Theoretical Biology and Medical Modelling

BioMed Central

Research

Bayesian profiling of molecular signatures to predict event times

Dabao Zhang and Min Zhang*

Address: Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, Indiana 47907-2067, USA

Email: Dabao Zhang - zhangdb@stat.purdue.edu; Min Zhang* - minzhang@purdue.edu

* Corresponding author

Published: 19 January 2007

Theoretical Biology and Medical Modelling 2007, **4**:3 doi:10.1186/1742-4682-4-3

Received: 24 September 2006

Accepted: 19 January 2007

This article is available from: http://www.tbiomed.com/content/4/1/3

Abstract

Background: It is of particular interest to identify cancer-specific molecular signatures for early diagnosis, monitoring effects of treatment and predicting patient survival time. Molecular information about patients is usually generated from high throughput technologies such as microarray and mass spectrometry. Statistically, we are challenged by the large number of candidates but only a small number of patients in the study, and the right-censored clinical data further complicate the analysis.

Results: We present a two-stage procedure to profile molecular signatures for survival outcomes. Firstly, we group closely-related molecular features into linkage clusters, each portraying either similar or opposite functions and playing similar roles in prognosis; secondly, a Bayesian approach is developed to rank the centroids of these linkage clusters and provide a list of the main molecular features closely related to the outcome of interest. A simulation study showed the superior performance of our approach. When it was applied to data on diffuse large B-cell lymphoma (DLBCL), we were able to identify some new candidate signatures for disease prognosis.

Conclusion: This multivariate approach provides researchers with a more reliable list of molecular features profiled in terms of their prognostic relationship to the event times, and generates dependable information for subsequent identification of prognostic molecular signatures through either biological procedures or further data analysis.

Background

High-throughput biotechnologies such as microarray and mass spectrometry permit simultaneous measurements of enormous bodies of genomic, proteomic, and metabolic information to be made. Such information helps us understand the molecular basis of important clinical outcomes, and thus improves the efficiency as well as accuracy in clinical decision making. More specifically, a small subset of these molecules can be used as biomarkers in daily clinical practice for detecting disease at early stages, measuring disease progress, monitoring the efficacy of treatments, and potentially accelerating the drug discovery process. However, the promise of genomics, proteomics, and metabolomics in clinical medicine rests on identifying these disease-specific molecular signatures. Clinical and preclinical studies of patients' genomics and proteomics profiles usually present datasets that share common characteristics, i.e., many molecular features ("large p") collected from few individuals ("small n"). The statistical challenge is to mine prognostic signatures from thousands of candidates by efficiently extracting information from samples of limited size, i.e., "small n large p" datasets. Moreover, the clinical outcomes measured for certain patients, e.g., survival times of cancer patients, are

usually censored data, which further complicates the statistical analysis. There has been extensive research on the classification and prediction of cancer using gene expression information [1-3], but there has been less progress in identifying individual molecules that can be used to predict the clinical outcome. We devote this paper to developing a Bayesian approach to profile molecular features on the basis of their prognostic relations to event times.

The proportional hazard model [4] has a long history in modeling the association of risk factors to the right-censored event times observed in clinical study [5,6]. Through this model, it has been of special interest to develop a systematic approach to identifying molecular signatures for event times with "small n large p" datasets. However, the overwhelmingly larger number of molecular candidates compared to the number of individuals prohibits exhaustive variable selection because of the heavy computation and model-overfitting considerations. A variety of strategies have been proposed in the literature. The first is to reduce the list of genotypic candidates by univariately associating each of them with phenotypic clinical outcome [1,7], and then regress the clinical outcome on the selected candidates. The second employs principal component analysis (PCA) to build up "eigengenes" (i.e., linear combinations of genes) and associates these with phenotypic clinical outcomes, and the identification of molecular signatures is further explored on the basis of these [8]. The third strategy employs partial least squares (PLS) [9,10] to construct orthogonal "eigengenes" [11]. Other strategies have also been used to reveal interesting prognostic molecular signatures for certain event times [12-15]. Recently, Tadesse et al. [16] proposed a Bayesian error-in-variable survival model to identify genes of which the expression levels are associated with survival outcome. It is widely accepted that most genes measured in microarray experiments provide little information for predicting patient survival, so a necessary step in the analysis is to reduce the number of candidates before identifying prognostic molecular signatures with a relatively small sample. This reduction is usually carried out by ranking molecular features (either the original molecular candidates or the "eigengenes") according to either z scores [7] or Cox scores [17-19], which measure the univariate association of each molecular feature with the event time. Several top-ranked molecular features are further explored for their prognostic associations with the event time. As shown in our simulation study, employing the univariate Cox scores to profile molecular features can be misleading as it may miss many important candidates but select many false-prognostic ones. Indeed, molecular features with high univariate association to the event time may not necessarily predict the event time effectively when applied together. As shown by Sha et al. [20] and Tadesse et al. [21], the disease may often be affected jointly by subsets of the genes while each individual gene might have a relatively weak effect. This study focuses on developing an efficient yet robust approach to profiling molecular features on the basis of their prognostic associations with the event time, taking advantage of the Bayesian framework for the proportional hazard model proposed by Kalbfleisch [22].

We acknowledge the high correlation between some molecular features due to the complicated genetic architecture. For example, genes involved in the same metabolic pathway may be similarly or oppositely regulated. These closely-related molecular features can result in collinearity between the candidates, and should therefore be grouped together in order to address their prognostic associations with the event time properly. Here, we group closely-related molecular features into linkage clusters. A centroid "gene" is constructed to represent each linkage cluster and thus partially solve the collinearity issue. As univariate Cox scores are unable to account for the complicated correlation structures among molecular features, we employ the Bayesian approach to construct a natural framework for molecular feature profiling.

We first propose a two-stage procedure for profiling prognostic molecular signatures for event times, and present the construction of linkage clusters as well as their centroids. A Bayesian framework of the Cox proportional hazard model is specified for "large p small n" data and a profiling criterion is described accordingly. The performance of our approach is evaluated via a simulation study and application to data concerning diffuse large B-cell lymphoma (DLBCL) [15].

Results

Simulation study

To evaluate the performance of the proposed approach, we simulated 20 survival datasets, each having p = 1, 000 features and n = 125 independent individuals. The feature values were generated from an autoregressive process of order one, with autocorrelation ρ = 0.5 and unit variance white noise. The event times follow an exponential distribution of which the rate is determined by a linear combination of the 12 features with non-zero coefficients. Independent random censoring times were generated from standard exponential distributions, and this induced censoring of approximately 50% of the observed event times. Among the 1, 000 autocorrelated features, the indices of the 12 with non-zero coefficients are 150, 151, 300, 302, 450, 453, 600, 604, 750, 755, 900, 906, and their values alternate between 1 and -1. Such constant-magnitude coefficients were chosen in order to evaluate the effect of correlation among features on the profiling, as the correlations between the pairs of features, i.e., (150, 151), (300, 302), (450, 453), (600, 604), (750, 755), and (900,

906), decrease geometrically from 0.5 to 0.015625. As shown in Figure 1, these feature pairs have similar chances of being selected as top features while being ranked by the Bayesian approach. In this simulation study, the proposed Bayesian approach could select each non-zero coefficient feature with high probability (more than 0.8) when more than 12 features were selected in total. However, when univariate Cox scores are used, a feature pair with higher correlation is more likely to be among the selected top features, and in general, all 12 features are less likely to be correctly selected, as shown in Figure 2. The percentages of the 12 non-zero coefficient features selected into top features (i.e., success rates) are shown in Figure 3 when using the Bayesian approach, or the univariate Cox scores. The univariate Cox scores can lead to very high false discovery rates because the features with non-zero coefficients are usually ranked very low. Furthermore, as shown in Figure 3, the success rates of selecting features with non-zero coefficients are very low even when a large number of features are selected. On the other hand, when more than 12 features are selected using the Bayesian approach, the success rates are usually higher than 0.8 and approach 1 very quickly as more features are selected.

Application to a real dataset

We applied the proposed two-stage procedure to data on diffuse large B-cell lymphoma (DLBCL) [15]. These data include the expression levels of 7, 399 genes from a total of 240 patients. The genomic information for each patient was obtained at the beginning of the study, and the patients were followed up until death or the end of the project. The missing gene expression values were imputed using the nearest neighbor averaging approach [12,23]. Using the single linkage clustering approach in Cluster 3.0 [24], we identified 5,656 linkage clusters by pruning the hierarchical tree such that the node distances within branches are less than 0.2. There are 4,944 linkage clusters containing only one gene, while the largest has 186 genes. We then consider selecting prognostic molecular features

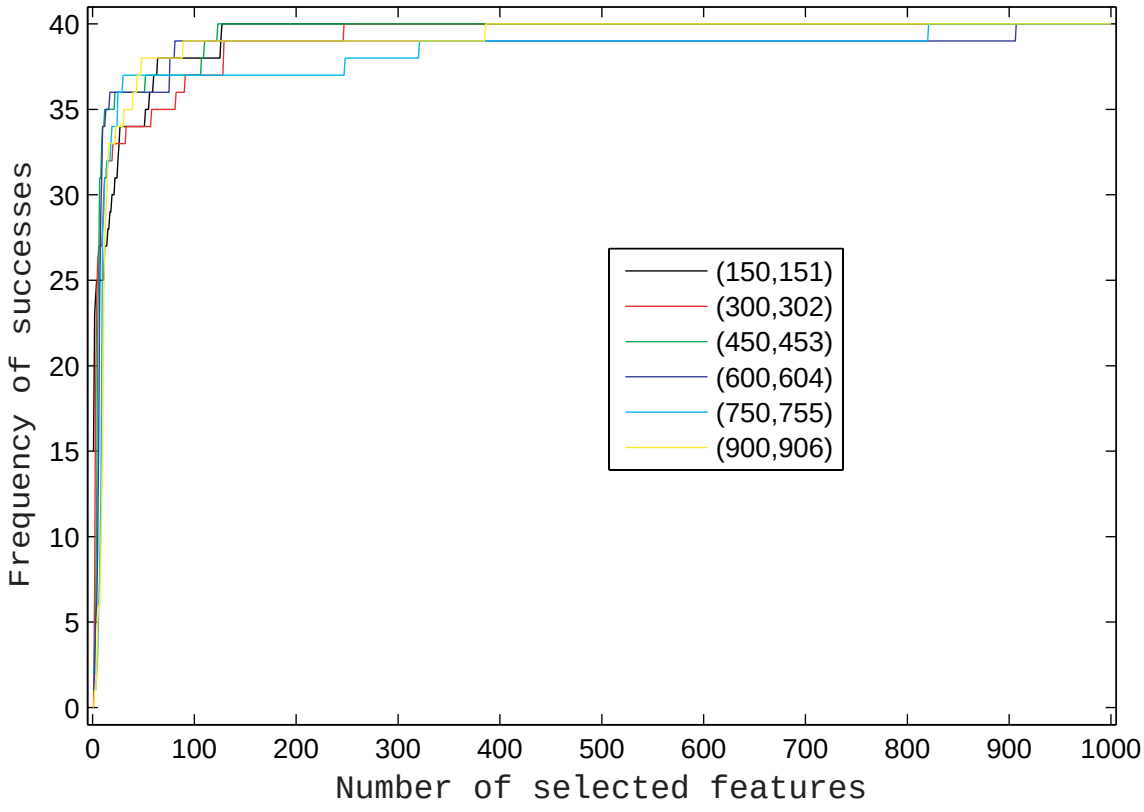

Figure 1

Frequency of successes using the Bayesian approach. For each of the six feature pairs, the frequency of successes (*y*-axis) is calculated as the total number of correct detections in the 20 simulated datasets when the Bayesian approach is used to select a certain number of features (*x*-axis).

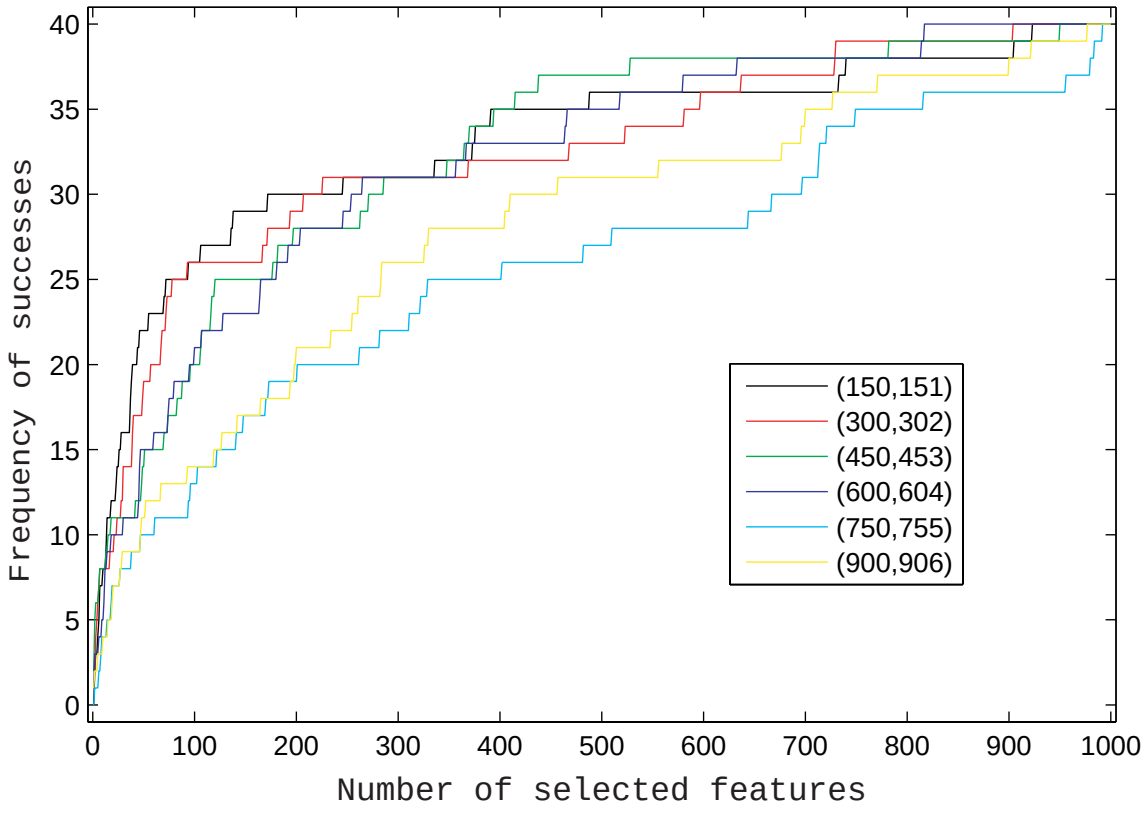

Figure 2
Frequency of successes using univariate Cox scores. For each of the six feature pairs, the frequency of successes (*y*-axis) is calculated as the total number of correct detections in the 20 simulated datasets when univariate Cox scores are used to select a certain number of features (*x*-axis).

from the 5, 656 candidates, each being the centroid of a linkage cluster.

The univariate Cox scores of all candidate clusters are calculated and shown in decreasing order in Figure 4. There are 761 candidates with Cox scores above the 95 percentile of the χ_1^2 distribution, and 290 candidates with Cox scores above the 99 percentile of the χ_1^2 distribution. We selected the top 100, 200, 300, and 500 candidates with the largest Cox scores and applied our Bayesian method to profile them. The top 25 of the 500 candidates are listed in Table 1.

Employing our Bayesian approach to profile the 500 candidates with the largest Cox scores, the posterior probabilities, i.e., $\tilde{p}_k$ defined in (2), of the top 25 clusters range from 0.0538 to 0.9825. However, the ranks of these 25 clusters vary widely when their univariate Cox scores are used, and only five of those with the top 25 univariate Cox scores appear in this list. Therefore, it may be misleading to profile the clusters for their prognostic ability on the basis of their univariate Cox scores, since many false prognostic features can be highly ranked owing to the complicated correlation structure among features.

When fewer than 500 candidates, for example, 100, 200 or 300, are profiled with the Bayesian approach, most of those that appeared in the top 25 of the 500 profiled candidates are also among the top 25 clusters as long as they are profiled. Indeed, the only exception is the cluster with two features in gene NM_00176, which was ranked at 61 when 300 candidates were profiled by the Bayesian approach. However, the complicated correlation structure between clusters makes it preferable to profile a number of clusters sufficient to avoid missing critical prognostic features. An exploratory selection of prognostic features from the top 25 clusters shown in Table 1 implies that 16

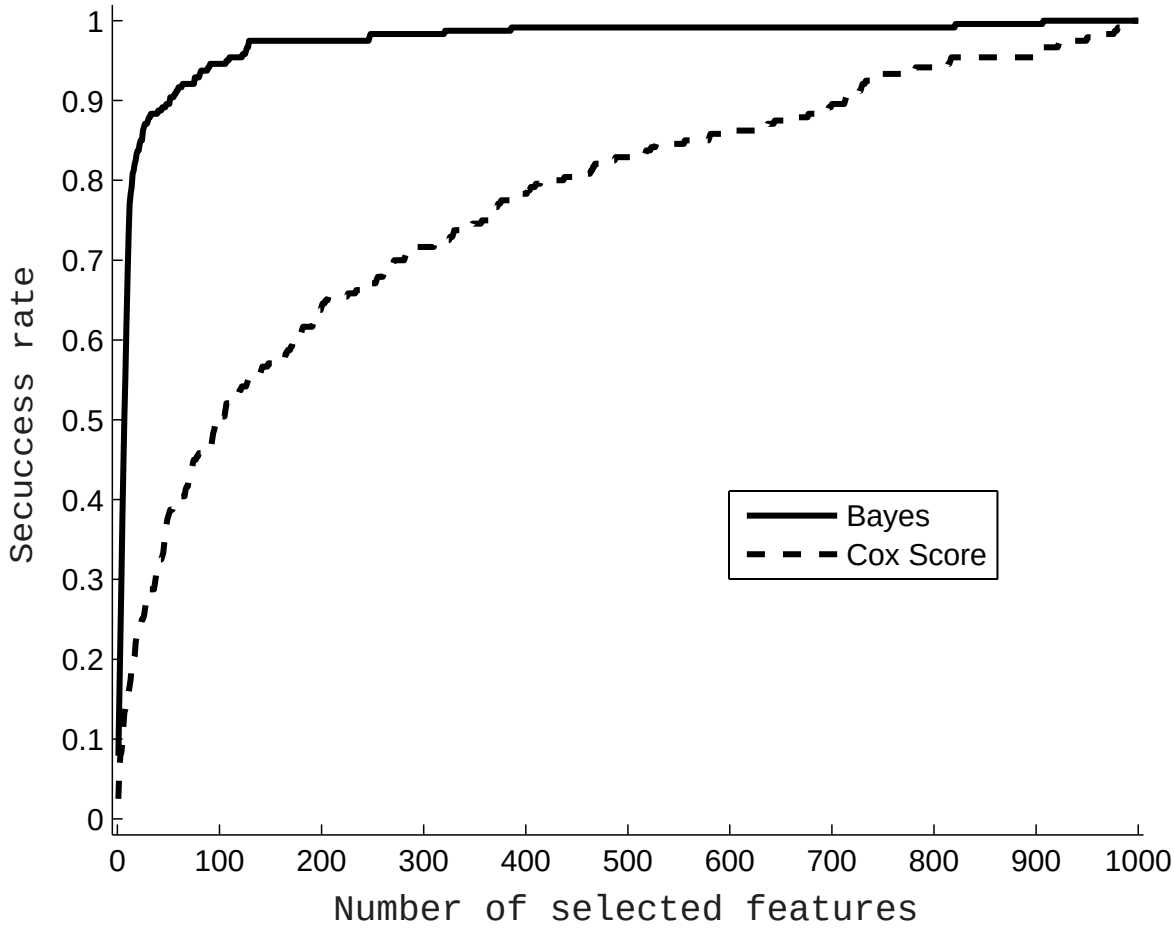

Figure 3
Comparison between the Bayesian approach and Cox scores. Shown as success rates (*y*-axis) are the true positive rates when a certain number of features (*x*-axis) are selected in each of the 20 simulated datasets. The solid line represents the results from the Bayesian approach and the dotted line represents the results using univariate Cox scores.

genes may be considered to construct prognostic features for the event time, and some of these features were ignored from the lists of 100, 200, and 300 candidates chosen on the basis of their univariate Cox scores. The cluster with 38 features from 11 genes is not one of the 16 selected, though all those genes except AK000170 belong to the MHC class II signature group defined by Rosenwald et al. [15]. D13666, which was reported by both Sha et al. [20] and Gui and Li [25], belongs to the lymph-node signature group, and BC012161 and AF134159 belong to the proliferation signature group (see Rosenwald et al. [15]). D42043, D88532, BC012161, and LC_33732 were also reported by Sha et al. [20]. It is interesting to observe that, among the 16 selected genes, AF414120 (gene CTLA4) is a member of the immunoglobulin superfamily and encodes a protein that transmits an inhibitory signal to T cells (Ling et al. [26]). AF127481, a lymphoid blast crisis oncogene (LBC), plays an important role in regulating the Rho/Rac GTPase cycle while the Rho/Rac family of small GTPases mediates cytoskeletal reorganization, gene transcription, and cell cycle progression through unique signal transduction pathways (Sterpetti et al. [27]). U46767 (gene CCL13) encodes a cytokine that plays a role in the accumulation of leukocytes during inflammation (Garcia-Zepeda et al. [28]). NM_000176 (gene NR3C1) encodes a receptor for glucocorticoids that can act as both a transcription factor and a regulator of other transcription factors. This protein can also be found in heteromeric cytoplasmic complexes along with heat shock factors and immunophilins (Subramaniam et al. [29]). X52186 (gene ITGB4) encodes the integrin beta 4 subunit, a receptor for the laminins, which tends to asso-

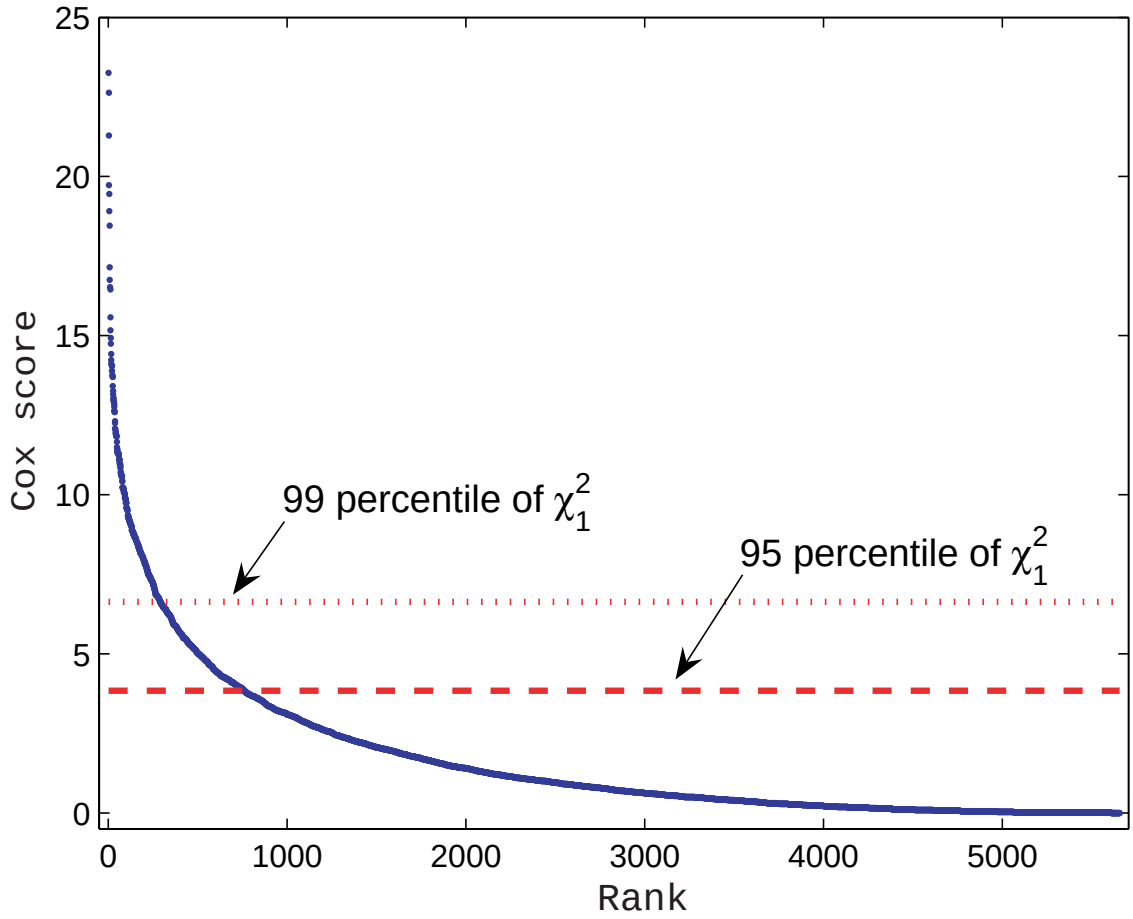

Figure 4

Cox Score plot for DLBCL data. This figure shows the descending Cox scores of 5,656 Candidates in the DLBCL data. The dotted and dashed lines indicate the 99 and 95 percentiles of the χ_1^2 distribution respectively.

ciate with the alpha 6 subunit and is likely to play a pivotal role in the biology of invasive carcinoma (Hogervorst et al. [30]).

Discussion

With high-throughput techniques now available, there has been extensive recent discussion of disease-specific molecular signatures [31,32]. The whole genome and proteome profiles for each of the limited number of patients presents an enormous number of molecular candidates with a complicated correlation structure. Here we group highly-correlated molecular features into linkage clusters in order to profile prognostic signatures. While the molecular features within the same linkage clusters are expected to have similar prognostic association with the event time, physical linkages and metabolic pathways will be able to provide confirmatory information. Thereupon, we strongly suggest that available genome, proteome, and metaboleme information be explored and combined with observed profiles to construct the linkage clusters. By doing this, we can improve the reliability significantly and establish the biological functionality of linkage clusters without overusing the limited number of profiles.

When an optimal subset with a prespecified number of candidates is targeted, classical model selection approaches may be employed to explore all possible subsets and identify the best one. However, "small n large p" datasets may still obstruct this practice because of the enormous computation and unidentifiable models involved. Indeed, when p diverges as $n \to \infty$, many classi-

Table 1: Bayesian Profiling of the DLBCL Data. The 16 starred genes are proposed by our exploratory selection. Bracketed are the numbers of features from the same gene, which are included in the same cluster.

GenBank Accession No.	Cox		100		200		300		500	
	Score	Rank	$\tilde{p}_k$	Rank	$\tilde{p}_k$	Rank	$\tilde{p}_k$	Rank	$\tilde{p}_k$	Rank
*D42043	18.91	6	0.99	2	0.99	1	0.99	1	0.98	1
*D88532	9.13	122	-	-	0.99	2	0.89	2	0.93	2
*U50196	6.82	272	-	-	-	-	0.12	6	0.32	3
*BC012161	21.28	3	0.81	4	0.55	4	0.61	4	0.31	4
*AF414120	5.77	387	-	-	-	-	-	-	0.27	5
*AF004709	5.27	460	-	-	-	-	-	-	0.20	6
AF127481	17.14	8	1.00	1	0.90	3	0.67	3	0.19	7
*AK025954	5.93	365	-	-	-	-	-	-	0.14	8
*AA504484 (2)	5.23	468	-	-	-	-	-	-	0.11	9
J00220	8.40	167	-	-	0.32	9	0.09	11	0.09	10
*AA837319	9.84	96	0.45	6	0.20	11	0.07	15	0.09	11
AA027985	11.44	50	0.30	9	0.18	13	0.09	10	0.09	12
*D13666	12.85	32	0.34	8	0.34	7	0.10	7	0.08	13
LC_33732	7.52	223	-	-	-	-	0.07	16	0.08	14
*AK000271	7.01	260	-	-	-	-	0.09	12	0.08	15
AF134159	14.11	18	0.43	7	0.20	10	0.08	13	0.07	16
*AA805749	5.72	393	-	-	-	-	-	-	0.07	17
*U46767	11.33	52	0.51	5	0.20	12	0.13	5	0.06	18
*AA804793	9.21	117	-	-	0.43	6	0.09	9	0.06	19
AA829241	12.76	33	0.25	10	0.12	21	0.06	19	0.06	20
*NM_000176 (2)	11.28	56	0.19	12	0.13	18	0.02	61	0.06	21
X00452 (5)	14.42	16	0.92	3	0.46	5	0.10	8	0.06	22
X00457 (3)										
X62744 (3)										
U15085 (4)										
M16276 (4)										
K01171 (5)										
M20430 (4)										
M83664 (2)										
K01144 (6)										
AK000170										
LC_24239										
*X52186	5.44	436	-	-	-	-	-	-	0.05	23
U59302	6.68	287	-	-	-	-	0.07	14	0.05	24
AA825732	7.67	215	-	-	-	-	0.05	21	0.05	25

cal approaches may not work even if $p < n$. Although univariate Cox scores are frequently utilized to profile candidates and accordingly select the subset, it is risky to identify prognostic signatures by this approach as it is easy to include false signatures but miss the true ones owing to the strong correlations among molecular features. For example, when a molecular feature is positively correlated with both true signatures, it may happen that the false one, instead of the two true ones, is selected. Ein-Dor et al. [33] discussed the discrepancies while using a univariate approach. Built upon the multivariate proportional hazard model (1), the proposed Bayesian approach is able to search all possible subsets of a certain size stochastically via Gibbs sampling. With restrictive priors for "small n large p" datasets, the posterior probability $\tilde{p}_k$ serves as a relative measure for profiling each candidate's prognostic association with the event time, accounting for other candidates. It is straightforward to extend this Bayesian approach to profile molecular signatures by controlling other clinical factors [3] and considering microenvironments [34].

The three-component prior for the coefficients in model (1) is crucial in constructing the profiling criterion. First, the prior probability of each component can be controlled with a uniform distribution on a subset of [0,1] to guarantee that the model is identifiable such that a Gibbs sampler can feasibly be employed to search the parameter

space stochastically. Putting these prior probabilities on a restricted interval allows various numbers of nonzero coefficients in model (1). Second, the three-component prior approach provides flexibility in the possible imbalance between the scales and/or sizes of positive and negative coefficients in the model. Third, the three-component prior automatically results in a three-component posterior distribution for each coefficient, with the posterior probability of each component available for further calculation. In summary, the profiling criterion, posterior probability ($\tilde{p}_k$), has a natural explanation and can be easily implemented in practice.

Methods

Construction of linkage clusters and their centroids

To facilitate pattern recognition and reveal otherwise hidden structures and functions, genes and proteins are usually clustered into groups based on different biological metrics, such as sequence similarity [35] or expression profiles [12,36]. With gene expression data only, many approaches have been applied to cluster genes that exhibit similar expression profiles across samples (see the review by Jörnsten and Yu, [36]). As we are interested in the prognostic relationships of genes to the event time, highly correlated genes are more likely to have the same power to predict the event time and therefore should be grouped into the same cluster. Here we define linkage clusters as groups of genes with large pairwise correlations in absolute value. As shown by Shaffer et al. [37], these linkage clusters of molecular features may reveal functional similarity/dissimilarity. Proteins regulated in the same metabolic pathway may also act similarly or oppositely. Although expressions of genes/proteins are usually observed with measurement errors, these linkage clusters can still be identified in experimental data. As shown in Figure 5, the molecular features can be mutually correlated as highly as correlation coefficient $|\rho| \geq 0.96$. From a statistical point of view, these closely-related molecular features can cause collinearity or near collinearity in multivariate identification of prognostic signatures and therefore destabilize the identification result if all these closely-related molecular features are included in the prognostic model.

Hierarchical clustering approaches (e.g., the complete linkage clustering or single linkage clustering approach in Cluster 3.0 by de Hoon et al. [24]) can be used to construct these linkage clusters. With the mutual correlation coefficients estimated from the data, we use the absolute values of correlation coefficients as the similarity scores, i.e., with the distance measure $d = 1 - |\rho|$. We prune the hierarchical tree with a prespecified value for the distance, e.g., $d \leq 0.2$ (hence absolute values of the correlation coef-

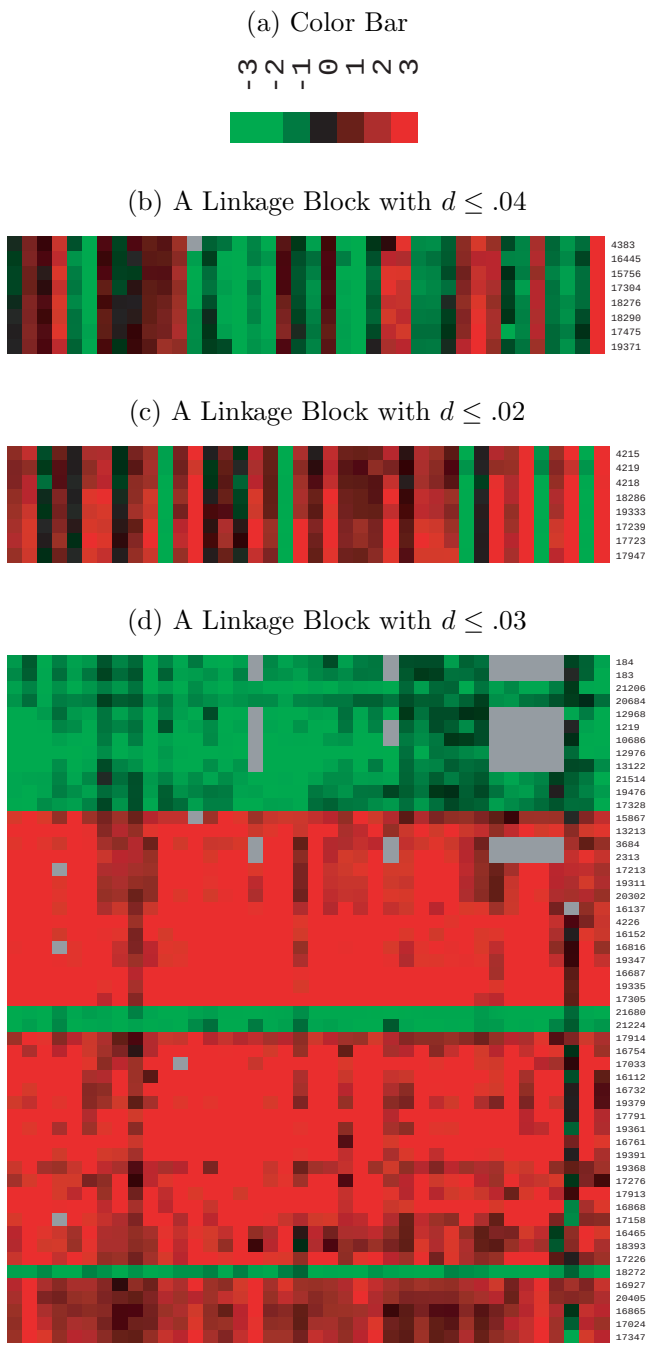

Figure 5
Illustration of Linkage Blocks. Part (a) indicates the color bar used in the other parts. Parts (b), (c) and (d) represent three different linkage blocks with correlation coefficients being 0.96, 0.98, and 0.97 respectively, where each row corresponds to a gene and each column corresponds to an individual.

ficients are no less than 0.8). The molecular features within the same branch are assumed to be within the same linkage cluster. The centroid of the linkage cluster is used to represent all the elements within the linkage cluster, and subsequent identification of prognostic signatures proceeds by associating these centroids only with the event time.

The expression levels of the centroids are calculated by standardizing expressions of all genes. More specifically, for each linkage cluster, we first randomly select one gene and reverse the expression signs of all genes within the cluster that are negatively correlated with this gene. Then the expression level of the centroid is calculated by averaging the expression levels of all the genes within the same linkage cluster. Meanwhile, the measurement errors in expression levels are attenuated after averaging the gene expression levels within the same linkage cluster.

Bayesian framework of proportional hazard model

Suppose that p linkage clusters are identified and therefore the expressions of the p centroids are calculated for each of the n individuals. The observed event time for the i-th individual is denoted γ_i, with δ_i indicating whether it was right-censored owing to loss in follow-up (i.e., $\delta_i = 0$ if right-censored and $\delta_i = 1$ otherwise). Accordingly, the expressions of the centroids are denoted $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})$. We use the popular Proportional Hazard Model [4] to associate the molecular features with the event time, i.e., the hazard function is modeled as follows:

$$\lambda(t|\boldsymbol{z}_i) = \lambda_0(t)\exp(\boldsymbol{z}_i\beta), \quad (1)$$

where β includes the p coefficients for all the centroids, and $\lambda_0(\cdot)$ is an unspecified baseline hazard function.

Further, let $\mathcal{D} = \{(\gamma_i, \delta_i, z_i) : i = 1, 2, \ldots, n\}$ be the observed data, and $\mathcal{R}(t) = \{i : \gamma_i \geq t\}$ be the risk set at time t. Following Kalbfleisch [22], we construct the Bayesian framework to estimate β by considering only the partial likelihood function,

$$PL(\boldsymbol{\beta}\,|\mathcal{D}) = \prod_{i=1}^{n}\left\{\frac{\exp\{\boldsymbol{z}_i\boldsymbol{\beta}\}}{\sum_{j\in\mathcal{R}(\gamma_i)}\exp\{\boldsymbol{z}_j\boldsymbol{\beta}\}}\right\}^{\delta_i},$$

which avoids the nuisance baseline hazard function $\lambda_0(t)$.

With a large number of available linkage clusters, the time-to-event of interest may be associated with a relatively small number of linkage clusters. On the other hand, the available "large p small n" data sets hamper us in detecting linkage clusters with too weak effects on the time-to-event of interest, and we expect to be able to identify those linkage clusters with strong effects. We therefore incorporate this important prior information by considering the following prior distribution for each β_k:

$$\beta_k \sim (1 - w_+ - w_-)\delta_{\{0\}} + w_+N_+(0,\ \sigma_+^2) + w_-N_-(0,\ \sigma_-^2),$$

where $N_+(\mu, \sigma^2)$ and $N_-(\mu, \sigma^2)$ are the truncated Gaussian distributions with only positive and negative parts, respectively. As shown by Zhang et al. [38] and Zhang et al. [39], this three-component prior has some theoretical properties and allows a possible imbalance between scales and/or sizes of positive and negative coefficients in model (1).

A priori, the hyperparameters σ_+^2 and σ_-^2 are assumed to follow inverse gamma distributions as $IG(1, \phi_+)$ and $IG(1, \phi_-)$, respectively. Here, sufficiently large ϕ_+ and ϕ_- are recommended to approximate the noninformative priors $1/\sigma_+^2$ and $1/\sigma_-^2$, respectively. For the prior distribution of (w_+, w_-), a noninformative prior (such as $Dirichlet(1, 1, 1)$) is not applicable since the model is not identifiable with $p \gg n$. As shown in Zhang et al. [40], the number of reliably identified significant predictors is limited by the sample size n. Following Zhang et al. [39] and Zhang et al. [41], we let

$$w_+ + w_- \sim Unif(0,\ 2\sqrt{n}/p),$$

which guarantees the model to be identifiable. On the other hand, as the number of candidate predictors is large, the upper bound, $2\sqrt{n}/p$, on $(w_+ + w_-)$ can be so restrictive that the resultant posterior probability for a true predictor to be significant can be very small. However, these posterior probabilities, as relative measures of significance, play an important role in profiling all features for their prognostic relations to the event time.

The Gibbs sampler

In view of the large number of parameters to be estimated, we consider a Gibbs sampler to obtain the posterior distributions of the parameters and make inferences. The Gibbs sampler can be developed by iteratively sampling each parameter from its full conditional distribution.

For simplicity, let β_{-k} include all components of β except β_k, and write $g_k(\beta_k|\beta_{-k}, \mathcal{D}) = PL(\beta|\mathcal{D})$ when β_k is of particular interest. Then the full conditional distribution of β_k is a mixture of a point mass at zero and two continuous distributions, i.e.,

$$\beta_k | w_+, w_-, \sigma_+^2, \sigma_-^2, \mathcal{D}$$

$$\sim (1 - \tilde{w}_{k+} - \tilde{w}_{k-})\delta_{\{0\}} + \tilde{w}_{k+}\tilde{F}_{k+} + \tilde{w}_{k-}\tilde{F}_{k-},$$

where $\tilde{F}_{k+}$ and $\tilde{F}_{k-}$ zare the distributions corresponding to the following probability density functions,

$$\tilde{f}_{k+}(x) = \frac{g_k(x \mid \boldsymbol{\beta}_{-k}, \mathcal{D})}{w_{k+}\sqrt{2\pi\sigma_+^2}} \exp\left\{-\frac{x^2}{2\sigma_+^2}\right\} I[x > 0],$$

$$\tilde{f}_{k-}(x) = \frac{g_k(x \mid \boldsymbol{\beta}_{-k}, \mathcal{D})}{w_{k-}\sqrt{2\pi\sigma_-^2}} \exp\left\{-\frac{x^2}{2\sigma_-^2}\right\} I[x < 0],$$

and the probabilities for β_k to be positive and negative are, respectively,

$$\tilde{w}_{k+} = \frac{2w_+ w_{k+}}{(1 - w_+ - w_-)g_k(0 \mid \boldsymbol{\beta}_{-k}, \mathcal{D}) + 2w_+ w_{k+} + 2w_- w_{k-}},$$

$$\tilde{w}_{k-} = \frac{2w_- w_{k-}}{(1 - w_+ - w_-)g_k(0 \mid \boldsymbol{\beta}_{-k}, \mathcal{D}) + 2w_+ w_{k+} + 2w_- w_{k-}}.$$

Here, w_{k+} and w_{k-} are normalization coefficients, which can be calculated as

$$w_{k+} = \int_0^\infty \frac{g_k(x \mid \boldsymbol{\beta}_{-k}, \mathcal{D})}{\sqrt{2\pi\sigma_+^2}} \exp\left\{-\frac{x^2}{2\sigma_+^2}\right\} dx,$$

$$w_{k-} = \int_{-\infty}^0 \frac{g_k(x \mid \boldsymbol{\beta}_{-k}, \mathcal{D})}{\sqrt{2\pi\sigma_-^2}} \exp\left\{-\frac{x^2}{2\sigma_-^2}\right\} dx.$$

The full conditional distribution of w_+ and w_- is

$$(w_+, w_-, 1 - w_+ - w_-) | \beta$$

$$\sim Dirichlet(\tilde{p}_+, \tilde{p}_-, p - \tilde{p}_+ - \tilde{p}_-), \quad w_+ + w_- \le \frac{2\sqrt{n}}{p},$$

where $\tilde{p}_+ = \#\{k : \beta_k > 0\}$ and $\tilde{p}_- = \#\{k : \beta_k < 0\}$. Finally, the full conditional distribution of σ_+^2 and σ_-^2 are

$$\sigma_+^{-2} \mid \boldsymbol{\beta} \sim Gamma\left(\frac{\tilde{p}_+}{2}, \left(\frac{1}{\phi_+} + \frac{1}{2}\sum_{k=1}^{p} \beta_k^2 I[\beta_k > 0]\right)^{-1}\right),$$

$$\sigma_-^{-2} \mid \boldsymbol{\beta} \sim Gamma\left(\frac{\tilde{p}_-}{2}, \left(\frac{1}{\phi_-} + \frac{1}{2}\sum_{k=1}^{p} \beta_k^2 I[\beta_k < 0]\right)^{-1}\right).$$

The parameters were initialized on the basis of estimators from univariate approaches. After the initial burn-in period (5,000 in the following analysis), the next 5,000 iterations in the Markov chain were used for inference without thinning. Convergence of the algorithm was checked by the diagnostic tools in Cowles and Carlin [42].

Profiling criterion

The significance of each centroid in model (1) is determined by one pair of parameters. They are, for the *j*-th centroid, the posterior probabilities $p_{k+} = P(\beta_k > 0 \mid \mathcal{D})$ and $p_{k-} = P(\beta_k < 0 \mid \mathcal{D})$. Given data $\mathcal{D}$, the marginal posterior distribution of β_k is still a mixture of three components, i.e., being positive with probability $p_{k+} = E[\tilde{w}_{k+} \mid \mathcal{D}]$, being negative with probability $p_{k-} = E[\tilde{w}_{k-} \mid \mathcal{D}]$, and having a point mass at zero with probability $1 - p_{k+} - p_{k-}$. The two parameters p_{k+} and p_{k-} can be estimated from the Markov chains of $\tilde{w}_{\beta k+}$ and $\tilde{w}_{\beta k-}$ drawn from the above Gibbs sampler. With moderately large *p*, the upper bound $2\sqrt{n}/p$ on $(w_+ + w_-)$ may not be restrictive and β_k can be estimated with the median value of its posterior probability. However, *p* is usually much larger than *n* in gene expression data and, as a result, p_{k+} and p_{k-} may be heavily shrunk to zero. Therefore, identifying significant prognostic centroids with the posterior median values will be too conservative. Instead, we suggest profiling the prognostic association of *k*-th centroid to the event time by

$$\tilde{p}_k = \max\{p_{k+}, p_{k-}\}, \quad (2)$$

which is a relative measure when *p* is much larger than *n*. As demonstrated in the simulation study, this profiling criterion performs much better than the popular univariate Cox score.

Authors' contributions

DZ and MZ both contributed to the development of the modeling method. DZ wrote the Matlab code and did the simulation study. MZ analyzed the real data. Both authors read and approved the final manuscript.

Acknowledgements

Thanks to Derick Peterson for providing the 20 simulated datasets. This research was partially supported by a startup fund from the Department of Statistics of Purdue University, and a summer faculty grant from the Purdue

Research Foundation. The authors thank the editor, and two referees for insightful comments that greatly improved the manuscript.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression profiling.** *Science* 1999, **286:**531-537.
2. Li H, Gui J: **Partial Cox regression analysis for high-dimensional microarray gene expression data.** *ISMB04/Bioinformatics* 2004, **20:**i208-i215.
3. Li L: **Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information.** *Bioinformatics* 2006, **22:**466-471.
4. Cox DR: **Regression models and life tables.** *J R Stat Soc Ser B* 1972, **39:**264-296.
5. Al-katib A: **Treatment of diffuse poorly differentiated lymphocytic lymphoma: an analysis of prognostic variables.** *Cancer* 1984, **53:**2404-2412.
6. Papatestas AE, Miller SR, Pertsemlidis D, Fagerstrom R, Lesnick G, Aufses AH: **Association between prognosis and hormone receptors in women breast cancer.** *Cancer Detection and Prevention* 1986, **9:**303-310.
7. Lossos IS, Czerwinski DK, Alizadeh AA, Wechser MA, Tibshirani R, Botstein D, Levy R: **Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes.** *N Eng J Med* 2004, **350:**1828-1837.
8. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JAJ, Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proc Natl Acad Sci USA* 2001, **98:**11462-11467.
9. Wold H: **Estimation of principal components and related models by iterative least squares.** In *Multivariate Analysis* Edited by: Krishnaiaah PR. London: Academic Press; 1966:391-420.
10. Garthwaite PH: **An interpretation of partial least squares.** *J Am Stat Assoc* 1994, **89:**122-127.
11. Nguyen DV, Rocke DM: **Tumor classification by partial least squares using microarray gene expression data.** *Bioinformatics* 2002, **18:**39-50.
12. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P: **'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns.** *Genome Biology* 2000.
13. Li H, Luan Y: **Kernel Cox regression models for linking gene expression profiles to censored survival data.** *Pacific Symposium of Biocomputing* 2003, **8:**65-76.
14. Park PJ, Tian L, Kohane IS: **Linking expression data with patient survival times using partial least squares.** *Bioinformatics* 2002, **18:**1625-1632.
15. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, Lopez-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM, Project LMP: **The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.** *New Engl J Med* 2002, **346:**1937-1947.
16. Tadesse MG, Ibrahim JG, Gentleman R, Chiaretti S, Ritz J, Foa R: **Bayesian error-in-variable model for the analysis of genechip arrays.** *Biometrics* 2005, **61:**488-497.
17. Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JMG, Iannetton MD, Orringer MB, Hanash S: **Gene-espression profiles predict survival of patients with lung adenocarcinoma.** *Nature Medicine* 2002, **8:**816-824.
18. Bair E, Tibshirani R: **Semi-supervised methods to predict patient survival from gene expression data.** *PLoS Biology* 2004, **2:**511-522.
19. Li H, Luan Y: **Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data.** *Bioinformatics* 2005, **21:**2403-2409.
20. Sha N, Tadesse MG, Vannucci M: **Bayesian variable selection for the analysis of microarray data with censored outcomes.** *Bioinformatics* 2006, **22:**2262-2268.
21. Tadesse MG, Sha N, Vannucci M: **Bayesian variable selection in clustering high-dimensional data.** *J Am Stat Assoc* 2005, **100:**602-617.
22. Kalbfleisch JD: **Nonparametric Bayesian analysis of survival time data.** *J R Stat Soc Ser B* 1978, **40:**214-221.
23. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.
24. De Hoon MJL, Imoto S, Nolan J, Miyano S: **Open source clustering software.** *Bioinformatics* 2004, **20:**1453-1454.
25. Gui J, Li H: **Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data.** *Bioinformatics* 2005, **21:**3001-3008.
26. Ling V, Wu PW, Finnerty HF, Sharpe AH, Gray GS, Collins M: **Complete sequence determination of the mouse and human CTLA4 gene loci: cross-species DNA sequence similarity beyond exon borders.** *Genomics* 1999, **60:**341-355.
27. Sterpetti P, Hack AA, Bashar MP, Park B, Cheng SD, Knoll JHM, Urano T, Feig LA, Toksoz D: **Activation of the Lbc Rho exchange factor proto-oncogene by truncation of an extended C terminus that regulates transformation and targeting.** *Molec Cell Biol* 1999, **19:**1334-1345.
28. Garcia-Zepeda EA, Combadiere C, Rothenberg ME, Sarafi MN, Lavigne F, Hamid Q, Murphy PM, Luster AD: **Human monocyte chemoattractant protein (MCP)-4 is a novel CC chemokine with activities on monocytes, eosinophils, and basophils induced in allergic and nonallergic inflammation that signals through the CC chemokine receptors (CCR)-2 and -3.** *J Immunol* 1996, **157:**5613-5626.
29. Subramaniam M, Colvard D, Keeting PE, Rasmussen K, Riggs BL, Spelsberg TC: **Glucocorticoid regulation of alkaline phosphatase, osteocalcin, and proto-oncogenes in normal human osteoblast-like cells.** *J Cell Biochem* 1992, **50:**411-424.
30. Hogervorst F, Kuikman I, von dem Borne AE, Sonnenberg A: **Cloning and sequence analysis of beta-4 cDNA: an integrin subunit that contains a unique 118 kd cytoplasmic domain.** *EMBO J* 1990, **9:**765-770.
31. Campbell CJ, Ghazal P: **Molecular signatures for diagnosis of infection: application of microarray technology.** *Journal of Applied Microbiology* 2003, **96:**18-23.
32. Mocellin S, Wang E, Panelli M, Pilati P, Marincola FM: **DNA array-based gene profiling in tumor immunology.** *Clinical Cancer Research* 2004, **10:**4597-4606.
33. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **2:**171-178.
34. Strausberg RL: **Tumor microenvironments, the immune system and cancer survival.** *Genome Biology* 2005, **6:**211.1-211.4.
35. Henikoff S, Henikoff JG: **Automated assembly of protein blocks for database searching.** *Nuclei Acids Research* 1991, **19:**6565-6572.
36. Jörnsten R, Yu B: **Simultaneous gene clustering and subset selection for sample classification via MDL.** *Bioinformatics* 2002, **19:**1100-1109.
37. Shaffer AL, Rosenwald A, Hurt EM, Giltnane JM, Lam LT, Pickeral OK, Staudt LM: **Signatures of the immune response.** *Immunity* 2001, **15:**375-385.
38. Zhang M, Zhang D, Wells MT: **Generalized shrinkage estimators adaptive to sparsity and asymmetry of high dimensional parameter spaces.** *Submitted* 2007.
39. Zhang M, Montooth KL, Wells MT, Clark AG, Zhang D: **Mapping multiple quantitative trait loci by Bayesian classification.** *Genetics* 2005, **169:**2305-2318.
40. Zhang M, Zhang D, Wells MT: **Variable selection for large *p* small *n* regression models with incomplete data: mapping QTL with epistases.** *Submitted* 2007.
41. Zhang M: **Inference for sparse and asymmetric signals in high dimensional data with applications to statistical genomics.** In *PhD thesis* Cornell University, Department of Biological Statistics and Computational Biology; 2005.
42. Cowles MK, Carlin BP: **Markov Chain Monte Carlo convergence diagnostics: a comparative review.** *J Am Stat Assoc* 1996, **91:**883-904.