addition, incorporating connectivity in the detection of differentially expressed genes can lead to significantly improved reproducibility. For integrity of this study, we describe the WGCNA algorithm below and refer to (WGCNA) for more details.

1. Assume that there are d genes. For genes k and $j(=1, \ldots, d)$, compute $\mathrm{cor}(k,j)$, the Pearson correlation coefficient of their expressions. Compute the similarity measure $S(k,j)=|\mathrm{cor}(k,j)|$.
2. Compute the adjacency function $a_{k,j}=S^b(k,j)$, where the adjacency parameter b is chosen using the scale-free topology criterion. In our data analysis, we find that $b=6$, which has been suggested in several published studies, lead to satisfactory results.
3. For gene k, compute its connectivity $C_k=\sum_u a_{k,u}$.
4. For gene $k(=1, \ldots, d)$, compute the topological overlap-based dissimilarity measure $d_{k,j}=1-w_{k,j}$, where $w_{k,j}=(l_{k,j}+a_{k,j})/(\min(C_k,C_j)+1-a_{k,j})$ and $l_{k,j}=\sum_u a_{k,u}a_{j,u}$. Define the dissimilarity matrix D, whose (k,j)th element is $d_{k,j}$.
5. Identify network modules using matrix D and the hierarchical clustering approach. Apply the dynamic tree cut approach (Langfelder *et al.*, 2008) to cut the clustering tree (dendrogram), and identify

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

the resulting branches as modules. Denote M as the number of modules and $S(m)$ as the size of module $m(=1, \ldots, M)$.

Several quantities are defined in the above algorithm. In the downstream analysis, we use the 'final product' – modules constructed in Step 5. As can be seen from the algorithm, the construction of WGCNA is computationally simple. A user-friendly R package is available for implementation (http://cran.r-project.org/web/packages/WGCNA/index.html). A significant advantage of WGCNA is that it is completely inferred from gene expression measurements of a single study and hence does not demand a large amount of biological experiments. On the negative side, it is built on the estimated covariance matrix. In a typical cancer gene expression study, with the sample size significantly smaller than the number of genes, the uniform consistency of the covariance matrix estimation is debatable. Thus, unlike some other ways of describing gene interplay (e.g. pathways), the weighted co-expression network structure may vary across studies with comparable setup.

(ii) *Statistical modelling*

Let T_i be the logarithm of survival time and X_i be the d-dimensional gene expressions for the ith subject in a random sample of size n. The AFT model assumes that

$$T_i = \alpha + X_i'\beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

where α is the intercept, $\beta \varepsilon R^d$ is the unknown regression coefficient and ε_i is the random error. Under right censoring, one observation consists of (Y_i, δ_i and X_i), where $Y_i = \min(T_i, C_i)$, C_i is the logarithm of censoring time and $\delta_i = I(T_i \leq C_i)$ is the censoring indicator. In the AFT model, the logarithm transformation can be replaced with another monotone transformation. The log transformation is the most commonly adopted in the literature and generates reasonable results with data analysed in this study. When the distribution of random error is known, the parametric likelihood function can be easily constructed, and likelihood-based approaches are more efficient than the one described below. Here, we consider the more flexible case with an unknown random error distribution.

The AFT model provides a flexible alternative to the Cox proportional hazards model (Wei, 1992). It assumes a linear function for the log-transformed survival time and may provide a more straightforward description of gene effects on survival than alternatives (e.g. the Cox model, which describes the survival hazard). There are multiple approaches for estimating the AFT model with an unspecified error distribution. Examples include the Buckley–James estimator, which adjusts censored observations using the Kaplan–Meier estimator, and the rank-based estimator, which is motivated by the score function of the partial likelihood function. With high-dimensional gene expression data, those approaches suffer a high computational cost. A computationally more feasible approach is the weighted least squares approach (Stute, 1993). Denote F_n as the Kaplan–Meier estimator of F, the distribution function of T. It can be computed as $F_n(y) = \sum_{i=1}^{n} w_i I(Y_{(i)} \leq y)$. Here, w_i/s are the jumps in the Kaplan–Meier estimator computed as $w_1 = \delta_{(1)}/n$ and $w_i = (\delta_{(i)}/(n-i+1)) \prod_{j=1}^{i-1} ((n-j)/(n-j+1))^{\delta_{(j)}}$, $i = 2, \ldots, n$. w_i/s have also been referred to as the Kaplan–Meier weights (Stute, 1993). $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the order statistics of Y_i/s, $\delta_{(1)}, \ldots, \delta_{(n)}$ are the associated censoring indicators, and $X_{(1)}, \ldots, X_{(n)}$ are the associated gene expressions. The weighted least squares loss function is

$$\frac{1}{2} \sum_{i=1}^{n} w_i (Y_{(i)} - \alpha - X_{(i)}'\beta)^2.$$

We centre $X_{(i)}$ and $Y_{(i)}$ using their corresponding w_i-weighted means, respectively. Let $\bar{X}_w = \sum_{i=1}^{n} w_i X_{(i)} / \sum_{i=1}^{n} w_i$ and $\bar{Y}_w = \sum_{i=1}^{n} w_i Y_{(i)} / \sum_{i=1}^{n} w_i$. Denote $X^*_{(i)} = w_i^{1/2}(X_{(i)} - \bar{X}_w)$ and $Y^*_{(i)} = w_i^{1/2}(Y_{(i)} - \bar{Y}_w)$. We can rewrite the weighted least squares loss function as

$$l(\beta) = \frac{1}{2} \sum_{i=1}^{n} (Y^*_{(i)} - X^{*'}_{(i)}\beta)^2.$$

The simple form of this loss function makes it computationally affordable and suitable for high-dimensional gene expression data.

(iii) *Gene selection with NSBoost*

The proposed NSBoost approach belongs to the family of boosting approaches. Boosting assembles a strong learner using a set of weak learners (Anjum *et al*., 2009; Hastie *et al*., 2009; Ma *et al*., 2012; Schapire & Freund, 2012). It is appropriate for cancer genomic data as individual genes usually have weak effects, but combined together, they may have strong effects. NSBoost is a variable selection approach and thus can have better interpretability than dimension reduction approaches. Compared with thresholding regularization, it has a better defined statistical framework. Compared with penalization, it may have lower computational cost.

(a) *Rationale*

With NSBoost, marker selection is achieved in two steps. This two-step strategy shares a similar spirit with that in Ma *et al*. (2007). With WGCNA, genes can be separated into non-overlapping modules (note that the proposed approach is also applicable to network construction methods with overlapping

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

modules). In the first step, each module is analysed separately. Genes within different modules tend to have different biological functions. Thus, it is sensible to analyse each module separately in the sense that different biological functionalities should be considered separately. On the other hand, genes within the same modules never have identical biological functions. Thus, we propose conducting within-module selection and search for genes that are associated with cancer prognosis within a group of functionally related genes. For a specific module, this step of selection can not only remove noises but also lead to the construction of a *super marker*, which is a linear combination of selected genes and can represent effects of all genes within this module. The introduction of super marker shares a similar spirit with that in Ma *et al.* (2011). In the second step, we consider the joint effects of all super markers and hence all modules. In whole-genome studies, it is reasonable to expect that only a subset of modules is cancer associated. It is thus necessary to conduct the second step of selection and discriminate cancer-associated modules from noises. With the proposed two-step approach, we may identify which modules are cancer-associated as well as which genes are cancer-associated within selected modules.

In both steps, marker selection is achieved using a sparse boosting approach. In high-dimensional data analysis, boosting may be preferred because of its low-computational cost, flexibility and satisfactory empirical performance. With ordinary boosting, when the stopping rule is properly chosen, the resulted strong learner may enjoy a certain degree of sparsity, and so marker selection can be achieved. This can be seen from Dettling & Buhlmann (2003) and follow-up studies as well as our numerical study. However, recent studies (Buhlmann & Yu, 2006; Huang *et al.*, 2011*b*) suggest that with high-dimensional data, ordinary boosting may not be 'sparse enough'. That is, it may identify a considerable number of false positives. The sparse boosting approach adopted here has been motivated by Buhlmann & Yu (2006). In particular, the objective function used for boosting and stopping has two terms. The first term measures goodness-of-fit, which is the same as ordinary boosting. The second, additional term measures model complexity. In particular, we adopt a Bayesian information criterion (BIC) for model complexity measure. As ordinary boosting only considers goodness-of-fit, it may introduce noisy variables (false positives) that happen to be able to slightly improve goodness-of-fit. With sparse boosting, the introduction of the model complexity measure can lead to sparser models and hence reduce the number of false positives. On the negative side, sparse boosting can be computationally more expensive than ordinary boosting as the model complexity measure and hence the whole objective function is not differentiable and cannot be minimized using gradient-based approaches. The sparse boosting approach adopted in this study differs from those in Buhlmann & Yu (2006) and Huang *et al.* (2011*b*). Particularly, previous studies focus on continuous and categorical data, whereas we analyse censored survival data, which can be more complicated. The adopted BIC has been more commonly adopted as a model complexity measure than the MDL. In addition, by conducting multi-step sparse boosting, the proposed approach can effectively accommodate the network module structure. The detailed algorithm is as follows.

(b) *Algorithm*

We first rearrange gene expressions so that $\beta=(\beta^{1\prime}, \ldots, \beta^{m\prime})'$, where β^m is the length $S(m)$ vector of regression coefficients for all genes within module m. Denote β_j^m as the jth component of β^m and $X_{(i)}^{*m}$as the component of $X_{(i)}^*$that corresponds to β^m.

Step I: Within-module boosting. For $m=1, \ldots, M$, consider the objective function $\frac{1}{2}\sum_{i=1}^n(Y_{(i)}^*-X_{(i)}^{*m\prime}\beta^m)^2$, which is $l(\beta)$ evaluated only on genes within the mth module. This is equivalent to the objective function obtained from fitting an AFT model using only the mth module.

(a) Initialization. Set $k=0$ and $\beta^{m[k]}=0$ (componentwise).

(b) Fit and update. $k=k+1$. Compute $\hat{s}=\arg\min_{1\le s\le S(m)}\arg\min_\gamma\sum_{i=1}^n\frac{1}{2}(Y_{(i)}^*-X_{(i)}^{*m\prime}\beta^{m[k-1]}-\gamma X_{(i),s}^{*m})^2+\log(n)\sum_{1\le s\le S(m)}I(\beta_s^{m[k-1]}+\gamma\neq 0)$. Compute $\hat{\gamma}=\arg\min_\gamma\sum_{i=1}^n\frac{1}{2}(Y_{(i)}^*-X_{(i)}^{*m\prime}\beta^{m[k-1]}-\gamma X_{(i),\hat{s}}^{*m})^2$. Update $\beta_s^{m[k]}=\beta_s^{m[k-1]}$ for $s\neq\hat{s}$ and $\beta_{\hat{s}}^{m[k]}=\beta_{\hat{s}}^{m[k-1]}+\nu\hat{\gamma}$, where ν is the step size. As suggested in Buhlmann & Yu (2006) and references therein, the choice of ν is not critical as long as it is small. In our numerical study, we set $\nu=0.1$. In numerical study, we have experimented with a few other step size values and reached almost identical results.

(c) Iteration. Repeat step (b) for K iterations.

(d) Stopping. At iteration k, compute the objective function $F^m(k)=\sum_{i=1}^n\frac{1}{2}(Y_{(i)}^*-X_{(i)}^{*m\prime}\beta^{m[k]})^2+\log(n)\sum_{1\le s\le S(m)}I(\beta_s^{m[k]}\neq 0)$. Estimate the stopping iteration by $\tilde{k}^m=\arg\min_{1\le k\le K}F^m(k)$. For subject i, define its module m 'super marker' as $Z_{(i)}^m=X^{*m\prime}\beta_{(i)}^{m[\tilde{k}^m]}$.

Step II: Module-wise boosting. Consider the objective function $\frac{1}{2}\sum_{i=1}^n(Y_{(i)}^*-Z'_{(i)}\tau)^2$, where $Z_{(i)}=(Z_{(i)}^1, \ldots, Z_{(i)}^M)'$ and $\tau=(\tau_1, \ldots, \tau_M)'$ is the unknown regression

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

Table 1. *Simulation study: median (SD) of the number of identified genes (T) and true positives (TP) computed over 200 replicates. Under each scenario, the first (second) row contains the summary statistics for gene (module) identification. Correlation structure: auto-regressive (auto), banded, and compound symmetry (comp)*

		Enet		Boost		SBoost		NBoost		NSBoost	
Corr.	Rho	T	TP	T	TP	T	TP	T	TP	T	TP
True positives: 4 modules, 20 genes											
Auto	0·3	32 (5·3)	20 (0)	37 (8·1)	20 (0)	32 (2·4)	19 (0·6)	98 (12·5)	15 (1·7)	24 (3·8)	20 (1·7)
		7 (1·2)	4 (0)	8 (2·0)	4 (0)	8 (2·3)	4 (0)	6 (1·6)	3 (1·1)	5 (0·6)	4 (0)
	0·7	22 (2·7)	20 (0)	36 (4·6)	20 (0)	23 (1·6)	17 (1·2)	84 (7·9)	14 (1·8)	24 (3·8)	20 (1·2)
		5 (0·7)	4 (0)	6 (1·2)	4 (0)	5 (0·9)	4 (1·3)	7 (2·6)	4 (1·6)	4 (0·9)	4 (0)
Banded	0·2	23 (1·7)	20 (0)	34 (4·9)	20 (0·3)	22 (2·3)	16 (2·6)	76 (8·6)	11 (1·8)	22 (3·4)	19 (0·9)
		5 (0·9)	4 (0)	6 (1·3)	4 (0)	5 (0·7)	4 (1·2)	7 (2·0)	3 (1·2)	4 (0·7)	4 (0)
	0·33	25 (3·2)	20 (0)	37 (5·6)	20 (0)	24 (2·2)	16 (1·9)	86 (10·8)	16 (1·9)	23 (5·4)	20 (1·3)
		5 (1·7)	4 (0)	7 (1·4)	4 (0)	5 (1·0)	4 (0·9)	9 (2·1)	4 (1·1)	4 (0·6)	4 (0)
Comp	0·3	33 (5·5)	20 (0)	47 (6·9)	20 (0·3)	26 (2·3)	15 (1·7)	96 (8·3)	14 (1·9)	29 (5·1)	20 (1·6)
		7 (1·4)	4 (0)	9 (2·4)	4 (0)	6 (1·2)	3 (1·1)	11 (2·9)	3 (1·3)	5 (1·1)	4 (0)
	0·7	35 (4·6)	20 (0)	44 (5·2)	20 (0·9)	18 (2·8)	11 (3·8)	80 (8·0)	11 (2·2)	22 (2·7)	17 (1·2)
		7 (1·9)	4 (0)	8 (2·2)	4 (0)	6 (1·9)	3 (1·0)	10 (2·7)	3 (0·9)	4 (1·6)	4 (0)
True positives: 2 modules, 10 genes											
Auto	0·3	18 (2·4)	10 (0)	20 (3·9)	10 (0)	17 (1·4)	10 (0·4)	44 (5·1)	7 (1·0)	14 (1·6)	10 (0·7)
		4 (0·8)	2 (0)	4 (0·9)	2 (0)	4 (1·0)	2 (0)	4 (1·0)	2 (0·7)	3 (0·4)	2 (0)
	0·7	13 (1·5)	10 (0)	20 (2·2)	10 (0)	13 (0·9)	9 (0·9)	39 (4·2)	7 (1·0)	13 (1·5)	10 (0·8)
		3 (0·4)	2 (0)	3 (0·6)	2 (0)	3 (0·6)	2 (0·8)	4 (1·4)	2 (0·6)	2 (0·5)	2 (0)
Banded	0·2	13 (1·0)	10 (0)	16 (2·6)	10 (0·2)	13 (1·2)	9 (1·1)	41 (3·9)	8 (0·8)	14 (1·8)	9 (0·4)
		3 (0·6)	2 (0)	3 (0·9)	2 (0)	3 (0·6)	2 (0·4)	5 (0·9)	2 (0·5)	2 (0·3)	2 (0)
	0·33	16 (1·4)	10 (0)	21 (3·0)	10 (0)	16 (1·6)	5 (1·0)	38 (4·9)	8 (1·0)	15 (2·3)	10 (0·5)
		3 (1·0)	2 (0)	4 (0·8)	2 (0)	3 (0·6)	2 (0·3)	4 (1·2)	2 (0·5)	2 (0·3)	2 (0)
Comp	0·3	18 (2·5)	10 (0)	23 (4·0)	10 (0·1)	13 (1·4)	8 (1·1)	50 (4·4)	6 (1·0)	17 (2·0)	10 (0·8)
		4 (0·8)	2 (0)	5 (1·1)	2 (0)	3 (0·7)	2 (0·6)	6 (1·5)	2 (0·8)	3 (0·6)	2 (0)
	0·7	17 (2·3)	10 (0)	24 (2·6)	10 (0·6)	10 (1·5)	6 (1·4)	41 (4·2)	7 (1·2)	13 (1·7)	8 (0·8)
		4 (1·3)	2 (0)	4 (1·5)	2 (0)	3 (1·1)	2 (0·6)	6 (1·2)	2 (0·4)	2 (0·6)	2 (0)

coefficient. That is, in the least squares objective function, we use the super markers, which represent the effects of all genes within the same modules, to replace the original gene expressions.

(a) Initialization. Set $k=0$ and $\tau^{[k]}=0$ (component-wise).

(b) Fit and update $k=k+1$. Compute $\hat{s}=\arg\min_{1\le s\le M}\arg\min_{\gamma}\sum_{i=1}^{n}\frac{1}{2}(Y^{*}_{(i)}-Z_{(i)}'\tau^{[k-1]}-\gamma Z_{(i),s})^{2}+\log(n)\sum_{1\le s\le M}I(\tau_{s}^{[k-1]}+\gamma\neq 0)$. Compute $\hat{\gamma}=\arg\min_{\gamma}\sum_{i=1}^{n}\frac{1}{2}(Y^{*}_{(i)}-Z_{(i)}'\tau^{[k-1]}-\gamma Z_{(i),\hat{s}})^{2}$. Update $\tau_{s}^{[k]}=\tau_{s}^{[k-1]}$ for $s\neq\hat{s}$ and $\tau_{\hat{s}}^{[k]}=\tau_{\hat{s}}^{[k-1]}+v\hat{\gamma}$ where $v=0.1$ is the step size.

(c) Iteration. Repeat Step (b) for K iterations.

(d) Stopping. At iteration k, compute the objective function $F(k)=\sum_{i=1}^{n}(Y^{*}_{(i)}-Z'_{(i)}\tau^{[k]})^{2}+\log(n)\sum_{1\le s\le M}I(\tau_{s}^{[k]}\neq 0)$. Estimate the stopping iteration by $\hat{k}=\arg\min_{1\le k\le K}F(k)$.

$\sum_{m=1}^{M}\tau_{m}^{\hat{k}}Z_{(i)}^{m}=\sum_{m=1}^{M}\tau_{m}^{\hat{k}}\{X^{*m\prime}_{(i)}\beta^{m[\tilde{k}^{m}]}\}$ is the resulted strong learner for $Y^{*}_{(i)}$. Genes and modules with non-zero regression coefficients in the strong learner are identified as associated with cancer.

(c) *Parameter path*

Parameter path, which is the graphical presentation of the estimates as a function of number of iterations, may provide further insights into NSBoost. Consider the simulation setting corresponding to row 1 of Table 1. For a better view, we simplify the simulation setting and consider four modules with four genes per module. The first two modules are cancer associated, within which there are two cancer-associated genes. Thus, among the 16 simulated genes, four are associated with cancer. For comparison, we also study Network Boosting (NBoost, details described in Section 3). For a randomly generated dataset, the parameter paths are shown in Figs 1 (NSBoost) and 2 (NBoost), respectively.

Within each module, the parameter paths of NSBoost are similar to those of other regularized variable selection approaches (Hastie *et al.*, 2009). By considering model complexity in boosting, the NSBoost parameter paths are 'smoother' than their NBoost counterparts. NBoost does not consider model complexity in boosting and thus may have a

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

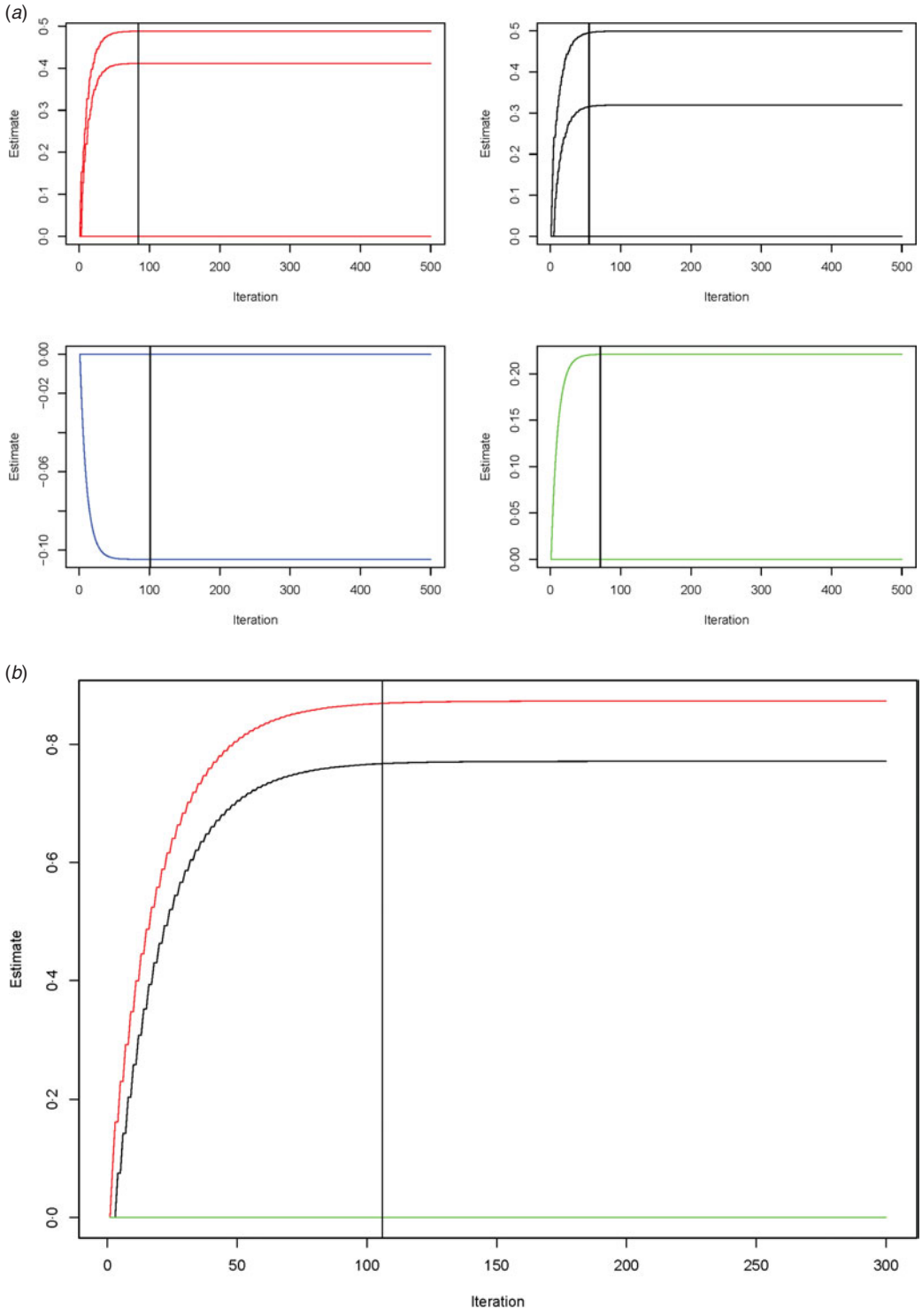

Fig. 1. Parameter path of NSBoost: estimates as a function of number of iterations. (a) The four panels correspond to four modules in Step 1 of boosting. (b) The panel corresponds to four super markers in Step 2 of boosting. Vertical lines correspond to the selected number of iterations.

risk of false positives. For example, in the top right panel of Fig. 2*a*, NBoost has one false positive, while NSBoost does not. Our limited numerical study suggests that, in the within-module boosting step, NSBoost may identify 'signals' even with purely noisy modules. Thus, the module-level boosting is conducted, which can effectively remove noisy modules as a whole (Fig. 1*b*). With a combination of the two boosting steps, NSBoost can be sparser than NBoost at both within-module level and module

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

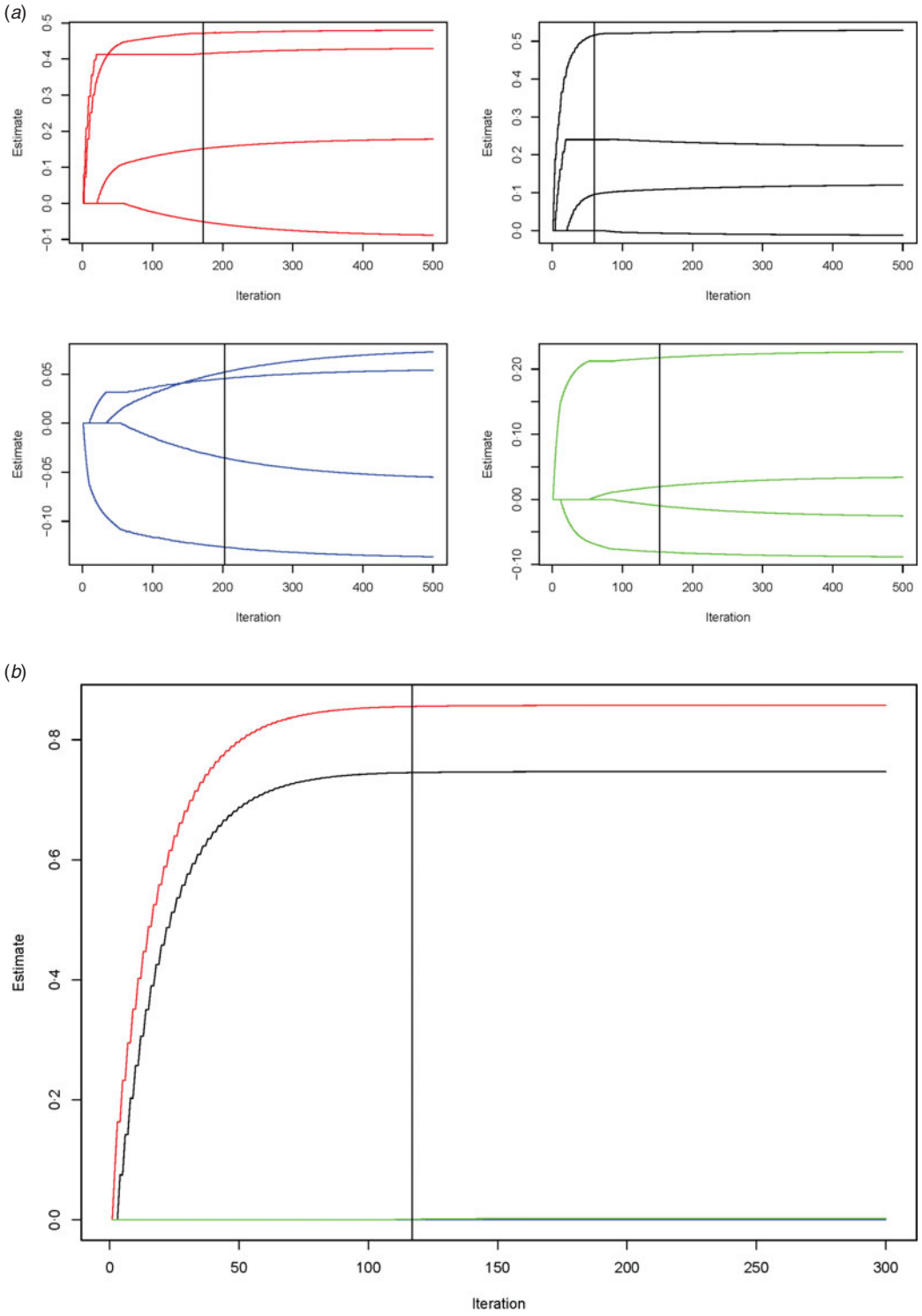

Fig. 2. Parameter path of NBoost: estimates as a function of number of iterations. (a) The four panels correspond to four modules in Step 1 of boosting. (b) The panel corresponds to four super markers in Step 2 of boosting. Vertical lines correspond to the selected number of iterations.

level. We note that the parameter paths are presented for a small dataset and are only meant to provide a graphical presentation. More meaningful comparisons with larger datasets are presented in Sections 3 and 4.

3. Simulation

We conduct simulation to better gauge properties of the proposed approach. In each simulated dataset, there are 100 subjects. We simulate 50 gene clusters

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

Table 2. *Description of datasets. Gene: number of genes profiled*

Data	Disease	Platform	Gene	Sample
D1: Sorlie *et al.* (2001)	Breast cancer	cDNA	8102	58
D2: Huang *et al.* (2003)	Breast cancer	Affymetrix	12 625	71
D3: Rosenwald *et al.* (2002)	DLBCL	cDNA	7399	240
D4: Rosenwald *et al.* (2003)	MCL	cDNA	8810	92

with 20 genes in each cluster. Gene expressions have marginally standard normal distributions. Genes within different clusters have independent expressions. For genes within the same clusters, their expressions have the following correlation structures: (1) auto-regressive correlation, where expressions of genes j and k have correlation coefficient $\rho^{|j-k|}$. $\rho=0{\cdot}3$ or 0·7, corresponding to weak and strong correlations; (2) banded correlation, where expressions of genes j and k have correlation coefficient $\max(0,1-|j-k|\rho)$. $\rho=0{\cdot}2$ or 0·33; and (3) compound symmetry, where expressions of genes j and k have correlation coefficient ρ when $j\neq k$, $\rho=0{\cdot}3$ or 0·7. With each simulated dataset, we generate network modules using WGCNA. We find that when the within-cluster correlation is strong, the resulted modules tend to be correlated with the simulated clusters. On the other hand, when the within-cluster correlation is weak, there are considerable discrepancies between the WGCNA modules and simulated clusters. We consider two scenarios for the prognosis-associated genes. Within each of the first four (or two) modules, the first five genes are associated with survival. There are thus a total of 20 (or 10) cancer-associated genes, and the rest are noises. For cancer-associated genes, we generate their regression coefficients from Unif[0·5, 1·5]. Thus, some genes have large effects, and others have moderate to small effects. For a subject, we generate the logarithm of survival time from the AFT model with intercept equal to zero. The logarithm of censoring time is independently generated from a normal distribution with variance one. We adjust the mean of the censoring distribution by trial and error so that the average censoring rate is about 40%. To better gauge performance of the proposed approach, we also consider the following alternatives:

1. Enet (elastic net) (Zou & Hastie, 2005), which is a penalization approach and has been extensively used in the analysis of gene expression data.
2. Boost, which is an ordinary boosting approach and takes goodness-of-fit as the only criterion for choosing weaker learners. A BIC similar to that with NSBoost is adopted for stopping.
3. SBoost, which is a sparse boosting approach and considers goodness-of-fit as well as model complexity measured using a BIC in boosting and stopping.
4. NBoost, which is a network boosting approach and has a two-step algorithm similar to that with the proposed approach. The difference is that in boosting only goodness-of-fit is considered when choosing weaker learners.

Among the four alternative approaches, the first three ignore the network structure and treat all gene effects as interchangeable. The NBoost approach respects the network structure, however, puts less emphasis on sparsity. We are aware that a large number of approaches can be used to analyse the simulated data. The above four approaches are chosen for comparison, as Enet is one of the most extensively used penalization approaches and particularly includes Lasso and ridge penalization as special cases and, as the Boost, SBoost and NBoost approaches have boosting frameworks closest to that of NSBoost.

Summary statistics, including medians and standard deviations, based on 200 replicates are presented in Table 1. We can see that the Enet and Boost approaches can identify all of the true positives. However, under some scenarios, they may identify a considerable number of false positives. SBoost, which considers model complexity in boosting but ignores the network structure, is 'overly sparse' by having a considerable number of false negatives. Without accounting for model complexity in boosting, NBoost identifies a large number of false positives. Under all simulated scenarios, NSBoost has the best performance in terms of module and gene identification accuracy. It is capable of identifying the majority or all of the true positives while having a small number of false positives. We have also experimented with a few other simulation settings and reached similar conclusions.

4. Data analysis

We collect four cancer prognosis studies with gene expression measurements. Brief descriptions are provided in Table 2 and below. We refer to the original publications for more details.

D1. Breast cancer is the second leading cause of cancer death among women in the United States. Sorlie *et al.* (2001) conducted a gene expression profiling study, investigating whether it was feasible to classify breast carcinomas based on gene expression patterns. cDNA profiling of 85 samples was

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

conducted, showing that breast cancer could be classified into a basal epithelial-like group, an ERBB2-overexpressing group, and a normal breast-like group. Among the 85 samples, 58 had survival information available and will be analysed in this study.

D2. Despite major progress in treatment, the ability to predict metastasis of breast tumours remains limited. Huang *et al*. (2003) reported a study investigating metastatic states and relapses in breast cancer patients. Affymetrix genechips were used for the profiling of 71 samples. Both D1 and D2 are on breast cancer prognosis. However, as suggested in multiple publications great heterogeneity may exist across studies, we choose to analyse D1 and D2 separately.

D3. Diffuse large B-cell lymphoma (DLBCL) is a cancer of the B-cell. It accounts for $\sim$40% of all non-Hodgkin lymphoma (NHL) cases. A DLBCL gene expression study was reported in Rosenwald *et al*. (2002). This study retrospectively collected tumour biopsy specimens and clinical data for 240 patients with untreated DLBCL. The median follow up was 2·8 years, with 138 observed deaths. Lymphochip cDNA microarray was used to measure the expressions of 7399 genes.

D4. Mantle cell lymphoma (MCL) accounts for $\sim$8% of all NHLs. Rosenwald *et al*. (2003) reported a gene expression study of MCL survival. Among 101 untreated patients with no history of previous lymphoma, 92 were classified as having MCL based on morphologic and immunophenotypic criteria. Survival times of 64 patients were available, and the rest were censored. The median survival time was 2·8 years. Lymphochip DNA microarrays were used to quantify mRNA expressions in the lymphoma samples. Gene expression data on 8810 cDNA elements were available.

Among the four studies, one used Affymetrix and three used cDNA for profiling. We process the datasets as follows. We conduct normalization using the lowess approach for cDNA data and the robust multi-array (RMA) approach for Affymetrix data. Missing measurements are imputed using the K-nearest neighbours approach. Affymetrix gene expression measurements are log 2 transformed. We select the 500 genes with the largest variances for downstream analysis. Here, the pre-screening may serve multiple purposes. First in cancer gene expression studies, usually genes with higher variations are of more interest. Second, it is expected that the number of cancer prognosis-associated genes is far smaller than 500. Pre-screening may remove genes that are highly unlikely to be cancer-associated and significantly reduce computational cost. More importantly, as described above, the WGCNA approach involves estimating covariance matrices. Pre-screening may significantly reduce the dimensionality of such matrices and improve estimation accuracy. Note that the number of screened genes can be somewhat subjective. In the pre-screening literature, there is still a lack of guideline on how many genes should be screened. With the selected genes, we normalize their expressions to have median zero and variance one.

With datasets D1–D4, the WGCNA approach constructs 5, 4, 6 and 6 modules, respectively. For dataset D4, we show the module construction result in Fig. A.1 (Appendix I). Results for other datasets are available with the authors.

We apply the NSBoost approach. The identified genes and corresponding estimates are provided in Appendix II. Searching PubMed suggests that some of the identified genes have been suggested as cancer markers in published studies. Detailed interpretations of the identified genes are provided in Appendix III. Note that here we investigate the implications of identified genes individually. With the proposed approach, we conduct selection at the module level as well as the individual gene level. Thus, gene level interpretation is meaningful. In addition, research that links network modules to specific biological functions is extremely limited. As there is no strong correspondence between network modules and pathways, pathway analysis may not be very sensible. Evaluation of the biological implications deserves more investigation in future research.

(i) *Analysis with alternative approaches*

We apply the four alternative approaches described in the Simulation section. Summary analysis results for all approaches are presented in Table 3. By conducting the module-level sparse boosting and hence encouraging sparsity at the module level, NSBoost identifies the smallest number of modules, which may lead to more focused hypotheses for downstream analysis. Genes identified by NSBoost differ significantly from those identified using Enet, Boost and SBoost. For example for dataset D1, the numbers of overlapped genes are 4, 5 and 3, respectively. The sets of genes identified by NBoost and NSBoost are more similar, which is as expected, as both approaches use boosting for marker selection and account for the module structure. For example for dataset D1, the number of overlapped genes is 23. Although discussions in Appendix III show that the NSBoost identified genes are biologically meaningful, with our limited understanding of cancer genomics, we are unable to determine whether they are 'more meaningful' than the other sets of identified genes. As an alternative, we examine the prediction performance of different approaches, which proceeds as follows: (1) randomly split data into a training set and a testing set with sizes 3 : 1; (2) analyse the training data and identify markers. A natural by-product of the

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

Table 3. *Data analysis results*

		D1	D2	D3	D4
Enet	Gene	29	39	82	60
	Overlap	2	3	6	0
	Module	4	3	5	2
	Logrank	0·089	8·931	3·405	5·629
Boost	Gene	70	74	17	12
	Overlap	5	4	1	0
	Module	4	4	3	4
	Logrank	1·704	2·478	1·642	7·976
SBoost	Gene	31	26	22	12
	Overlap	3	1	1	0
	Module	3	2	4	2
	Logrank	0·063	0·128	5·961	6·662
NBoost	Gene	102	91	44	35
	Overlap	23	21	13	14
	Module	3	2	5	2
	Logrank	0·266	0·318	8·996	17·015
NSBoost	Gene	31	30	21	22
	Module	2	1	1	1
	Logrank	2·863	11·504	15·613	18·937

Gene, number of genes identified; Overlap, number of genes overlapped with NSBoost; Module, number of modules identified; logrank, prediction logrank statistic.

proposed approach is a prediction model; (3) make prediction for subjects in the testing set. The predictive model can lead to a predicted risk score $X'\beta$ for each subject. Dichotomize the risk scores at the median and create two risk groups. Compute the logrank statistic, which measures the survival difference between the two groups; (4) to avoid an extreme partition, repeat steps (1)–(3) 100 times, and compute the average logrank statistic. Table 3 shows that with the four analysed datasets, NSBoost has the largest logrank statistics and hence the best performance in separating subjects into groups with different survival risks.

5. Discussion

In cancer genomic studies, an important goal is to identify markers associated with prognosis. There exists inherent coordination among genes, and network provides an effective way of describing such coordination. In this study, we adopt the weighted gene co-expression network and develop a two-step sparse boosting approach to account for the network structure in cancer marker selection. The proposed approach is intuitively reasonable. Simulation and data analysis demonstrate its satisfactory performance.

As shown in multiple published studies, network modules may have important biological implications. The proposed approach respects the network module structure and can be more informative than alternatives that ignore such structure. Another advantage of the proposed approach is its computational affordability. As can be seen from the algorithm, only simple calculations are involved. In the within-module boosting, the number of genes per module can be much smaller than the total number of genes. In addition, this step can be carried out in a parallel manner. Thus, the first step of boosting has computational cost much smaller than that of ordinary boosting with all genes. With WGCNA, the number of modules (and hence super markers) is usually not large. Numerical studies in Ma *et al.* (2010*b*, 2011) suggest less than 20. Thus, the computational cost of the second step of boosting is almost negligible. The proposed approach also has the advantage that its applicability is relatively 'independent' of the model setup and network construction procedure. It is applicable to other survival models and other types of data, for example diagnosis studies with categorical response variables and generalized linear models, with very minor modifications.

As described in Section 1, there are multiple ways of describing the interplay among genes. To the best of our knowledge, there is a lack of definitive evidence on the relative performance of different network construction procedures. Our analysis shows that with WGCNA, the proposed NSBoost may improve cancer marker selection. It is possible that in practical data analysis, adopting other network construction methods can further improve prediction and selection. As the focus of this study is on the development of NSBoost, a more comprehensive examination of its performance under different networks is beyond our scope. We adopt the AFT model to describe gene effects on survival. Compared with alternatives such as the Cox model, this model may have more lucid interpretations and lower computational cost. Model diagnostics is not conducted, as there is a lack of diagnostic tools for survival analysis with high-dimensional gene expression data. The satisfactory prediction performance may partly support the validity of this model. NSBoost can effectively account for the 'module-gene' two-level hierarchical structure, which is not the complete information contained in a network. WGCNA and other networks contain other information, for example the connectedness measure between any two genes within the same modules. It may be possible to extend the proposed approach and accommodate the connectedness measure in marker selection. However, as discussed above, with $n=d$, the uniform estimation consistency of $d(d-1)/2$ connectedness measures is questionable. In contrast, the module structure can be much more reliable. Thus, we focus on the module structure in our research. The simulation settings considered

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

in this study are simpler than what is encountered in practical data analysis. We intentionally choose such settings as they may favour simple approaches such as Enet and Boost. In data analysis, we conclude that NSBoost may be preferred as it identifies a smaller number of modules and genes and has better prediction performance. Analysis of independent validation studies may be needed to fully confirm performance of NSBoost and identified markers.

We thank the editor and referees for careful review and insightful comments. This study was supported by awards DMS-0904181 from NSF, CA142774 and CA165923 from NIH, USA and Fujian Social Science Fund (2011C042), China.

Appendix I

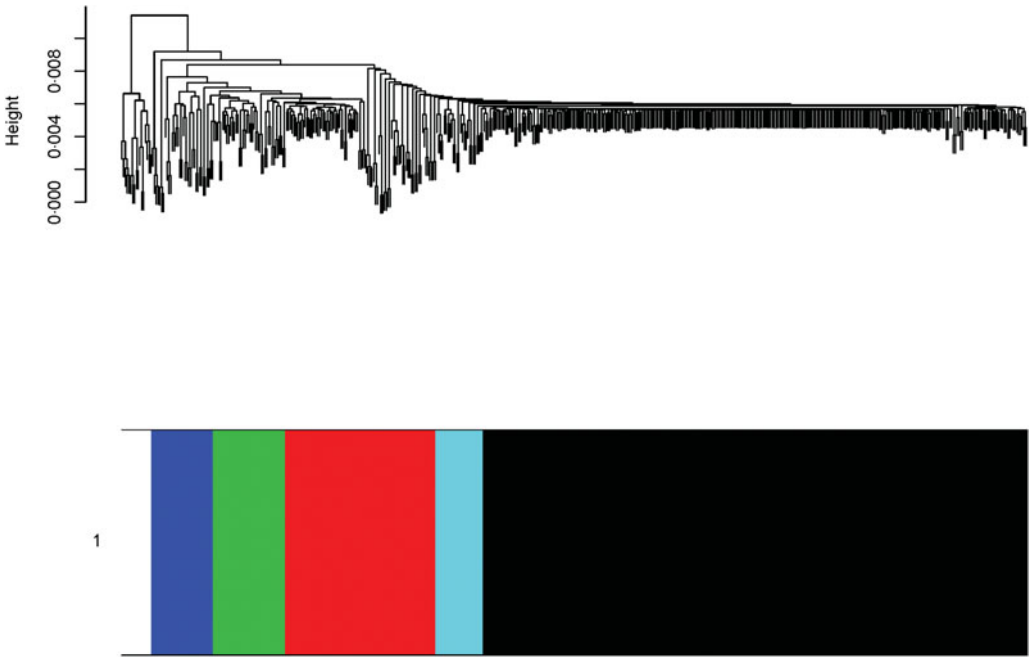

Fig. A.1. Module construction result for dataset D4.

Appendix II. Details on the identified genes and their estimates for dataset D1–D4.

Data D1

Gene ID	Gene Name	Module	Estimate
Hs.154387	Tetratricopeptide repeat domain 35 (TTC35)	1	0·055
Hs.169330	Transgelin 3 (TAGLN3)	1	−0·160
Hs.180946	Family with sequence similarity 69, member A (FAM69A)	1	−0·332
Hs.24734	Oxysterol binding protein (OSBP)	1	0·458
Hs.25351	Iroquois homeobox 5 (IRX5)	1	−0·125
Hs.267632	TATA element modulatory factor 1 (TMF1)	1	−0·147
Hs.2719	WAP four-disulfide core domain 2 (WFDC2)	1	−0·028
Hs.27916	Transcribed locus	1	−0·954
Hs.30743	Preferentially expressed antigen in melanoma (PRAME)	1	0·114
Hs.418506	Insulin-like 4 (placenta) (INSL4)	1	−0·086
Hs.45743	Adenosine A2b receptor (ADORA2B)	1	−0·345
Hs.5344	Adaptor-related protein complex 1, gamma 1 subunit (AP1G1)	1	−0·051
Hs.621	Lectin, galactoside-binding, soluble, 3 (LGALS3)	1	−0·538
Hs.687	Cytochrome P450, family 4, subfamily B, polypeptide 1 (CYP4B1)	1	−0·034
Hs.73793	Vascular endothelial growth factor A (VEGFA)	1	0·057
Hs.74592	SATB homeobox 1 (SATB1)	1	−0·240
Hs.75206	Protein phosphatase 3, catalytic subunit, gamma isozyme (PPP3CC)	1	0·114
Hs.75400	Family with sequence similarity 168, member A (FAM168A)	1	0·481
Hs.78452	Solute carrier family 20 (phosphate transporter), member 1 (SLC20A1)	1	0·340

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

Data D1 *(Cont.)*

Gene ID	Gene Name	Module	Estimate
Hs.80642	Signal transducer and activator of transcription 4 (STAT4)	1	−0·300
Hs.82921	Chromosome 6 open reading frame 165 (C6orf165)	1	0·508
Hs.83347	Angio-associated, migratory cell protein (AAMP)	1	−0·521
Hs.89582	Glutamate receptor, ionotropic, AMPA 2 (GRIA2)	1	0·508
Hs.93913	Interleukin 6 (interferon, beta 2) (IL-6)	1	−0·596
Hs.96063	Insulin receptor substrate 1 (IRS1)	1	−0·287
Hs.166994	FAT tumour suppressor homologue 1 (*Drosophila*) (FAT1)	2	−0·059
Hs.2256	Matrix metallopeptidase 7 (matrilysin, uterine) (MMP7)	2	−0·117
Hs.296634	Ceruloplasmin (ferroxidase) (CP)	2	−0·164
Hs.5716	SEC16 homologue A (*Saccharomyces cerevisiae*) (SEC16A)	2	−0·203
Hs.75275	Ubiquitination factor E4A (UFD2 homologue, yeast) (UBE4A)	2	0·141
Hs.75737	Pericentriolar material 1 (PCM1)	2	0·124

Data D2

Gene ID	Gene name	Estimate
Hs.13321	Rearranged L-myc fusion (RLF)	−1·740
Hs.177584	3-oxoacid CoA transferase 1 (OXCT1)	−0·089
Hs.180610	Splicing factor proline/glutamine-rich (SFPQ)	−0·531
Hs.182626	Transmembrane protein 184B (TMEM184B)	−0·263
Hs.184693	Transcription elongation factor B (SIII), polypeptide 1 (15 kDa, elongin C) (TCEB1)	−0·134
Hs.21595	A kinase (PRKA) anchor protein 17A (AKAP17A)	−1·065
Hs.24594	Ubiquitination factor E4B (UBE4B)	−0·613
Hs.2488	Lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76 kDa) (LCP2)	0·197
Hs.25363	Presenilin 2 (Alzheimer disease 4) (PSEN2)	1·032
Hs.2706	Glutathione peroxidase 4 (phospholipid hydroperoxidase) (GPX4)	−0·276
Hs.282975	Carboxylesterase 2 (CES2)	0·088
Hs.284244	Fibroblast growth factor 2 (basic) (FGF2)	−0·567
Hs.28914	Adenine phosphoribosyltransferase (APRT)	−1·699
Hs.290070	Gelsolin (GSN)	−0·909
Hs.297681	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1 (SERPINA1)	−1·176
Hs.30954	Phosphomevalonate kinase (PMVK)	−0·079
Hs.343586	Zinc finger protein 36, C3H type, homologue (mouse) (ZFP36)	−0·069
Hs.348935	Immunoglobulin lambda-like polypeptide 1 (IGLL1)	−0·586
Hs.406186	Splicing factor 3b, subunit 4, 49 kDa (SF3B4)	0·397
Hs.4980	LIM domain binding 2 (LDB2)	−1·418
Hs.5716	SEC16 homolog A (*S. cerevisiae*) (SEC16A)	−0·076
Hs.75643	Nuclear factor (erythroid-derived 2), 45 kDa (NFE2)	−0·327
Hs.76780	Protein phosphatase 1, regulatory (inhibitor) subunit 1A (PPP1R1A)	0·218
Hs.7912	Neuronal cell adhesion molecule (NRCAM)	−1·076
Hs.79391	Huntingtin (HTT)	−0·307
Hs.84746	Regulator of chromosome condensation 1 (RCC1)	−1·377
Hs.8769	Transmembrane protein 47 (TMEM47)	−0·310
Hs.93183	Vasodilator-stimulated phosphoprotein (VASP)	−1·558
Hs.95821	Osteoclast stimulating factor 1 (OSTF1)	−0·074
Hs.98938	Protocadherin alpha cluster, complex locus (PCDHA)	−0·753

Data D3

Gene name	Estimate
CASP2 and RIPK1 domain containing adaptor with death domain (CRADD)	0·011
Diacylglycerol kinase, delta (130 kDa) (DGKD)	−0·017
Topoisomerase (DNA) II binding protein (TOPBP1)	−0·012
ESTs	−0·017
Surfeit 1	0·021
CDC7 cell division cycle 7-like 1 (*S. cerevisiae*)	−0·042

Data D3 *(Cont.)*

Gene name	Estimate
Hypothetical protein FLJ10509	−0·014
Bromodomain adjacent to zinc finger domain, 1B (BAZ1B)	0·017
Septin 6 (SEPT6)	−0·01
Complement component (3b/4b) receptor 1, including Knops blood group system (CR1)	0·01
Alanyl (membrane) aminopeptidase (ANPEP)	0·016
GRAMD1A GRAM domain containing 1A	−0·054
Osteoblast specific factor 2 (fasciclin I-like) (POSTN)	0·014
Suppression of tumourigenicity 13 (colon carcinoma) (Hsp70 interacting protein) (ST13P4)	−0·019
T-cell receptor delta locus (TRA)	0·01
Myosin, light polypeptide 2, regulatory, cardiac, slow (MYL2)	0·008
Ankyrin 1, erythrocytic (ANK1)	0·014
ESTs	−0·009
LCOR ligand-dependent nuclear receptor corepressor	−0·009
Immunoglobulin superfamily receptor translocation associated 1 (FCRL4)	0·035
Immunoglobulin superfamily receptor translocation associated 1 (FCRL4)	0·011

Data D4

Gene ID	Gene name	Estimate
15870	Interferon-induced protein with tetratricopeptide repeats 1 (IFIT1B)	0·02
15977	SHC (Src homology 2 domain containing) transforming protein 1 (SHC1)	−0·032
16847	Special AT-rich sequence binding protein 1 (binds to nuclear matrix/ scaffold-associating DNAs) (SATB1)	−0·023
17312	Neuroblastoma RAS viral (v-ras) oncogene homologue (NRAS)	−0·03
19261	Inhibitor of DNA binding 2, dominant negative helix-loop-helix protein (ID2)	0·026
24473	Similar to Williams–Beuren syndrome critical region protein 19 (LOC442608)	−0·026
24635	Meningioma expressed antigen 5 (hyaluronidase) (MGEA5)	0·026
26475	Chemokine (C–C motif) ligand 3 (CCL19)	−0·032
27108	CD24 antigen (small cell lung carcinoma cluster 4 antigen)	0·064
27199	Prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase) (PTGS2)	0·024
28027	Activating transcription factor 2	0·066
28973	Tissue inhibitor of metalloproteinase 3 (Sorsby fundus dystrophy, pseudoinflammatory)	−0·026
29286	Haematopoietically expressed homeobox (HHEX)	−0·009
30596	Immunoglobulin superfamily receptor translocation associated 1	0·017
31298	Zinc finger protein 592	0·059
31543	SH3-domain kinase binding protein 1 (SH3KBP1)	0·041
31702	Hypothetical protein FLJ90709	0·022
31731	CDNA FLJ41270 fis, clone BRAMY2036387	0·036
31979	Kelch-like 14 (*Drosophila*) (KLHL14)	0·039
32902	MRNA full-length insert cDNA clone EUROIMAGE 1534000	−0·027
33017	Chromosome 3 open reading frame 14 (C3orf14)	0·045
33424	Chromosome 3 open reading frame 1	0·017

Appendix III. Biological implications of the identified genes

D1. Among the identified genes, gene TTC35 is one of the identified breast cancer markers according to G2SBC (http://www.itb.cnr.it/breastcancer/php/browse.php). Gene FAM69A encodes a member of the FAM69 family of cysteine-rich type II transmembrane proteins. It has been implied in the development of multiple sclerosis and ovarian cancer, suggesting that it may play an essential role in cancer development. It is also involved in the development of mental disorders. Gene IRX5 encodes a member of the iroquois homeobox gene family, which are involved in several embryonic developmental processes. Studies with knockout mice lacking this gene show that it is required for retinal cone bipolar cell differentiation, and that it negatively regulates potassium channel gene expression in the heart to ensure coordinated cardiac repolarization. Gene WFDC2

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

encodes a protein that is a member of the WFDC domain family. This gene is expressed in pulmonary epithelial cells and is also found to be expressed in ovarian cancer, which shares multiple genetic markers with breast cancer. Gene PRAME encodes an antigen that is predominantly expressed in multiple cancer tissues such as melanomas and that is recognized by cytolytic T-lymphocytes. It is not expressed in normal tissues, except testis. Gene INSL4 encodes the insulin-like 4 (INSL4) protein, a member of the insulin superfamily. Its involvement in breast cancer development is proposed in Burger *et al*. (2005). Gene Adenosine A2b receptor (ADORA2B) is over-expressed in cancer tissues under a hypoxic state, promoting cancer cell growth (Ma *et al*., 2010*a*). Gene LGALS3 encodes a member of the galectin family of carbohydrate binding proteins. This protein plays a role in numerous cellular functions including apoptosis, innate immunity, cell adhesion and T-cell regulation. Gene CYP4B1 encodes a member of the cytochrome P450 superfamily of enzymes. In rodents, the homologous protein has been shown to metabolize certain carcinogens. Gene VEGFA is a member of the PDGF/VEGF growth factor family and encodes a protein that is often found as a disulphide-linked homodimer. This protein is a glycosylated mitogen that specifically acts on endothelial cells and has various effects, including mediating increased vascular permeability, inducing angiogenesis, vasculogenesis and endothelial cell growth, promoting cell migration and inhibiting apoptosis. Published studies have suggested gene SATB1 as a marker for breast cancer, gastric cancer and non-small cell lung cancer. The protein encoded by gene SLC20A1 is a sodium-phosphate symporter that absorbs phosphate from interstitial fluid for use in cellular functions such as metabolism, signal transduction and nucleic acid and lipid synthesis. The encoded protein is also a retroviral receptor, causing human cells to be susceptible to infection by gibbon ape leukaemia virus, simian sarcoma-associated virus, feline leukaemia virus subgroup B, and 10A1 murine leukaemia virus. The protein encoded by gene STAT4 is a member of the STAT family of transcription factors. In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor-associated kinases, and then form homo- or heterodimers that translocate to the cell nucleus where they act as transcription activators. Gene angio-associated migratory cell protein (AAMP) is found to be expressed strongly in endothelial cells, cytotrophoblasts and poorly differentiated colon adenocarcinoma cells found in lymphatics. Gene IL-6 encodes a cytokine that functions in inflammation and the maturation of B-cells. The functioning of this gene is implicated in a wide variety of inflammation-associated disease states, including susceptibility to diabetes mellitus and systemic juvenile rheumatoid arthritis. Gene FAT1 is an orthologue of the *Drosophila* fat gene, which encodes a tumour suppressor essential for controlling cell proliferation. Its product functions as an adhesion molecule and/or signalling receptor, and is likely to be important in developmental processes and cell communication. Proteins of the matrix metalloproteinase (MMP) family are involved in the breakdown of extracellular matrix (ECM) in normal physiological processes, such as embryonic development, reproduction and tissue remodelling, as well as in disease processes, such as arthritis and metastasis. The protein encoded by gene CP is a metalloprotein that binds most of the copper in plasma and is involved in the peroxidation of Fe(II)transferrin to Fe(III) transferrin. Mutations in this gene cause aceruloplasminemia, which results in iron accumulation and tissue damage. The protein encoded by gene PCM1 is a component of centriolar satellites, which are electron dense granules scattered around centrosomes. Chromosomal aberrations involving this gene are associated with papillary thyroid carcinomas and a variety of haematological malignancies, including atypical chronic myeloid leukaemia and T-cell lymphoma, suggesting that this gene plays an essential role in cancer development.

D2. Gene RLF is widely expressed in foetal and adult tissues, suggesting that it has a general role in transcriptional regulation. It encodes a Zn-15 related zinc finger protein and plays a role in deregulating the tightly controlled expression of the L-myc oncogene. Gene OXCT1 encodes a member of the 3-oxoacid CoA-transferase gene family. The encoded protein is a homodimeric mitochondrial matrix enzyme that plays a central role in extrahepatic ketone body catabolism by catalysing the reversible transfer of coenzyme A from succinyl-CoA to acetoacetate. Gene TMEM184B is one of the breast cancer markers identified by Bonnefoi *et al*. (2007). Gene TCEB1 encodes the protein elongin C, which is a subunit of the transcription factor B (SIII) complex. It belongs to the KEGG pathway in cancer, organism-specific biosystem. The modification of proteins with ubiquitin is an important cellular mechanism for targeting abnormal or short-lived proteins for degradation. Gene UBE4B is the strongest candidate in the neuroblastoma tumour suppressor genes. Gene PSEN2 is one of the confirmed Alzheimer's disease (AD) susceptibility genes. A negative correlation between the occurrence of AD and cancer has been observed. Gene CES2 encodes a member of the carboxylesterase large family. The protein encoded by this gene is the major intestinal enzyme and functions in intestine drug clearance. It has been identified as a cancer marker in Cai *et al*. (2009). The protein encoded by gene FGF2 is a member of the fibroblast growth factor (FGF) family. FGF family members bind heparin and

possess broad mitogenic and angiogenic activities. This protein has been implicated in diverse biological processes, such as limb and nervous system development, wound healing, and tumour growth. The GSN encoded calcium-regulated protein functions in both assembly and disassembly of actin filaments. Defects in this gene are a cause of familial amyloidosis Finnish type (FAF). The protein encoded by gene SERPINA1 is secreted and is a serine protease inhibitor whose targets include elastase, plasmin, thrombin, trypsin, chymotrypsin and plasminogen activator. Defects of this gene are associated with the development of breast cancer and lung cancer. Gene ZFP36 has been implied in the development of colon cancer, head and neck squamous cell carcinoma, tissue inflammation, cervical cancer and colorectal cancer, indicating its generic role in cancer development. The protein encoded by the LDB2 gene is capable of binding to a variety of transcription factors and is likely to function at enhancers to bring together diverse transcription factors and form higher order activation complexes or to block formation of such complexes (Jurata & Gill, 1997). The fact that LIM domain-binding factors are likely to be involved in the coordination of the transcriptional activity of many diverse factors may implicate them in human phenotypes characterized by multiple affected sites. Cell adhesion molecules (CAMs) are members of the mmunoglobulin superfamily. This gene encodes a neuronal CAM with multiple immunoglobulin-like C2-type domains and fibronectin type-III domains. This ankyrin-binding protein is involved in neuron–neuron adhesion and promotes directional signalling during axonal cone growth. This gene is also expressed in non-neural tissues and may play a general role in cell–cell communication via signalling from its intracellular domain to the actin cytoskeleton during directional cell migration. Vasodilator-stimulated phosphoprotein (VASP) is associated with filamentous actin formation and likely plays a widespread role in cell adhesion and motility. VASP may also be involved in the intracellular signalling pathways that regulate integrin-ECM interactions. VASP is regulated by the cyclic nucleotide-dependent kinases PKA and PKG. The protocadherin alpha gene cluster is one of three related clusters tandemly linked on chromosome five. These neural adhesion proteins most likely play a critical role in the establishment and function of specific cell–cell connections.

D3. The protein encoded by gene CRADD is a death domain (CARD/DD)-containing protein and has been shown to induce cell apoptosis. Through its CARD domain, this protein interacts with, and thus recruits, caspase 2/ICH1 to the cell death signal transduction complex that includes tumour necrosis factor receptor 1 (TNFR1A), RIPK1/RIP kinase and numbers of other CARD domain-containing proteins. Gene DGKD encodes a cytoplasmic enzyme that phosphorylates diacylglycerol to produce phosphatidic acid. Diacylglycerol and phosphatidic acid are two lipids that act as second messengers in signalling cascades. Their cellular concentrations are regulated by the encoded protein, and so it is thought to play an important role in cellular signal transduction. The TopBP1 protein includes eight BRCT domains (originally identified in BRCA1) and has homology with BRCA1 over the carboxyl terminal half of the protein. Gene CDC7 encodes a cell division cycle protein with kinase activity that is critical for the G_1/S transition. The yeast homologue is also essential for initiation of DNA replication as cell division occurs. Overexpression of this gene product may be associated with neoplastic transformation for some tumours. Gene BAZ1B encodes a member of the bromodomain protein family. It is expressed in multiple tumour tissues, including adrenal tumour, breast tumour, cervical tumour, chondrosarcoma, head and neck tumour, leukaemia, lymphoma, prostate cancer and several others. Gene CR1 is a member of the receptors of complement activation (RCA) family and is located in the 'cluster RCA' region of chromosome 1. The gene encodes a monomeric single-pass type I membrane glycoprotein. The protein mediates cellular binding to particles and immune complexes that have activated complement. Decreases in expression of this protein and/or mutations in its gene have been associated with gallbladder carcinomas, mesangiocapillary glomerulonephritis, systemic lupus erythematosus and sarcoidosis. Gene ANPEP has been identified as a marker for lung cancer and prostate cancer. In the small intestine aminopeptidase N plays a role in the final digestion of peptides generated from hydrolysis of proteins by gastric and pancreatic proteases. Gene POSTN is involved in the development of gastric cancer and pancreatic cancer. Gene MYL2 encodes the regulatory light chain associated with cardiac myosin beta (or slow) heavy chain. Ca^+ triggers the phosphorylation of regulatory light chain that in turn triggers contraction. Ankyrins are a family of proteins that link the integral membrane proteins to the underlying spectrin-actin cytoskeleton and play key roles in activities such as cell motility, activation, proliferation, contact and the maintenance of specialized membrane domains. Gene FCRL4 encodes a member of the immunoglobulin receptor superfamily and is one of several Fc receptor-like glycoproteins clustered on the long arm of chromosome 1. This protein may play a role in the function of memory B-cells in the epithelia. Aberrations in the chromosomal region encoding this gene are associated with non-Hodgkin lymphoma and multiple myeloma.

D4. Gene IFIT1B is identified as a lymphoma susceptibility marker in Gaiser *et al.* (2002). Gene SHC1

https://doi.org/10.1017/S0016672312000419 Published online by Cambridge University Press

encodes three main isoforms that differ in activities and subcellular location. Although all three are adapter proteins in signal transduction pathways, the longest (p66Shc) may be involved in regulating life span and the effects of reactive oxygen species. The other two isoforms, p52Shc and p46Shc, link activated receptor tyrosine kinases to the Ras pathway by recruitment of the GRB2/SOS complex. Gene SATB1 has been identified as a marker for breast cancer, gastric cancer and non-small cell lung cancer, suggesting its fundamental role in cancer development. Gene NRAS is an N-ras oncogene encoding a membrane protein that shuttles between the Golgi apparatus and the plasma membrane. Mutations in this gene have been associated with somatic rectal cancer, follicular thyroid cancer, autoimmune lymphoproliferative syndrome, Noonan syndrome and juvenile myelomonocytic leukaemia. The protein encoded by gene ID2 belongs to the inhibitor of DNA binding family, members of which are transcriptional regulators that contain a helix-loop-helix (HLH) domain but not a basic domain. This protein may play a role in negatively regulating cell differentiation. Gene CCL19 is one of several CC cytokine genes clustered on the p-arm of chromosome 9. Cytokines are a family of secreted proteins involved in immunoregulatory and inflammatory processes. The CC cytokines are proteins characterized by two adjacent cysteines. The cytokine encoded by this gene may play a role in normal lymphocyte recirculation and homing. It also plays an important role in trafficking of T-cells in thymus, and in T-cell and B-cell migration to secondary lymphoid organs. Gene PTGS2 encodes the inducible isozyme. It is regulated by specific stimulatory events, suggesting that it is responsible for the prostanoid biosynthesis involved in inflammation and mitogenesis. Gene TIMP3 belongs to the TIMP gene family. The proteins encoded by this gene family are inhibitors of the MMPs, a group of peptidases involved in degradation of the ECM. It has been implied in the development of multiple cancers, including for example colorectal cancer, head and neck cancer and breast cancer. Gene HHEX encodes a member of the homeobox family of transcription factors, many of which are involved in developmental processes. Gene ZNF592 plays a role in a complex developmental pathway and the regulation of genes involved in cerebellar development. Gene SH3KBP1 encodes an adapter protein that contains three N-terminal Src homology domains, a proline-rich region and a C-terminal coiled-coil domain. The encoded protein facilitates protein–protein interactions and has been implicated in numerous cellular processes including apoptosis, cytoskeletal rearrangement, cell adhesion and in the regulation of clathrin-dependent endocytosis.

References

Anjum, S., Doucet, A. & Holmes, C. C. (2009). A boosting approach to structure learning of graphs with and without prior knowledge. *Bioinformatics* **25**, 2929–2936.

Bonnefoi, H., Potti, A., Delorenzi, M., Mauriac, L., Campone, M., Tubiana-Hulin, M., Petit, T., Rouanet, P., Jassem, J., Blot, E., Becette, V., Farmer, P., Andre, S., Acharya, C. R., Mukherjee, S., Cameron, D., Bergh, J., Nevins, J. R. & Iggo, R. D. (2007). Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial. *Lancet Oncology* **8**, 1071–1078.

Buhlmann, P. & Yu, B. (2006) Sparse boosting. *Journal of Machine Learning Research* **7**, 1001–1024.

Burger, H., Kemming, D., Helms, M., Feldmann, U., Matuschek, A., Bocker, W. & Brandt, B. (2005). Expression of early placenta insulin-like growth factor (EPIL) in breast cancer cells provides an autocrine loop with enhancement of predominantly HER-2-related invasivity. *Verhandlungen der Deutschen Gesellschaft für Pathologie* **89**, 201–205.

Cai, L., Tang, X., Guo, L., An, Y., Wang, Y. & Zheng, J. (2009). Decreased serum levels of carboxylesterase-2 in patients with ovarian cancers. *Tumori* **95**, 473–478.

Casci, T. (2010). Gene networks: meaningful connections. *Nature Reviews Genetics* **11**, 172–173.

Dettling, M. & Buhlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19**, 1061–1069.

Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

Gaiser, T., Thorns, C., Merz, H., Noack, F., Feller, A. C. & Lange, K. (2002) . Gene profiling in anaplastic large-cell lymphoma-derived cell lines with cDNA expression arrays. *Journal of Hematotherapy and Stem Cell Research* **11**, 423–428.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*. Berlin: Springer.

Huang, E., Cheng, S. H., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iversen, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R. & Huang, A. T. (2003). Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596.

Huang, J., Ma, S., Li, H. & Zhang, C. (2011*a*). The sparse Laplacian shrinkage estimator for high dimensional regression. *Annals of Statistics* **39**, 2021–2046.

Huang, Y., Huang, J., Shia, B. C. & Ma, S. (2011*b*). Identification of cancer genomic markers via integrative sparse boosting. *Biostatistics*, in press.

Jornsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T. E., Nordlander, B., Sander, C., Gennemark, P., Funa, K., Nilsson, B., Lindahl, L. & Nelander, S. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular System Biology* **7**, 486.

Jurata, L. W. & Gill, G. N. (1997). Functional analysis of the nuclear LIM domain interactor NLI. *Molecular and Cellular Biology* **17**, 5688–5698.

Knudsen, S. (2006). *Cancer Diagnostics with DNA Microarray*. Hoboken, NJ: Wiley.

Langfelder, P. & Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC System Biology* **1**, 54.

Langfelder, P., Zhang, B. & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720.

Ma, S., Song, X. & Huang, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* **8**, 60.

Ma, D. F., Kondo, T., Nakazawa, T., Niu, D. F., Mochizuki, K., Kawasaki, T., Yamane, T. & Katoh, R. (2010*a*). Hypoxia-inducible adenosine A2B receptor modulates proliferation of colon carcinoma cells. *Human Pathology* **41**, 1550–1557.

Ma, S., Shi, M., Li, Y., Yi, D. & Shia, B. C. (2010*b*). Incorporating gene co-expression network in identification of cancer prognosis markers. *BMC Bioinformatics* **11**, 271.

Ma, S., Kosorok, M. R., Huang, J. & Dai, Y. (2011). Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis. *BMC Medical Genomics* **4**, 5.

Ma, S., Huang, J., Xie, Y. & Yi, N. (2012). Identification of breast cancer prognosis markers using integrative sparse boosting. *Methods of Information in Medicine* **51**, 152–161.

Maathuis, M. H., Colombo, D., Kalisch, M. & Buhlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods* **7**, 247–248.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, K. H., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., Lopez-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. & Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* **346**, 1937–1947.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W. C., Connors, J. M., Campo, E., Gascoyne, R. D., Grogan, T. M., Muller-Hermelink, H. K., Smeland, E. B., Chiorazzi, M., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Henrickson, S., Yang, L., Powell, J., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Montserrat, E., Bosch, F., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Fisher, R. I., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Holte, H., Delabie, J. & Staudt, L. M. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* **3**, 185–197.

Saris, C. G., Horvath, S., van Vught, P. W., van Es, M. A., Blauw, H. M., Fuller, T. F., Langfelder, P., DeYoung, J., Wokke, J. H., Veldink, J. H., van den Berg, L. H. & Ophoff, R. A. (2009). Weighted gene co-expression network analysis of the peripheral blood from amyotrophic lateral sclerosis patients. *BMC Genomics* **10**, 405.

Schapire, R. E. & Freund, Y. (2012). *Boosting: Foundations and Algorithms*. Cambridge, MA: MIT Press.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Rijn van de, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lonning, P. & Borresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences USA* **98**, 10869–10874.

Stute, W. (1993). Consistent estimation under random censorship when covariables are available. *Journal of Multivariate Analysis* **45**, 89–103.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* **11**, 1871–1879.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* **67**, 301–320.