Published in final edited form as:
Comput Stat Data Anal. 2012 September 1; 56(9): 2718–2728. doi:10.1016/j.csda.2012.02.017.

Identification of Breast Cancer Prognosis Markers via Integrative Analysis

Shuangge Ma[1,*], **Ying Dai**[1], **Jian Huang**[2], and **Yang Xie**[3]

[1]School of Public Health, Yale University

[2]Department of Statistics & Actuarial Science and Biostatistics, University of Iowa

[3]Department of Clinical Sciences, UT Southwestern Medical Center

Summary

In breast cancer research, it is of great interest to identify genomic markers associated with prognosis. Multiple gene profiling studies have been conducted for such a purpose. Genomic markers identified from the analysis of single datasets often do not have satisfactory reproducibility. Among the multiple possible reasons, the most important one is the small sample sizes of individual studies. A cost-effective solution is to pool data from multiple comparable studies and conduct integrative analysis. In this study, we collect four breast cancer prognosis studies with gene expression measurements. We describe the relationship between prognosis and gene expressions using the accelerated failure time (AFT) models. We adopt a 2-norm group bridge penalization approach for marker identification. This integrative analysis approach can effectively identify markers with consistent effects across multiple datasets and naturally accommodate the heterogeneity among studies. Statistical and simulation studies demonstrate satisfactory performance of this approach. Breast cancer prognosis markers identified using this approach have sound biological implications and satisfactory prediction performance.

Keywords

Breast cancer prognosis; Gene expression; Marker identification; Integrative analysis; 2-norm group bridge

1. Introduction

Amongst women in the US, breast cancer is the most commonly diagnosed malignancy after skin cancer and the second leading cause of cancer deaths after lung cancer. According to the American Cancer Society, in 2009, an estimated 192,370 new cases of breast cancer were diagnosed, and 40,160 died from breast cancer. Women in the US have a 1 in 8 lifetime risk of developing invasive breast cancer, and a 1 in 33 overall chance of dying from it. Various prediction models have been constructed for breast cancer prognosis using clinical risk factors and environmental exposures. Despite their successes, it is now

*shuangge.ma@yale.edu.

commonly accepted that genomic markers have independent predictive power (Cheang et al. 2008; Knudsen 2006).

Multiple profiling studies have been independently conducted, searching for genes whose expressions are associated with breast cancer prognosis. “Breast cancer has probably been the carcinoma most intensively studied by gene expression profiling” (Cheang et al. 2008, p68). In this article, we limit the study to relapse-free survival. Overall and other types of survival are also of interest. However, they have different patterns and different genomic bases and need to be investigated separately. A representative gene expression study on breast cancer prognosis was reported in Sotiriou et al. (2006), which used Affymetrix U133A microarrays and identified 97 genes including UBE2C, PKNA2, TPX2, FOXM1, STK6, CCNA2, BIRC5 and MYBL2. Ivshina et al. (2006) reported similar findings from a concurrent, independent study. Researchers at the Netherlands Cancer Institute identified a 70-gene prognostic signature (van’t Veer et al. 2002). Many genes involving the hallmarks of cancer were included: cell cycle, metastasis, angiogenesis and invasion. This gene signature was then validated on an independent cohort of 295 patients (van de Vijver et al. 2002). We refer to Cheang et al. (2008) for a comprehensive review of related studies.

Published studies have suggested that different prognosis gene signatures may have only moderate or even little overlap. Our data analysis in Section 4 reconfirms this finding. The lack of reproducibility has prevented prognosis gene signatures from being routinely used in clinical practice. Multiple factors may contribute to the lack of reproducibility, including technical variations, functional similarities of multiple genes, incomparability of different studies and others. The most important reason is perhaps small sample sizes of individual studies. For example, the study reported in Sotiriou et al. (2003) profiled 7,650 genes on 98 subjects. Conducting large-scale studies, although ideal, can be prohibitively expensive and time-consuming. Because of the clinical importance of breast cancer prognosis, multiple studies have been independently conducted (Knudsen 2006). A cost-effective way to identify reproducible breast cancer prognosis markers is to pool data from multiple studies.

Available multi-datasets approaches include meta-analysis and integrative analysis approaches. With meta-analysis approaches, multiple datasets are analyzed separately. Then summary statistics (lists of identified markers, effect sizes, p-values) are pooled across multiple datasets. In contrast, integrative analysis approaches pool and analyze raw data from multiple studies. A family of integrative analysis approaches, called “intensity approaches”, search for transformations that make gene expression measurements in different studies (using possibly different platforms) fully comparable. After transformation, multiple datasets are combined and analyzed as if they were from a single study. Such approaches can be limited in that they need to be conducted on a case-by-case basis, and there is no guarantee that the desired transformations always exist.

The goal of this study is to identify important, reproducible markers associated with breast cancer prognosis. This study contains methodological development and integrative analysis of four breast cancer datasets. The proposed method contains the following two main steps. In the first step, we describe the relationship between breast cancer survival and gene expressions using the accelerated failure time (AFT) models. The AFT model describes event time directly and provides a useful alternative to the Cox model (Wei 1992). It has been adopted in Datta et al. (2007), Huang et al. (2006), Schmid and Hothorn (2008) and others for modeling prognosis data with gene expression measurements and shown to have satisfactory performance. A weighted least-squares approach, which is particularly suitable for high dimensional data, is adopted for estimation. In the second step, we adopt a penalization approach for marker selection. In recent statistical literature, there have been many studies investigating penalized marker selection. Commonly adopted penalties include

Lasso, bridge, SCAD, elastic net, MCP and their extensions. Despite their satisfactory properties, most existing approaches are designed for the analysis of single datasets and cannot accommodate the heterogeneity across multiple studies. In this study, we adopt the 2-norm group bridge penalty, which was proposed by Ma et al. (2010) for the analysis of binary data under the logistic regression model, for marker selection.

This study may advance from existing studies along the following directions. Compared with existing breast cancer prognosis studies, it analyzes more datasets, has more power, and thus can generate more reproducible markers. Compared with single-dataset penalization methods, the adopted method can better accommodate heterogeneity across multiple studies. In addition, this study also advances from Ma et al. (2010) by analyzing censored prognosis data under the AFT model. The rest of the article is organized as follows. In Section 2, we first describe the data and model setup and then marker identification using the penalization approach. We also develop an effective computational algorithm. In Section 3, we conduct simulation to better understand performance of the proposed method and compare with alternatives. In Section 4, we analyze four breast cancer prognosis studies, investigate biological implications of the identified markers and evaluate prediction performance. The article concludes with discussion in Section 5. Statistical properties of the proposed approach are described in Appendix.

2. Integrative Analysis and Penalized Marker Selection

2.1 Data and model settings

With data from multiple studies, our goal is to identify the set of genes showing consistent associations with prognosis across studies. In single-dataset analysis, it has been suggested that multiple sets of genes may have equal predictive power for prognosis. As reproducibility is of major concern, we reinforce that the same set of genes are identified. It has been suggested in Rhodes et al. (2004) and Rhodes and Chinnaiyan (2004) that such genes are more likely to represent the essential features of cancer. Although the multiple datasets analyzed share the same markers, it is not appropriate to directly combine them into a single dataset because of the heterogeneity among them. Particularly, in gene expression studies, measurements using different platforms are not directly comparable. One unit increase in cDNA measurement is not directly comparable to one unit increase in Affymetrix measurement. There is no guarantee that cross-platform normalization or transformation (which makes measurements comparable across multiple platforms/studies) always exists. In addition, other confounders may alter the relationship between gene expressions and prognosis.

Assume $M(>1)$ independent studies. For simplicity of notation, assume that the same d covariates (gene expressions) are measured in all studies. In what follows, we use the superscript "(m)" to denote the mth study. Let $T^{(1)}, \ldots, T^{(M)}$ be the logarithms of failure times, and $X^{(1)}, \ldots, X^{(M)}$ be length-d covariates. For $m = 1, \ldots, M$, assume the AFT model

$$T^{(m)}=\alpha^{(m)}+\beta^{(m)\prime}X^{(m)}+\varepsilon^{(m)}. \quad (1)$$

Here $\alpha^{(m)}$ is the unknown intercept, $\beta^{(m)}$ is the regression coefficient, $\beta^{(m)'}$ is the transpose of $\beta^{(m)}$, and $\varepsilon^{(m)}$ is the random error with an unknown distribution. Unlike alternatives such as the Cox or additive risk models, the AFT model describes event time directly and may have a more lucid interpretation. Denote $C^{(1)}, \ldots, C^{(M)}$ as the logarithms of random censoring times. Under right censoring, we observe $(Y^{(m)}, \delta^{(m)}, X^{(m)})$ for $m = 1 \ldots M$. Here $Y^{(m)} = \min(T^{(m)}, C^{(m)})$ and $\delta^{(m)} = I(T^{(m)} \leq C^{(m)})$.

Consider $\beta = (\beta^{(1)}, \ldots, \beta^{(M)})$, the $d \times M$ regression coefficient matrix. The main characteristics of β are as follows. First, $\beta^{(m)}$s are sparse in that only a subset are nonzero. This feature corresponds to the fact that out of a large number of genes surveyed, only a subset are associated with prognosis, and the rest are noises. Only cancer-associated genes have nonzero regression coefficients. Second, $\beta^{(1)}, \ldots, \beta^{(M)}$ have the same sparsity structure. That is, elements of β in the same row are either all zero or all nonzero. This feature corresponds to the fact that multiple studies share the same set of markers. Third, for cancer markers with nonzero coefficients, the values of regression coefficients may be different across studies, which can accommodate the heterogeneity across studies.

2.2 Weighted least squares estimation

In the literature, several approaches have been proposed for estimation with the AFT model (Buckley and James 1979; Ying 1993). Among them, the weighted least squares approach (Stute 1993) may have the least computational cost and is thus more suitable for high dimensional gene expression data. In study m(= 1, …, M), assume $n^{(m)}$ iid observations $(Y_i^{(m)}, \delta_i^{(m)}, X_i^{(m)})$, $i = 1 \ldots n^{(m)}$. Let $\hat{F}^{(m)}$ be the Kaplan-Meier estimate of $F^{(m)}$, the distribution function of $T^{(m)}$. It can be computed as $\hat{F}^{(m)}(y) = \sum_{i=1}^{n^{(m)}} w_i^{(m)} I(Y_{(i)}^{(m)} \le y)$. Here $Y_{(1)}^{(m)} \le \ldots \le Y_{(n^{(m)})}^{(m)}$ are the order statistics of $Y_i^{(m)}$s. Denote $\delta_{(1)}^{(m)}, \ldots, \delta_{(n^{(m)})}^{(m)}$ as the associated censoring indicators and $X_{(1)}^{(m)}, \ldots, X_{(n^{(m)})}^{(m)}$ as the associated covariates. $w_i^{(m)}$s are the jumps in the Kaplan-Meier estimate and can be computed as

$$w_1^{(m)} = \frac{\delta_{(1)}^{(m)}}{n^{(m)}}, \quad \text{and } w_{(i)}^{(m)} = \frac{\delta_{(i)}^{(m)}}{n^{(m)} - i + 1} \prod_{j=1}^{i-1} \left(\frac{n^{(m)} - j}{n^{(m)} - j + 1} \right)^{\delta_{(j)}^{(m)}} \quad \text{for } i = 2, \ldots, n^{(m)}.$$

For study m, the weighted least squares objective function is defined as

$$R^{(m)} = \frac{1}{2} \sum_{i=1}^{n^{(m)}} w_i^{(m)} \left(Y_{(i)}^{(m)} - \alpha^{(m)} - \beta^{(m)\prime} X_{(i)}^{(m)} \right)^2.$$

We center $X_{(i)}^{(m)}$ and $Y_{(i)}^{(m)}$ as

$$X_{(i)}^{(m)*} = \sqrt{w_i^{(m)}} \left(X_{(i)}^{(m)} - \frac{\sum_j w_j^{(m)} X_{(j)}^{(m)}}{\sum_j w_j^{(m)}} \right) \text{ and } Y_{(i)}^{(m)*} = \sqrt{w_i^{(m)}} \left(Y_{(i)}^{(m)} - \frac{\sum_j w_j^{(m)} Y_{(j)}^{(m)}}{\sum_j w_j^{(m)}} \right).$$

We define the overall loss function by

$$R(\beta) = \sum_{m=1}^{M} R^{(m)} = \sum_{m=1}^{M} \frac{1}{2} \sum_{i=1}^{n^{(m)}} \left(Y_{(i)}^{(m)*} - \beta^{(m)\prime} X_{(i)}^{(m)*} \right)^2. \quad (2)$$

2.3 Penalized marker selection

Denote $\beta_j^{(m)}$ as the jth component of $\beta^{(m)}$. $\beta_j = (\beta_j^{(1)}, \ldots, \beta_j^{(M)})$ is the jth row of β and represents the coefficients of covariate j across M studies. Define

$$\hat{\beta}=argmin_{\beta}\left\{R(\beta)+\lambda_n\sum_{j=1}^{d}J(\beta_j)\right\}, \quad (3)$$

where λ_n is a data-dependent tuning parameter. For the penalty function $J(\cdot)$, we adopt the 2-norm group bridge penalty recently proposed by Ma et al. (2010), where

$$J(\beta_j)=\|\beta_j\|_2^{\gamma}. \quad (4)$$

Here $\|\beta_j\|_2=\left[(\beta_j^{(1)})^2+\ldots+(\beta_j^{(M)})^2\right]^{1/2}$ and $0 < \gamma < 1$ is the fixed bridge index. In numerical studies, we set $\gamma = 1/2$. Ma et al. (2010) investigates the 2-norm group bridge penalty with binary data and logistic regression models. In this article, we extend it to prognosis data and AFT models.

Adopting the 2-norm group bridge penalty has been motivated by the following considerations. When $M = 1$, it simplifies to the bridge penalty, which has been shown to have the "oracle" properties in the analysis of single datasets (Huang et al. 2008). In integrative analysis, for a specific gene, we need to evaluate its overall effects in multiple datasets. To achieve such a goal, we treat its M regression coefficients as a *group* and conduct group-level selection. When $\gamma = 1$, the 2-norm group bridge penalty becomes the group Lasso (GLasso, Meier et al. 2008). Theoretical investigation in Appendix and simulation in Section 3 show that the 2-norm group bridge penalty has significantly better selection property than the GLasso. The 2-norm group bridge penalty is also related to but differs significantly from the 1-norm group bridge penalty in Huang et al. (2009). The 1-norm group bridge penalty is designed for *bi-level* selection. In integrative analysis, as multiple datasets share the same sparsity structure, the within-group selection is undesired. The most significant difference between this study and Huang and Ma (2010) and others is data structure. Most penalized marker selection studies focus on single datasets, whereas multiple heterogeneous datasets are analyzed in this study. In addition, in other studies, the grouping structure comes from dummy variables for single covariates or clusters of covariates. In contrast, in this study, one group represents the effects of one covariate in multiple studies.

2.4 Computational algorithm

As the 2-norm group bridge penalty is not convex, direct minimization of the objective function can be difficult. Consider the following computational algorithm. Denote $\eta = (1 - \gamma)/\gamma$. Define $S(\beta)=R(\beta)+\sum_{j=1}^{d}\theta_j^{-\eta}\|\beta_j\|_2+\tau\sum_{j=1}^{d}\theta_j$, where $\tau=\eta^{-1/\eta}(1+\eta)^{-(1+\eta)/\eta}\lambda_n^{(1+\eta)/\eta}$. $\hat{\beta}$ minimizes the objective function defined in (3) if and only if

$$\hat{\beta}=argmin_{\beta}S(\beta) \text{ subject to } \theta_j \geq 0\ (j=1,\ldots,d).$$

This result can be proved using Proposition 1 of Huang et al. (2009). Based on this result, we propose the following algorithm. For a fixed λ_n,

1. Initialize $\hat{\beta}$ as the GLasso estimate, i.e., estimate defined in (3) with $\gamma = 1$;
2. Compute $\theta_j=(\eta/\tau)^{1/(1+\eta)}\|\hat{\beta}_j\|_2^{1/(1+\eta)}$ for $j = 1, \ldots, d$;

3. Compute $\hat{\beta}=argmin\{R(\beta)+\sum_{j=1}^{d}\theta_j^{-\eta}\|\beta_j\|_2\}$;

4. Repeat steps 2–3 until convergence.

In Theorem 1 (Appendix), we show that with a high probability, the GLasso can select all true positives while effectively removing the majority of true negatives. In addition, it is estimation consistent. Thus, it is an appropriate choice for the initial estimate. However, the GLasso tends to over-select, and thus the downstream iterations are needed. In Step 3, we transform the group bridge-type minimization to a weighted GLasso-type minimization. The iteration continues until convergence. In our numerical studies, we use the ℓ_2 norm of the difference between two consecutive estimates less than 0.01 as the convergence criterion, and convergence is achieved within ten iterations. We use the coordinate descent algorithm described in Friedman et al. (2010) to compute the GLasso estimate. An interesting observation is that, for a fixed j and any $k \neq l$, $\partial^2 R(\beta)/\partial\beta_j^{(k)}\partial\beta_j^{(l)}=0$. Thus the Hessian for the coefficients in a single group is a diagonal matrix. The unique form of the objective function makes the coordinate descent algorithm computationally less expensive. Research code written in R is available from the authors.

The tuning parameter λ_n balances sparsity and goodness-of-fit. With a smaller λ_n, more genes are identified as associated with prognosis. We adopt V-fold cross validation for tuning parameter selection. We have numerically experimented with several other tuning parameter selection techniques, including BIC, AIC and Leave-One-Out cross validation (results omitted). We find that performance of other tuning parameter selection techniques is comparable to or worse than that of V-fold cross validation. We choose V-fold cross validation because of its computational simplicity. With V-fold cross validation, V can be viewed as another "tuning parameter". Our literature search does not suggest an objective way of selecting V. When the sample size is not too small, our limited experience suggests that $V = 4 – 10$ lead to similar results. In our numerical studies, we set $V = 4$. It is advised that small values of V should be considered when the sample size is small.

3. Simulation Studies

For simplicity of notation, we have assumed matched gene sets across multiple studies. When different sets of genes are measured in different studies, we use the following rescaling approach. Assume that gene 1 is measured only in the first $K(< M)$ studies. We set $\beta_1^{(K+1)}=\ldots=\beta_1^{(M)}=0$ and $J(\beta_1)=[(\beta_1^{(1)})^2+\ldots+(\beta_1^{(K)})^2]^{\gamma/2} \times (M/K)^{\gamma}$. The proposed approach and computational algorithm are then applicable with minor modifications.

We simulate data for four independent studies, each with 50 or 100 subjects. We simulate 50 or 100 gene clusters, with 20 genes in each cluster. Thus, the total number of gene expressions simulated is 1,000 or 2,000. We first simulate gene expressions from multivariate normal distributions with marginal means zero and variances one. Genes in different clusters have independent expressions. For genes within the same clusters, their expressions have the following correlation structures: (i) auto-regressive correlation, where expressions of genes j and k have correlation coefficient $\rho^{|j-k|}$; (ii) banded correlation, where expressions of genes j and k have correlation coefficient $\max(0, 1 - |j - k| \times \rho)$; and (iii) compound symmetry, where expressions of genes j and k have correlation coefficient ρ when $j \neq k$. Under each correlation scenario, we consider two different ρ values. We then add floor −3 and ceiling 3 to satisfy the boundedness requirement. Within each of the first four clusters, there are five genes associated with the responses. There are thus a total of twenty important genes, and the rest are noises. For important genes, we generate their regression coefficients from $Unif[-1, -0.5] \cup Unif[0.5, 1]$. 20% important and 10% noisy

genes are only measured in two studies. We generate the (log) event time from the AFT model with intercept equal to zero. The censoring time is generated independent of event. We adjust the censoring time so that the censoring rate is ~ 50%. The simulation settings closely mimic real cancer prognosis studies, where genes have the pathway structures. Genes within the same pathways tend to have correlated expressions, whereas genes within different pathways tend to have weakly correlated or independent expressions. Among a large number pathways, only a few are associated with prognosis. Within those important pathways, there are some important genes and others are noises.

To better gauge performance of the proposed approach, we also consider the following alternative approaches. (a) Meta-analysis. We analyze each dataset separately. Genes that are identified in at least one study are identified in meta-analysis. An alternative is to consider genes identified in all studies. However, we have examined all simulation settings and found that there are very few such genes. When analyzing each dataset, we consider the following approaches: (a.1) Lasso, (a.2) one-step and (a.3) bridge. When analyzing a single dataset using the bridge approach, it can be shown that the bridge estimate can be computed using an iterative approach similar to that described in Section 2.4. Following similar proof as for Theorem 2 (Appendix), it can be proved that the one-step estimate obtained after one iteration is selection consistent; (b) An intensity approach. Since all four datasets are generated under similar settings, we adopt an intensity approach, make transformations of gene expressions, combine the four datasets and analyze as if they were from a single study. For the combined dataset, we analyze using (b.1) Lasso, (b.2) one-step and (b.3) bridge approaches; (c) Integrative analysis with (c.1) GLasso and (c.2) one-step approaches. For all approaches, we select the tuning parameters using 4-fold cross validation.

Simulation suggests that the proposed approach is computationally affordable. Analysis of one replicate takes about five minutes on a regular desktop PC. Summary statistics based on 200 replicates are shown in Table 1. We can see that although meta-analysis approaches can identify the majority or all of the true positives, they also identify a large number of false positives. Intensity approaches can significantly outperform meta-analysis approaches. The satisfactory performance of intensity approaches is not surprising, considering that the four simulated datasets are very similar to each other – the degree of similarity is much higher than that encountered in practical data analysis. Integrative analysis approaches outperform alternatives by identifying the majority or all of the true positives and a smaller number of false positives. Among the three integrative analysis approaches, the proposed approach identifies the smallest number of false positives, at the price of a very small number of false negatives.

4. Identification of Breast Cancer Markers

We collect and analyze four breast cancer prognosis studies with microarray gene expression measurements. The same datasets have been analyzed in Shen et al. (2004) and Ma and Kosorok (2010). Previous studies have examined the study designs and concluded that they are comparable and can be pooled for analysis. Analysis in this study differs significantly from that in Shen et al. (2004) and Ma and Kosorok (2010). Specifically, the two previous studies focus on the *marginal* effects of genes. In contrast, in this study, we investigate the *combined* effects of multiple genes, which may better describe the biological mechanisms of breast cancer.

We provide brief descriptions of the four studies in Table 2 and refer to the original publications for more detailed information. Among the four datasets, two used cDNA, one used oligonucleotide arrays, and one used Affymetrix genechips for profiling. We first conduct normalization of gene expressions for each dataset separately, using a lowess

approach for cDNA data and an RMA (robust multichip average) approach for the others. With Affymetrix chips, the measurements are log2 transformed. We fill in missing expressions with means across samples. We then standardize each gene expression to have zero mean and unit variance. The proposed approach does not require the direct comparability of measurements from different studies. Thus additional cross-study transformation or normalization is not needed. We match genes in the four studies using their Unigene Cluster IDs. Although the proposed approach can accommodate partially matched gene sets, to improve reliability, we focus on the 2,555 genes that are measured in all four studies. As it is expected that the number of prognosis-related genes to be much smaller than 2,555, focusing on the common set is expected to have negligible impact.

4.1 Prognosis markers

We apply the proposed approach and identify 22 genes as breast cancer prognosis markers. Gene names and corresponding estimates are provided in Table 3. Two main factors may contribute to the small regression coefficients observed in Table 3. First, it has been suggested that even though gene expressions have independent predictive power, they can explain only a small fraction of variation in prognosis. Second, with penalization methods and extremely high dimensional data, shrinkage (towards zero) has been commonly observed. It is worth noting that when predicting relative survival risk, shrinkage is not of serious concern.

We search NCBI and find that some of those identified genes have sound biological implications. For example, gene PPOX encodes the penultimate enzyme of heme biosynthesis, which catalyzes the 6-electron oxidation of protoporphyrinogen IX to form protoporphyrin IX. Mutations in this gene cause variegate porphyria, an autosomal dominant disorder of metabolism. Gene MLLT4, also known as AF6, is a Ras target that regulates cell-cell adhesions downstream of Ras activation. It is fused with MLL in tumors caused by t(6; 11) translocations (Taya et al. 1998). It encodes the adadin protein. It has been shown that loss of adadin protein expression is associated with poor outcome in breast cancer (Letessier et al. 2007). Gene MEIS2 encodes a homeobox protein belonging to the TALE (three amino acid loop extension) family of homeodomain-containing proteins. TALE homeobox proteins are highly conserved transcription regulators, and several members have been shown to be essential contributors to cancer developmental programs. Gene MB encodes a member of the globin superfamily, which is a haemoprotein contributing to intracellular oxygen storage and transcellular facilitated diffusion of oxygen. Kristiansen et al. (2010) showed that myoglobin mRNA was found in a subset of breast cancer cell lines. In microdissected tumors, MB transcript was markedly upregulated. In addition, 71% breast tumors displayed MB protein expression, which is in significant correlation with a positive hormone receptor status and better prognosis. In silico data mining also confirmed higher MB levels in luminal-type breast cancer. The protein encoded by gene ENOX2 is a growth-related cell surface protein. It reacts with the monoclonal antibody KI in cells, such as the ovarian carcinoma line OVCAR-3. 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 tumor growth (Li and Jiang 2010). The protein encoded by gene GRB2 binds the epidermal growth factor receptor and contains one SH2 domain and two SH3 domains. This gene is similar to the Sem5 gene of C.elegans, which is involved in the signal transduction pathway. Expression of this gene has been implied in multiple cancers including endometrial cancer, non-small cell lung cancer and prostate cancer. Annexin I belongs to a family of Ca(2+)-dependent phospholipid binding proteins. Since phospholipase A2 is required for the biosynthesis of the potent mediators of inflammation, prostaglandins and leukotrienes, annexin I may have

potential anti-inflammatory activity. Maschler et al. (2010) identified Annexin A1 as having important functions in intracellular vesicle trafficking and as an efficient suppressor of EMT and metastasis in breast cancer. It was found that AnxA1 levels were strongly reduced in EMT of mammary epithelial cells, in metastatic murine and human cell lines and in metastatic mouse and human carcinomas. Gene IL6 encodes a cytokine that functions in inflammation and maturation of B cells. The functioning of this gene is implicated in a wide variety of inflammation-associated disease states. It has been identified as a susceptibility gene of multiple cancers.

For the identified genes, we search KEGG for their pathway information. We find that many hallmarks of cancer are presented, including metabolic pathways (KEGG: 01100), MAPK signaling pathway (KEGG: 04010), apoptosis (REACT: 578), Focal adhesion (KEGG: 04510), Signaling by EGFR (REACT: 9417), Pathways in cancer (KEGG: 05200), Toll-like receptor signaling pathway (KEGG: 04620) and others.

4.2 Analysis with alternative methods

We also analyze data using the alternative methods described in Section 3. The numbers of genes identified and overlap with the proposed approach are presented in Table 4. As seen in simulation, meta-analysis approaches identify a relatively large number of genes, with small overlap among the sets of genes identified in different datasets. Both the intensity and integrative analysis approaches identify a small number of genes. The proposed approach identifies genes significantly different from those identified using alternatives.

With practical data, it is difficult to objectively evaluate marker identification accuracy. As an alternative, we evaluate prediction performance, which may provide an indirect evaluation of gene identification accuracy. It is expected that if the identified markers are more meaningful, prediction using those markers is more accurate. Specifically, we first split each dataset randomly into a training set and a testing set with sizes 3:1. We construct the estimate using the training set only and then make prediction for subjects in the testing set. Based on the predicted $\hat{\beta}^{(m)'}X^{(m)}$, we generate two risk groups with equal sizes. The logrank statistic is computed to evaluate the difference between survival of the two groups. For each random split, we compute the mean logrank statistic over four datasets. To avoid an extreme split, we repeat the whole process 50 times, compute the mean logrank statistics and present the results in Table 4. The proposed approach has the best prediction performance with the logrank statistic equal to 5.930 (p-value 0.015).

5. Discussion

In breast cancer prognosis studies with gene expression measurements, markers identified from the analysis of single datasets have suffered a lack of reproducibility. Multiple factors may contribute to the low reproducibility, including technical variations, high correlations and functional similarities among genes, incomparability of cohorts, small sample sizes of individual studies and others. In this article, we pool and conduct integrative analysis with data from four independent studies. Analysis of multiple studies is inevitably more complicated. Additional considerations may include the selection of comparable studies, interpretation of analysis results and utilization of identified markers. We acknowledge the importance of those issues. However as there are established guidelines (Guerra and Goldstein 2009), we will not reiterate discussions on such issues. The four studies we analyze were conducted in a similar time period and with similar patient selection criteria. Although there are several other studies falling into the category of “breast cancer prognosis studies”, not all of them have data publicly available or have similar patients characteristics. We adopt the AFT model to describe prognosis. Compared with alternatives, the AFT model may have a more lucid interpretation. Extension to other survival models is nontrivial and

will be postponed to future studies. Because of a lack of model diagnostics techniques for extremely high dimensional data, the AFT models are not validated. For marker identification, we adopt the 2-norm group bridge penalization approach, which reinforces that multiple datasets identify the same set of markers. With data analyzed in this study, such a strategy can be reasonable. However, with other data, this can be too restricted. For example because of the heterogeneity caused by confounders, datasets generated under similar designs may have overlapping but different sets of markers. Different penalization methods will be needed to accommodate such a scenario.

Simulation study shows satisfactory performance of the proposed method. We note that the simulation settings are simpler than what is encountered in practice. As our goal is to demonstrate improvement over existing methods, such settings can be sufficient. In simulation, there are a relatively small number of signals. With the proposed method, the number of selected markers is limited by sample size. In the theoretical development (Appendix), it is assumed that the number of signals is fixed as the sample size and number of covariates increase. The proposed method is capable of accommodating a limited number of moderate to large signals, but not a very large number of small signals. This limitation is shared by many existing penalization methods. The proposed approach identifies 22 genes as prognosis markers, many of which have sound biological implications. We note that although some genes have been previously identified as breast cancer prognosis markers, this may be the first time they are identified in an integrative analysis context. In addition, there are also new findings that need further investigation. With limited knowledge of breast cancer genomics, it is still hard to objectively quantify the accuracy of marker identification. Cross validation-based prediction evaluation shows that the proposed approach and identified markers have satisfactory prediction performance. Although it does not use completely independent data, different approaches are compared on the same ground, and thus the comparison result is expected to be reasonably fair. As with other penalized marker identification studies, this study also has limitations. For example, the proposed integrative analysis approach cannot fully separate passenger genes from drivers. In addition, the identified markers need to be confirmed by independent prospective studies before any clinical usage. We study breast cancer relapse-free survival. Other types of breast cancer survival and other types of cancers can also be studied using the proposed integrative analysis method.

Acknowledgments

The authors would like to thank the associate editor and a referee for careful review and insightful comments. This study has been supported by awards CA120988, CA152301 and CA142774 from NIH and DMS-0904181 from NSF.

References

Buckley J, James I. Linear regression with censored data. Biometrika. 1979; 66:429–436.

Cheang M, van de Rijn M, Nielsen TO. Gene expression profiling of breast cancer. Annual Review of Pathology: Mechanisms of Disease. 2008; 3:67–97.

Datta S, Le-Rademacher J, Datta S. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. Biometrics. 2007; 63:259–271. [PubMed: 17447952]

Friedman, J.; Hastie, T.; Tibshirani, R. A note on the group lasso and a sparse group lasso. 2010. http://arxiv.org/abs/1001.0736

Guerra, R.; Goldstein, DR. Meta-Analysis and Combining Information in Genetics and Genomcis. Chapman and Hall/CRC; 2009. 2008

Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT. Gene expression predictors of breast cancer outcomes. Lancet. 2003; 361:1590–1596. [PubMed: 12747878]

Huang J, Horowitz JL, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Annals of Statistics. 2008; 36:587–613.

Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge penalty. Lifetime Data Analysis. 2010; 16:176–195. [PubMed: 20013308]

Huang J, Ma S, Xie H. Regularized estimation in the accelerated failure time model with high dimensional covariates. Biometrics. 2006; 62:813–820. [PubMed: 16984324]

Huang J, Ma S, Xie H, Zhang C. A group bridge approach for variable selection. Biometrika. 2009; 96:339–355. [PubMed: 20037673]

Ivshina AV, George J, Senko O, Mow B, Putti TC, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Research. 2006; 66:10292–10301. [PubMed: 17079448]

Knudsen, S. Cancer Diagnostics with DNA microarrays. Wiley; 2006.

Kristiansen G, Rose M, Geisler C, et al. Endogenous myoglobin in human breast cancer is a hallmark of luminal cancer phenotype. Br J Cancer. 2010; 102:1736–1745. [PubMed: 20531416]

Letessier A, Garrido-Urbani S, Ginestier C, Fournier G, Esterni B, Monville F, Adelaide J, Geneix J, Xerri L, Dubreuil P, Viens P, Charafe-Jauffret E, Jacquemier J, Birnbaum D, Lopez M, Chaffanet M. Correlated break at PARK2/FRA6E and loss of AF-6/Afadin protein expression are associated with poor outcome in breast cancer. Oncogene. 2007; 26:298–307. [PubMed: 16819513]

Li T, Jiang S. Effect of bFGF on invasion of ovarian cancer cells through the regulation of Ets-1 and urokinase-type plasminogen activator. Pharm Biol. 2010; 48:161–165. [PubMed: 20645833]

Ma S, Huang J. Regularized gene selection in cancer microarray meta-analysis. BMC Bioinformatics. 2009; 10:1. [PubMed: 19118496]

Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional datasets. Biostatistics. 2010 *In Press*.

Maschler S, Gebeshuber CA, Wiedemann EM, Alacakaptan M, Schreiber M, Custic I, Beug H. Annexin A1 attenuates EMT and metastatic potential in breast cancer. EMBO Mol Med. 2010; 2:401–414. [PubMed: 20821804]

Meier L, van de Geer S, Buhlmann P. The group Lasso for logistic regression. JRSSB. 2008; 70:53–71.

Rhodes D, Chinnaiyan AM. Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers. Annals of the New York Academy of Sciences. 2004; 1020:32–40. [PubMed: 15208181]

Rhodes D, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. Large-scale meta-analysis of cancer microarray data identified common transcriptional profiles of neoplastic transformation and progression. PNAS. 2004; 101:9309–9314. [PubMed: 15184677]

Schmid M, Hothorn T. Flexible boosting of accelerated failure time models. BMC Bioinformatics. 2009; 9:269. [PubMed: 18538026]

Shen R, Ghosh D, Chinnaiyan AM. Prognostic meta signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC Genomics. 2004; 5:94. [PubMed: 15598354]

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS. 2001; 98:10869–10874. [PubMed: 11553815]

Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET. Breast cancer classification and prognosis based on gene expression profiles from a population based study. PNAS. 2003; 100:10393–10398. [PubMed: 12917485]

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. JNCI. 2006; 98:262–272. [PubMed: 16478745]

Stevens JR, Doerge RW. Meta-analysis combines Affymetrix microarray results across laboratories. Comparative and Functional Genomics. 2005; 6:116–122. [PubMed: 18629222]

Stute W. Consistent estimation under random censorhip when covariables are available. Journal of Multivariate Analysis. 1993; 45:89–103.

Taya S, Yamamoto T, Kano K, Kawano Y, Iwamatsu A, Tsuchiya T, Tanaka K, Kanai-Azuma M, Wood SA, Mattick JS, Kaibuchi K. The Ras target AF-6 is a substrate of the fam deubiquitinating enzyme. J Cell Biol. 1998; 142:1053–1062. [PubMed: 9722616]

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–536. [PubMed: 11823860]

van de Vijver MJ, He YD, van't Veer MJ, Dai H, Hart AA, et al. A gene expression signature as a predictor of survival in breast cancer. NEJM. 2002; 347:1999–2009. [PubMed: 12490681]

Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in Medicine. 1992; 11:1871–1879. [PubMed: 1480879]

Ying ZL. A large sample study of rank estimation for censored regression data. Annals of Statistics. 1993; 21:76–99.

Appendix: Asymptotic Properties

Here we establish the selection consistency property of the proposed approach. In the computational algorithm described in Section 2.4, the GLasso estimate is used as the starting value. Covariates not selected by the GLasso will not be selected by the proposed approach. In what follows, we first state properties of the GLasso estimate. Main results include estimation consistency and that with probability converging to one, all important covariates are selected. Then we state the result that the one-step estimate obtained after one iteration is selection consistent. Following this result, it can be proved that the estimate from a finite number of iterations is selection consistent.

The GLasso estimate

The GLasso estimate is defined as

$$\tilde{\beta}=argmin_{\beta}\left\{R(\beta)+\lambda_n\sum_{j=1}^{d}\|\beta_j\|_2\right\}$$

with $\tilde{\beta}=(\tilde{\beta}_1^{'},\ldots,\tilde{\beta}_d^{'})'$. Let $\tilde{A}_1=\{j:\|\tilde{\beta}_j\|_2\neq 0\}$ be the set of GLasso selected covariates. Define $n=\sum_m n^{(m)}$. For $m=1,\ldots,M$ and $j=1,\ldots,d$, let $X_{\cdot j}^{(m)}=(X_{(1)j}^{(m)*},\ldots,X_{(n^{(m)})j}^{(m)*})'$ be the $n^{(m)}\times 1$ vector of the jth centered and standardized covariate vector in the mth dataset. For any $A\subseteq\{1,\ldots,d\}$, denote $X_A^{(m)}=(X_{\cdot j}^{(m)*}:j\in A), 1\le m\le M,$ and $X_A=(X_A^{(1)'},\ldots,X_A^{(M)'})'$. When $A=\{1,\ldots,d\}$, we simply write $X^{(m)}$ for $X_A^{(m)}$ and X for X_A. Define $\Sigma_A=n^{-1}X_A^{'}X_A=n^{-1}(X_A^{(1)'}X_A^{(1)}+\cdots+X_A^{(M)'}X_A^{(M)})$, $A\subseteq\{1,\ldots,d\}$. Denote the cardinality of A by $|A|$. Define $c_{\min}(l)=\min_{|A|=l,\|\nu\|_2=1}\nu'\Sigma_A\nu$, $c_{\max}(l)=\max_{|A|=l,\|\nu\|_2=1}\nu'\Sigma_A\nu$, where $\nu\in\mathbb{R}^q$. Matrix X satisfies the sparse Riesz condition (Zhang and Huang 2008), or SRC, with rank r and spectrum bounds $0<c_*<c^*<\infty$ if

$$c_*\le c_{\min}(r)\le c_{\max}(r)\le c^*,\forall A \text{ with } |A|=r. \quad (1)$$

Since $\|X_A\nu\|^2/n = \nu'\Sigma_A\nu$, all the eigenvalues of Σ_A are inside the interval $[c_*, c^*]$ under (1) when the size of A is not greater than q. The SRC ensures that the models with dimension lower than q are identifiable. Let ρ_n be the maximum of eigenvalues of matrices $X^{(m)}X^{(m)'}$, $1 \le m \le M$. Let β_0 be the true value of β. Denote the set of nonzero regression coefficients by $A^o = \{j : \|\beta_{0j}\|_2 \neq 0, 1 \le j \le d\}$. Let $q = |A^o|$ and let $b_{n2} = \max\{\|\beta_{0j}\|_2 : j \in A^o\}$ be the largest L_2 norm of the nonzero β_{0j}s. We assume

(A1) The number of nonzero coefficients q is finite.

(A2) (a) The observations $(Y_i^{(m)}, X_i^{(m)}, \delta_i^{(m)})$, $1 \le i \le n^{(m)}$ are independent and identically distributed; (b) The errors $\varepsilon_1^{(m)}, \ldots, \varepsilon_{n^{(m)}}^{(m)}$ are independent and identically distributed with mean 0 and finite variance. Furthermore, they are subgaussian, in the sense that there exist K_1, $K_2 > 0$ such that the tail probabilities of ε_i satisfy $P(|\varepsilon_i^{(m)}| > x) \le K_2 \exp(-K_1 x^2)$ for all $x \ge 0$ and all i and m.

(A3) (a) The errors $(\varepsilon_1^{(m)}, \ldots, \varepsilon_{n^{(m)}}^{(m)})$ are independent of the Kaplan-Meier weights $(w_1^{(m)}, \ldots, w_{n^{(m)}}^{(m)})$; (b) The covariates are bounded. That is, there is a constant $C > 0$ such that $|X_{ij}^{(m)}| \le C$, $1 \le i \le n^{(m)}$, $1 \le j \le d$, $1 \le m \le M$.

(A4) The covariate matrix satisfies the sparse Riesz condition (SRC) with rank q^*: there exist constants $0 < c_* < c^* < \infty$, such that for $q^* = (3 + 4C)q$ and $C = c^*/c_*$, with probability converging to 1, $c_* \le \dfrac{\nu'\Sigma_A\nu}{\|\nu\|_2^2} \le c^*$, $\forall A$ with $|A| = q^*$ and $\nu \in \mathbb{R}^{q^*}$.

Theorem 1. Suppose that assumptions (A1)–(A4) hold. (i) Let $\tilde{A} = \{j : \|\tilde{\beta}_j\|_2 \neq 0\}$. *Then, with probability one,* $|\tilde{A}| \le C_1 q$ *for a constant* $C_1 > 0$. *(ii) If* $\lambda_n = O(\sqrt{\log(d)/n})$, *then* $\|\tilde{\beta} - \beta_0\|_2 \equiv \left\{\sum_{m=1}^{M}\sum_{j=1}^{d}(\tilde{\beta}_j^{(m)} - \beta_{0j}^{(m)})^2\right\}^{1/2} = O_p(\sqrt{\log d/n})$. *(iii) Let* $b_{n1} = \min\{\|\beta_{0j}\|_2 : j \in A^o\}$. *Suppose that* $b_{n1}/\sqrt{\log d/n} \to \infty$. *Then all covariates with nonzero coefficients are selected by the GLasso with probability converging to one.*

Part (i) provides an upper bound for the dimension of the GLasso model. In particular, the number of nonzero estimates is at most a finite multiply of the number of nonzero coefficients. Part (ii) shows that the rate of convergence is $\sqrt{\log d/n}$ with a proper choice of λ_n. Part (iii) implies that all covariates with nonzero coefficients are selected with a high probability. This justifies using the GLasso as the initial estimate in the proposed computational algorithm.

Proof. A main tool used in the proof is the maximal inequality stated in the following lemma. For $1 \le m \le M$, let $\tau^{(m)} = (\tau_1^{(m)}, \ldots, \tau_{n^{(m)}}^{(m)})'$ where $\tau_i^{(m)} = w_i^{(m)}\varepsilon_{(i)}^{(m)}$.

Lemma 1. Suppose that conditions (A2) and (A3) hold. Let $\xi_j^{(m)} = \sum_{i=1}^{n^{(m)}} X_{ij}^{(m)}\tau_i^{(m)}$, $1 \le j \le d$. Let $\xi_n = \max_{1 \le m \le M, 1 \le j \le d}|\xi_j^{(m)}|$. Then

$$E(\xi_n) \le C_1\sqrt{\log(d)}\left(\sqrt{2C_2 n\log(d)} + 4\log(2d) + C_2\right)^{1/2},$$

where C_1 and C_2 are two positive constants. In particular, when $\log(d)/n \to 0$,

$$\mathrm{E}(\xi_n)=O(1)\sqrt{n\log d}.$$

Proof of this lemma can be found in Huang and Ma (2010).

Part (i) of Theorem 1 mainly follows from the proof of Theorem 1 of Wei and Huang (2010). The difference is that here we use the sub-gaussian assumption to control certain tail probabilities, as opposed to the normality condition assumed in Wei and Huang (2010). Since sub-gaussian random variables have the same tail behavior as normal random variables, the argument of Wei and Huang (2010) goes through. Part (ii) follows from part (iii) and the assumption that the number of nonzero coefficients is fixed. Thus the absolute values of the nonzero coefficients are bounded away from 0 by a positive constant independent of n. Part (iii) can be proved in a way similar to that in the proof of Theorem 1 in Huang and Ma (2010).

The iterative estimate

Consider $\hat{\beta}$, the one-step estimate after one iteration in the algorithm described in Section 2.4. For simplicity of notation, set $\gamma = 1/2$.

Simple algebra shows that θ_j computed in Step 2 of the proposed algorithm is $\theta_j=(2/\lambda_n)\|\tilde{\beta}_j\|_2^{1/2}$. Thus the one-step estimator is

$$\hat{\beta}=argmin_\beta\left\{R(\beta)+\frac{\lambda_n}{2}\sum_{j=1}^{d}\|\tilde{\beta}_j\|_2^{-1/2}\|\beta_j\|_2\right\}.$$

With the convention $0\times\infty=0$, $\hat{\beta}_j$ will be set as 0 if $\|\tilde{\beta}_j\|_2=0$.

In addition to (A1)–(A4), we further assume

(A5) Denote $r_n=\sqrt{\log d/n}$. $\{q, d, b_{n1}, r_n, \lambda_n\}$ satisfy

$$\frac{\sqrt{\log q}}{b_{n1}\sqrt{n}}+\frac{\sqrt{n\log(d-q)}}{\lambda_n r_n}+\frac{\sqrt{q}\lambda_n}{nb_{n1}}\to 0.$$

This condition restricts the number of covariates with zero and nonzero coefficients, the penalty parameter and the smallest norm of nonzero coefficients. Since only the logarithm of d enters the equation, the results are applicable to models with dimension much larger than n. There are two special cases where condition (A5) is especially simple. The first is in a conventional model with fixed d. Then b_{n1} is bounded away from zero. In this case, (A5) is satisfied if $\lambda_n/n\to 0$ and $\lambda_n r_n/n^{1/2}\to\infty$. The second case is when the number of nonzero coefficients q is fixed, but d is larger than n. This is a reasonable assumption for cancer genomic studies where the total number of genes surveyed is larger than n, but the number of genes associated with cancer is small. In this case, b_{n1} is bounded away from zero. (A5) is satisfied if $\lambda_n/n\to 0$ and $\log d=o(1)(\lambda_n r_n/n^{1/2})^2$. Therefore, depending on r_n and λ_n, the total number of covariates can be as large as $\exp(n^a)$ for some $0<a<1$.

For $\hat{\beta}\equiv(\hat{\beta}_1',\ldots,\hat{\beta}_d')'$ and $\beta_0\equiv(\beta_{01}',\ldots,\beta_{0d}')'$, $\mathrm{sgn}_0(\hat{\beta})=\mathrm{sgn}_0(\beta_0)$ if $\mathrm{sgn}_0(\|\hat{\beta}_j\|_2)=\mathrm{sgn}_0(\|\beta_{0j}\|_2)$, $1\le j\le d$, where $\mathrm{sgn}_0(u)=1$ if $|u|>0$, and $=0$ if $u=0$.

Theorem 2. Suppose that (A1)–(A5) hold and the matrix Σ_{A^o} *is positive definite. Then* $P(sgn_0(\hat{\beta}) = sgn_0(\beta_0)) \to 1$.

Theorem 2 establishes that the estimate from one-step iteration can consistently distinguish covariates with zero and nonzero regression coefficients. Following similar arguments, we can prove that the iterative estimate is selection consistent. With the selection consistency and finite q, estimation consistency can be easily obtained.

Proof. This theorem can be proved analogously to the proof of the selection consistency of the adaptive group Lasso in Wei and Huang (2010) and the adaptive Lasso in Huang et al. (2009). The key is to note that (a) the initial estimate is estimation consistent, and (b) the dimensionality of the initial estimate is controlled by the sample size and a multiply of the number of true positives.

References

Huang J, Ma S, Zhang CH. Adaptive Lasso for high-dimensional regression models. Statistica Sinica. 2008; 18:1603–1618.

Wei F, Huang J. Consistent group selection in high-dimensional linear regression. Bernoulli. 2010; 16:1369–1384. [PubMed: 22072891]

Zhang CH, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. Annals of Statistics. 2008; 36:1567–1594.

Table 1

Simulation: summary based on 200 replicates. There are 4 independent studies with 50 (100) samples per study. Correlation structures include auto-regressive (auto), banded (band) and compound symmetry (comp). ρ: correlation coefficient. Number of true positives: 20. P: number of covariates identified. TP: number of true positives identified.

				meta-analysis						intensity approach						integrative analysis					
				Lasso		one-step		bridge		Lasso		one-step		bridge		GLasso		one-step		Proposed	
sample	#cov	cor	ρ	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP
50	1000	auto	0.3	158	20	148	20	114	19	148	20	36	20	22	20	119	17	34	17	20	17
			0.7	163	20	139	20	102	20	88	20	30	20	21	20	99	20	23	18	20	18
		band	0.2	164	20	147	20	98	20	75	20	30	20	22	20	82	20	21	17	21	17
			0.33	165	20	150	20	87	19	110	20	37	20	23	20	122	19	28	18	19	18
		comp	0.3	139	20	120	20	84	19	127	18	56	17	32	15	107	18	32	18	18	17
			0.7	120	20	91	20	55	20	104	20	42	17	25	14	108	19	24	18	17	17
100	2000	auto	0.3	181	20	134	20	60	20	95	20	32	20	22	20	151	20	21	20	20	20
			0.7	127	20	86	20	38	20	156	20	84	20	25	20	51	20	20	20	20	20
		band	0.2	187	20	82	20	28	20	165	20	88	20	43	20	32	20	20	20	20	20
			0.33	196	20	83	20	39	20	167	20	80	20	39	20	92	20	20	20	20	20
		comp	0.3	101	20	124	20	51	20	143	20	78	20	21	20	117	20	27	20	20	20
			0.7	152	20	68	20	25	20	129	20	54	19	29	17	106	20	23	20	20	20

Table 2

Breast cancer prognosis studies.

Reference	Platform	Gene	Sample
Huang et al. (2003)	Affymetrix	12625	71
Sorlie et al. (2001)	cDNA	8102	58
Sotiriou et al. (2003)	cDNA	7650	98
van't Veer et al. (2002)	Oligonucleotide	24481	78

Table 3

Breast cancer markers identified and corresponding estimates.

Gene	D1	D2	D3	D4
Protoporphyrinogen oxidase (PPOX)	0.0284	0.0032	0.0436	0.0069
Myeloid/lymphoid or mixed-lineage leukemia 4 (MLLT4)	0.0006	0.0009	0.0013	0.0005
Meis homeobox 2 (MEIS2)	0.0002	−0.0225	−0.0724	0.0017
Myoglobin (MB)	0.0007	−0.0001	−0.0002	0.0002
Carnitine O-acetyltransferase (CRAT)	0.0023	0.0014	−0.0055	−0.0002
Tax1 (human T-cell leukemia virus type I) binding protein 3 (TAX1BP3)	−0.0052	−0.0006	−0.0043	−0.0017
Rearranged L-myc fusion (RLF)	−0.0039	−0.0013	−0.0006	−0.0007
Tyrosine kinase, non-receptor, 2 (TNK2)	0.0011	0.0005	0.0004	0.0010
Ecto-NOX disulfide-thiol exchanger 2 (ENOX2)	−0.0182	−0.0039	−0.0039	−0.0013
Complement component 3a receptor 1 (C3AR1)	0.0188	0.0093	0.0237	0.0042
Transportin 1 (TNPO1)	0.0196	−0.0014	−0.0042	0.0011
Transcribed locus	−0.0110	−0.0116	−0.0089	−0.0061
Lysine (K)-specific demethylase 1A (KDM1A)	−0.0071	−0.0036	0.0015	0.0003
Transcribed locus	−0.0138	−0.0096	0.0274	0.0085
Oxysterol binding protein (OSBP)	−0.0052	−0.0040	−0.0003	−0.0010
Transcribed locus	−0.0036	−0.0057	−0.0010	−0.0052
Fibroblast growth factor 2 (basic) (FGF2)	0.0043	0.0004	0.0005	0.0005
Gelsolin (GSN)	−0.0021	−0.0005	−0.0008	−0.0002
Growth factor receptor-bound protein 2 (GRB2)	0.0005	0.0002	0.0001	0.0001
Phosphatidylinositol glycan anchor biosynthesis, class C (PIGC)	−0.0022	−0.0011	−0.0017	−0.0010
Annexin A1 (ANXA1)	−0.0093	−0.0020	−0.0061	−0.0001
Interleukin 6 (interferon, beta 2) (IL6)	−0.0029	−0.0024	−0.0011	−0.0011

Table 4

Data analysis results using different approaches. With meta-analysis approaches, numbers in "()" are the number of genes identified with each individual datasets. A logrank statistic 3.84 corresponds to p-value 0.05.

Approach		Gene	Overlap	Logrank
Meta-analysis	Lasso	81 (25, 20, 24, 13)	5	2.661
	one-step	84 (26, 21, 29, 15)	8	1.481
	bridge	68 (25, 21, 24, 13)	10	1.391
Intensity approach	Lasso	32	2	1.884
	one-step	22	1	1.799
	bridge	33	6	1.523
Integrative analysis	GLasso	42	4	2.100
	one-step	21	5	3.910
	proposed	22	–	5.930