Published in final edited form as:
Stat Probab Lett. 2010 September 7; 80(17-18): 1313–1319. doi:10.1016/j.spl.2010.04.011.

An Integrative Pathway-based Clinical-genomic Model for Cancer Survival Prediction

Xi Chen, PhD[1,*], **Lily Wang, PhD**[2], and **Hemant Ishwaran, PhD**[3]

[1]Division of Cancer Biostatistics, Department of Biostatistics, Vanderbilt University, Nashville, TN 37232, USA

[2]Department of Biostatistics, Vanderbilt University, Nashville, TN 37232, USA

[3]Department of Quantitative Health Sciences, The Cleveland Clinic, 9500 Euclid Ave. Cleveland, OH 44195, USA

Abstract

Prediction models that use gene expression levels are now being proposed for personalized treatment of cancer, but building accurate models that are easy to interpret remains a challenge. In this paper, we describe an integrative clinical-genomic approach that combines both genomic pathway and clinical information. First, we summarize information from genes in each pathway using Supervised Principal Components (SPCA) to obtain pathway-based genomic predictors. Next, we build a prediction model based on clinical variables and pathway-based genomic predictors using Random Survival Forests (RSF). Our rationale for this two-stage procedure is that the underlying disease process may be influenced by environmental exposure (measured by clinical variables) and perturbations in different pathways (measured by pathway-based genomic variables), as well as their interactions. Using two cancer microarray datasets, we show that the pathway-based clinical-genomic model outperforms gene-based clinical-genomic models, with improved prediction accuracy and interpretability.

Keywords

microarrays; gene expression; pathway analysis; survival prediction; random survival forests

1. INTRODUCTION

Cancer is a heterogeneous complex disease, influenced by both genetic background and environmental exposure. Therefore, when weighing treatment options it is important to use accurate prognostic models. Towards this end, it is common for clinical information such as age, tumor size, histopathologic grade, and lymph node involvement to be used when modeling cancer prognosis. Tumor specific markers are also increasingly used in modeling. For example, estrogen receptor and human epidermal growth factor receptor 2 (HER2) are used for breast cancer, and prostate specific antigen (PSA) for prostate cancer.

[*]Corresponding Author: Xi Chen, PhD, Division of Cancer Biostatistics, Department of Biostatistics, Vanderbilt-Ingram Cancer Center, 2220 Pierce Avenue, 571 Preston Building, Nashville, TN 37232-6848, Phone: 615-936-2785, Fax: 615-936-2602, steven.chen@vanderbilt.edu.

However, while clinical prognostic factors are useful at a population level for identifying risk, because of the heterogeneity and complexity of cancer, they are far from being accurate predictors of an individual's clinical course. Often, cancer patients may exhibit the same clinical pattern, but prognosis can vary significantly. The recent development of high-throughout microarrays for measuring gene expression has given investigators a more accurate way of identifying cancer subtypes by making use of gene expression profiles. This approach has been successfully applied to tumor classification and to making prognosis in lung, breast, colon and other cancers (Beer et al., 2002; Perou et al., 2000, Alon et al., 1999). For predicting survival outcome (death or recurrence) approaches based on the Cox proportional hazard model have been proposed, including partial least square (Nguyen and Roche, 2002), L1-penalized regression (Segal, 2006, Datta et al., 2007), L2-penalized regression (Hastie and Tibshirani, 2004), supervised principal component analysis (Bair and Tibshirani, 2004). By using high dimensional genomic information, these methods have been shown to improve cancer prognosis compared to models based on clinical predictors alone.

Although gene-based prediction models is a promising approach, some disturbing inconsistencies among gene expression profiles have been reported (Ein-Dor et al., 2005). It has been suggested that one reason for these discrepancies may be that in complex diseases many genes are associated with outcomes, each with only a small marginal effect. Therefore, many real but weak signals could be missed, especially when sample sizes are small (Mootha et al., 2003). Because activities within pathways are key components for cancer development (Wood et al., 2007), pathway analyses that borrow information from genes within a pathway and associate groups of genes instead of individual genes with clinical outcome, have become a more popular alternative (Mootha et al., 2003; Wang et al. 2008). Increased power comes from combining weak signals from a number of genes within each pathway. Signatures from oncogenic pathways have been shown to be not only effective markers for identifying tumor subtypes, but also a valuable guide for targeting therapies (Bild et al., 2006). Several recent papers have successfully integrated microarray data with prior pathway knowledge for disease status prediction (Lee et al., 2008; Chen and Wang, 2009) and have shown that pathway based prediction models improve accuracy and increase reproducibility (Manoli et al., 2006).

In this paper, we propose an integrative clinical-genomic model that combines both genomic pathway *and* clinical information. More specifically, we integrate genomic predictors based on pathway information with clinical variables using Supervised Principal Components (SPCA) and Random Survival Forests (RSF) (Ishwaran et al., 2008). First, we summarize genomic information from each pathway using supervised principal components (Section 2.1); these are the "supergenes". Next, using both supergenes and clinical variables as predictors, we use RSF for prediction. Doing so allows us to approximate the underlying functional gene network by allowing interactions between pathway-pathway and pathway-clinical (or environmental) factors. One of the important properties of RSF is that it is highly adept at identifying interactions.

Section 2 describes the details of our proposed approach. Section 3 considers two publicly available breast cancer microarray data sets and compare the performance of our proposed model with three widely used approaches: supervised principal components, L1-penalized Cox regression, and Cox-based boosting.

2. METHOD

Our method involves two main steps:

1. Summarize information from genes in each pathway using SPCA to obtain pathway-based genomic predictors.
2. Build prediction model based on clinical variables and pathway-based genomic predictors using RSF.

Our rationale for this approach is that the underlying disease process may be influenced by environmental exposure (measured by clinical variables), perturbations in different pathways (measured by pathway-based genomic variables) as well as their interactions. We discuss the SPCA method and RSF in more details in the next two sections.

2.1 SPCA

Principal Component Analysis (PCA) is a popular dimension reduction technique for summarizing information from a group of genes, such as those from the same pathway (Tomfohr et al., 2005; Bild et al., 2006). The first principal component has been called the "eigengene" or "metagene" (Alter et al., 2000). However, because pathways such as those from Gene Ontology (Ashburner et al., 2001) are defined *a priori* in a particular experiment, typically only a subset of genes from a pathway work together to influence changes in a biological process, which then brings about changes in outcome. When all the genes in a pathway are used to estimate the principal components, the resulting "eigengene" may be affected by noisy signals from genes unrelated to outcome. Therefore, we adopt a semi-supervised approach: SPCA to do gene shaving and dimension reduction within pathways.

SPCA was proposed by Bair and Tibshirani (2004) and Bair et al. (2006) to predict survival outcomes using genome-wide gene expression data. Instead of using all genes as in standard PCA, a subset of genes having strong correlations with survival time are selected for PCA. The estimated principal component scores are used as predictors in a Cox proportional hazard model for survival prediction.

Although SPCA is an effective prediction model, it can be difficult to interpret because of the large number of genes often selected in constructing the eigengenes. Rather than assuming a few eigengenes at the genome-wide level, it is more reasonable to assume an eigengene for each pathway. Towards this end, SPCA was successfully modified and applied to pathway analysis (Chen at al., 2008; Chen and Wang, 2009). Along the same lines, in this paper, we use the estimated eigengene (i.e., first principal component score) to represent the latent variable associated with underlying biological process in a pathway.

2.2 Random Survival Forests

Random forests (RF) is a state of the art ensemble learning method, which was introduced by Breiman (2001), and further developed by Breiman and Cutler. RF grows deep, random trees, which are aggregated to form the ensemble learner. By growing a deep tree, the base learner has low bias. By growing the tree randomly (see details below), tree-correlation, and hence variance is kept low. These two competing forces enable RF to be an effective classification and regression procedure for high-dimensional data.

Random survival forests (RSF) is a new extension of Breiman's RF methodology to right-censored survival settings (Ishwaran et al., 2008). The core algorithm used by RSF and RF are similar. First, *ntree* bootstrap samples are drawn from the original data. For each bootstrap sample, a single random survival tree is grown. In growing the tree, at each tree node, *mtry* variables are randomly selected and the node is split by finding the variable that maximizes the log-rank test across *nsplit* randomly selected split points. Each survival tree is grown to full size under the constraint that the minimum number of unique event times in a node is no less than a predefined *nodesize value*. Each bootstrap sample excludes on average

36.7% of the data, called out-of-bag (OOB) data, which is used to construct an OOB ensemble used for estimating test set error (Ishwaran et al., 2008).

To evaluate the survival prediction performance of RSF as well as other procedures, we use Harrell's concordance index (C-index) (Harrell et al., 1982), which estimates the probability of concordance between predicted and observed survival. To measure the error rate, we use 1-*C*, which is bounded between 0 and 1. An error rate of 0 indicates perfect prediction, whereas an error rate of 0.5 indicates random guessing.

A key feature of RSF is the ability to assess variable importance (VIMP). VIMP is defined as the prediction error from the OOB ensemble subtracted from the prediction error of a new OOB ensemble derived when the variable in question is "noised up" (Ishwaran et al., 2008). A large positive VIMP indicates a predictive variable.

2.3 Analysis Details: Work-Flow

The following algorithm describes the work-flow used to construct the clinical-genomic model:

1. Randomly partition the data into training and test sets of sizes n_1 and n_2, respectively.
2. Link gene identifiers from microarrays with those from the pathway databases Gene Ontology and KEGG (Ashburner et al., 2000; Kanehisa et al., 2002). Group genes into different pathways. For genes not assigned to any gene categories, rather than discarding those genes, perform *K*-means clustering to group them based on their expression patterns. To determine the optimal number of cluster, use the Gap statistic (Tibshirani et al., 2001). Genes grouped this way are referred to as "pseudo-pathways".
3. Using the training set, use SPCA to select the subset of genes most associated with survival outcome for each pathway (or pseudo-pathway). Using the gene expressions of the selected genes, the "supergene" for a pathway is estimated by the first principal component score. A super-gene expression matrix is constructed using the supergenes. If there is a total of m pathways (including pseudo-pathways), the training set super-gene matrix is of dimension $m \times n_1$.
4. Using survival outcomes as the response, use RSF with the pathway super-gene matrix and clinical information as predictors. Only training data is used.
5. To assess performance, construct a super-gene matrix for the test data (of dimension $m \times n_2$). These calculations use the eigenvectors estimated from the training data alone. Using the training set derived forest, determine the accuracy of the resulting predictor on the test set using the test set super-gene matrix and test set clinical variables.

3. RESULTS

We studied the performance of our method using two breast cancer microarray datasets. Our first example is the widely used benchmark microarray dataset from Miller et al. (2005). It included 251 microarray samples (i.e., patients) obtained from Affymetrix U133A and U133B platforms (GEO accession no. GSE3494). Of the 251 samples, only 236 have follow-up information; only these data were used for our analysis. In addition to gene expression data, clinical predictors used included: P53 status, Elston-grade, ER, PgR, age, tumor size and lymph node status. The second dataset included 255 early stage estrogen receptor (ER) positive breast cancer samples from patients receiving tamoxifen adjuvant

treatment (Loi et al., 2008). Three Affymetrix platforms, U133A, U133B and U133PLUS2 were used (GEO accession no. GSE6532). The survival endpoint was time until first distant metastatic event (distant metastasis free survival). Clinical-pathological predictors were histological grade, tumor size, age, nodal status, ER (high vs. low expression), PgR (high vs. low expression) and HER2 (high vs. low expression).

3.1 Survival Prediction Performance

Each dataset was randomly split into training and testing sets using a 2:1 ratio. For the Miller dataset, we mapped 13,441 genes to Gene Ontology "Biological Process" (GO-BP) categories. There were 10,695 genes belonging to 1570 GO-BP categories with gene set sizes larger than two. The remaining 2,746 genes were split into groups with similar gene expression patterns using *K*-means clustering in tandem with the Gap Statistic (Tibshirani et al, 2001). There were a total of 1576 pathways (1570 based on GO categories and 6 based on *K*-means clustering). These were then used to derive 1576 supergenes. For the Loi dataset, we mapped 11,553 genes to 1656 GO-BP categories, and the remaining 4184 genes were divided into 12 clusters, yielding a total of 1668 supergenes.

RSF was applied as described in Section 2.3. All forests were comprised of *ntree* = 5000 survival trees, with each tree grown under random log-rank splitting with an *nsplit* value of 10. All RSF applications in this paper were implemented using the R-package, *randomSurvialForest* (Ishwaran and Kogalur, 2007). Default values for the package were used in all examples, excepting those just listed.

We compared performance with three other popular procedures: (i) SPCA (Bair et al., 2006); (ii) L_1-penalized Cox regression (LASSO) (Park and Hastie, 2007); and (iii) Cox-likelihood based boosting (Binder and Schumacker, 2008). These procedures were implemented using the R-packages: superpc, glmpath and Coxboost, respectively. Table 1 lists test set errors for all procedures using 1-C (Section 2.2). All results are averaged over 10 independent replicates of the procedure outlined above. Pathway-based models used supergenes and clinical variables as predictors, whereas gene-based models used individual genes and clinical variables as predictors. Five-fold cross-validation was used to select tuning parameters for the comparison procedures (for SPCA, the threshold for selecting genes; for Cox Lasso, the L_1 regularization parameter; and for Cox Boosting, the optimal number of boosting steps). For all procedures, as expected, the pathway-based approach performed better than the gene-based one. This is not surprising, as these models have incorporated additional prior biological knowledge. Among all methods, the RSF pathway-model had lowest prediction error over both datasets.

To better understand why the RSF pathway based approach worked so well, we investigated three sources potentially contributing to its success: (1) selection of subset of genes for constructing supergenes within each gene set; (2) gene categories from clustering of genes not annotated in pathway databases; and (3) clinical variables. For each of these factors, the RSF based model was constructed by removing the factor being investigated and keeping all other procedures unchanged. For example, to evaluate the impact of (1), all genes instead of a selected subset of genes were used to estimate the supergenes for each pathway. For (2) and (3), supergenes from *K*-means clusters and clinical variables were omitted from the RSF model, respectively.

Table 2 shows that the selection of a subset of genes for estimating supergenes had the largest impact on prediction error. This supports our assumption that within each gene set, only a subset of genes play an important role. By removing noisy signals from non-relevant genes, the pathway-based RSF model improves both prediction performance and biological interpretation.

Finally, we should remark that while we used K-means and the Gap statistic to form clusters for genes without pathway annotation in our approach, this may not necessarily be the optimal method for clustering. In deciding what method might be best, one could use the R-packages “clValid” and “RankAggreg” to evaluate the performances of different clustering algorithms and to select the optimal approach (Datta and Datta, 2003). However, we note that informal experimentation with different clustering procedures showed that prediction performance for RSF remained very stable. We believe this is because most of the predictors in the model were derived based on a priori defined pathway information, which lends stability to our approach.

3.2 Predictors Identified by RSF

A VIMP analysis was used to identify key variables for predicting survival outcome. All predictors, including supergenes representing the pathways and clinical variables, were ranked by VIMP. To increase precision, bootstrap resampling was used. We drew 200 independent bootstrap samples and calculated VIMP for each sample. The estimated VIMP for a variable was calculated as the mean over all bootstrap samples divided by the standard deviation.

Table 3 shows the top 10 variables in terms of standardized VIMP for the Miller et al. (2005) dataset. The list includes one clinical variable and nine supergenes. The pathways for the supergenes are involved in different biological processes such as cell proliferation, neuron development, cell cycle, ion transport and amino acid metabolism. The most significant predictor is endothelial cell proliferation (GO: 0001935). It is well known that tumor angiogenesis, which is the development of new blood vessels and a critical process in tumor progression, is dependent on endothelial cell proliferation. It has been reported that estrogen directly modulates angiogenesis through endothelial cells and estrogen receptor antagonists can inhibit angiogenesis in breast tumors (Glgliardi and Collins, 1993). In addition, two GO categories (GO: 0000281, GO: 0000077) are related to cell cycle, which is closely related to cancer, which results from uncontrolled division and growth of cells.

The nine supergenes listed in Table 3 were derived from a total of 39 genes (last column). Prior literature has shown several of these to be directly related to breast cancer. For example, PTEN is a tumor suppressor gene working through the action of its phosphatase protein product. Inactivating mutations or deletions of the PTEN gene can lead to resistance to chemotherapy and hormone therapy (Pandolfi, 2004).

Supplementary Table 1 shows the 10 predictors by standardized VIMP for the Loi et al. (2008) data. The first pathway, negative regulation of apoptosis (GO: 0043066), is substantially more predictive than other genomic or clinical predictors. It has been shown that Tamoxifen (TAM) and its active metabolite, 4-hydroxytamoxifen (OHT) can induce apoptotic cell death through ER-dependent and ER-independent pathways (Mandlekar et al., 2000; Obrero et al., 2002). Different studies have confirmed that multiple non-ER-mediated mechanisms such as MAP kinases, calmodulin and calcium signaling, caspases, TGF-beta involve TAM-induced apoptosis (Mandlekar and Kong, 2001). The genes in negative regulation of apoptosis and several other top pathways are closely related to these functions. For example, VEGFA is a pivotal gene in breast tumor angiogenesis and metastases and elevated VEGFA level is known to be associated with reduced disease-free survival for TAM treated patients (Ryden et al., 2005).

3.3 Interactions Identified by RSF

In addition to ranking and identifying important predictors individually, RSF can also be used to identify important interactions between variables. For a pair of variables, the joint

VIMP is defined to be the difference between the prediction error when both predictors are noised up and the prediction error without noising up. The VIMP for each single variable is calculated and the sum of two single variable VIMPs is the additive importance. A large difference between the joint VIMP and the additive importance indicates a potential interaction between two variables (Ishwaran, 2007).

We used the top 30 predictors with largest single variable VIMPs and evaluated all pairwise interactions of these variables using the above approach. To increase precision we used a bootstrap standardized measure for the interactions using 200 independent bootstrap draws.

Supplementary Table 2 shows the top 10 pairwise interactions for each dataset. For the Miller et al. (2005) data, we identified both pathway-pathway and pathway-clinical interactions. An interesting finding relates to the prostaglandin metabolic process (GO: 0006693). Although the individual VIMP ranking for this pathway is 28th, its pairwise interaction with several other pathways ranked high. Prostaglandin E2 and its receptors play a key role in cancer progression by activating signaling pathways that involve apoptosis, antiogenesis, migration and cell proliferation (Wang and DuBois, 2006). Several clinical trials have shown that non-steroidal anti-inflammatory drugs (NSAIDs), which inhibits prostaglandins mediated processes, can reduce the relative risk of developing different cancers such as breast, colorectal, bladder, etc. (Gupta and DuBois, 2001). Our results suggest that the genes related to the prostaglandin metabolic process exhibit high connectivity with genes in other pathways and are possible “hub” genes in breast cancer development.

For the Loi et al. (2008) data, a pathway involved in several important interactions is the very-long-chain fatty acid metabolic process (GO: 0000038), which ranked 25th by individual VIMP. This is probably because tamoxifen affects fatty acid metabolism. Actually, tamoxifen therapy is associated with an increased risk of developing fatty liver (steatosis) and it is reported that 43% of patients having tamoxifen treatment may develop steatosis within the first two years (Ogawa et al., 1998). The selected genes in this gene set include ELOVL2, HSD17B4, SLC27A2 and SLC27A6, and these genes are critical for triglyceride biosynthesis. This perturbed pathway may have effect on other metabolism processes in these tamoxifen treated breast cancer patients.

4. DISCUSSION

We have presented a novel approach to predicting survival outcomes by integrating gene expression profiles with prior biological knowledge and clinical factors. Because the underlying disease process for cancer may be dependent on perturbations of different pathways, prediction models based on pathways may approximate the true disease process more closely than models based on genes alone. We have shown that our pathway-based clinical-genomic model improves prediction accuracy over gene-based prediction models. Furthermore, we found in addition to grouping genes into pathways, within each pathway, the selection of the subset of genes most associated with the outcome is a critical step for accurate prediction performance. This agrees with our hypothesis that only a subset of genes from a pre-defined pathway may participate in the cellular process influencing survival outcome.

An attractive feature of our methodology is that RSF can handle a large number of clinical and genomic predictors with mixed types (categorical or continuous). In addition, RSF can automatically discover higher order and nonlinear interactions between predictors such as clinical-clinical, clinical-pathway and pathway-pathway interactions. This important feature enables us to closely approximate the underlying disease process, which is influenced by

multiple pathways, environmental effects, and pathway-environmental interactions. This in turn can shed light on the biological mechanisms behind a disease process. Although we have described a pathway-based clinical-genomic modeling for survival outcomes, the methodology is generalizable and can be easily extended to binary, multi-category, or continuous outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work of XC was partially supported by NCI grant 5P30CA068485-13. The work of LW was partially supported by NICHD grant 5P30 HD015052-25 and NIH grant 1 P50 MH078028-01A1.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA. 1999; 96:6745–6750. [PubMed: 10359783]

Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA. 2000; 97:10101–10106. [PubMed: 10963673]

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 2000; 25:25–29. [PubMed: 10802651]

Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Bio. 2004; 2:511–522.

Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. J. Amer. Stat. Assoc. 2006; 101:119–137.

Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med. 2002; 8:816–824. [PubMed: 12118244]

Bild AH, Potti A, Nevins JR. Linking oncogenic pathways with therapeutic opportunities. Nat. Rev. Cancer. 2006; 6:735–741. [PubMed: 16915294]

Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics. 2008; 9:14. [PubMed: 18186927]

Breiman L. Random forests. Machine Learning. 2001; 45:5–32.

Chen X, Wang L, Smith JD, Zhang B. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. Bioinformatics. 2008; 24:2474–2481. [PubMed: 18753155]

Chen X, Wang L. Integrating biological knowledge with gene expression profiles for survival prediction of cancer. J. Comput. Biol. 2009; 16:265–278. [PubMed: 19183004]

Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics. 2003; 19:459–466. [PubMed: 12611800]

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]

Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005; 21:171–178. [PubMed: 15308542]

Gagliardi A, Collins DC. Inhibition of angiogenesis by antiestrogens. Cancer Res. 1993; 53:533–535. [PubMed: 7678775]

Gupta RA, Dubois RN. Colorectal cancer prevention and treatment by inhibition of cyclooxygenase-2. Nat. Rev. Cancer. 2001; 1:11–21. [PubMed: 11900248]

Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. J. Am. Med. Assoc. 1982; 247:2543–2546.

Hastie T, Tibshirani R. Efficient quadratic regularization for expression arrays. Biostatistics. 2004; 5:329–340. [PubMed: 15208198]

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. Ann. Appl. Stat. 2008; 2:841–860.

Ishwaran H, Kogalur UB. Random survival forests for R. Rnews. 2007; 7/2:25–31.

Ishwaran H. Variable importance in binary regression trees and forests. Elect. J. Stat. 2007; 1:519–537.

Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002; 30:42–46. [PubMed: 11752249]

Loi S, Haibe-Kains B, Desmedt C, Wirapati P, Lallemand F, Tutt AM, Gillet C, Ellis P, Ryder K, et al. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. BMC Genomics. 2008; 9:239. [PubMed: 18498629]

Mandlekar S, Hebbar V, Christov K, Kong ANT. Pharmacodynamics of tamoxifen and its 4-hydroxy and N-desmethyl metabolites: Activation of caspases and induction of apoptosis in rat mammary tumors and in human breast cancer cell lines. Cancer Res. 2000; 60:6601–6606. [PubMed: 11118041]

Mandlekar S, Kong ANT. Mechanisms of tamoxifen-induced apoptosis. Apoptosis. 2001; 6:469–477. [PubMed: 11595837]

Manoli T, Gretz N, Grone HJ, Kenzelmann M, Eils R, Brors B. Group testing for pathway analysis improves comparability of different microarray datasets. Bioinformatics. 2006; 22:2500–2506. [PubMed: 16895928]

Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc. Natl. Acad. Sci. USA. 2005; 102:13550–13555. [PubMed: 16141321]

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, et al. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 2003; 34:267–273. [PubMed: 12808457]

Nguyen DV, Rocke DM. Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics. 2002; 18:1625–1632. [PubMed: 12490447]

Obrero M, Yu DV, Shapiro DJ. Estrogen receptor-dependent and estrogen receptor-independent pathways for tamoxifen and 4-hydroxytamoxifen-induced programmed cell death. J. Biol. Chem. 2002; 277:45695–45703. [PubMed: 12244117]

Ogawa Y, Murata Y, Nishioka A, Inomata T, Yoshida S. Tamoxifen-induced fatty liver in patients with breast cancer. Lancet. 1998; 351:725–725. [PubMed: 9504521]

Pandolfi PP. Breast cancer--loss of PTEN predicts resistance to treatment. N. Engl. J. Med. 2004; 351:2337–2338. [PubMed: 15564551]

Park MY, Hastie T. L-1-regularization path algorithm for generalized linear models. J. Roy. Stat. Soc. B. 2007; 69:659–677.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, et al. Molecular portraits of human breast tumours. Nature. 2000; 406:747–752. [PubMed: 10963602]

Ryden L, Stendahl M, Jonsson H, Emdin S, Bengtsson N, Landberg G. Tumor-specific VEGF-A and VEGFR2 in postmenopausal breast cancer patients with long-term follow-up. Implication of a link between VEGF pathway and tamoxifen response. Breast Cancer Res. TR. 2005; 89:135–143.

Segal MR. Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. Biostatistics. 2006; 7:268–285. [PubMed: 16284340]

Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J. Roy. Stat. Soc. B. 2001; 63:411–423.

Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics. 2005; 6:225. [PubMed: 16156896]

Wang D, Dubois RN. Prostaglandins and cancer. Gut. 2006; 55:115–122. [PubMed: 16118353]

Wang L, Zhang B, Wolfinger RD, Chen X. An Integrated Approach for the Analysis of Biological Pathways using Mixed Models. PLoS Genet. 2008; 4:e1000115. [PubMed: 18852846]

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007; 318:1108–1113. [PubMed: 17932254]

Table 1

Test set errors for RSF, SPCA, LASSO and CoxBoost using 1-C. Values reported are averaged over 10 independent experiments.

	Method	RSF	SPCA	LASSO	CoxBoost
Miller's data	Gene-based	0.3066	0.3029	0.3900	0.3634
	Pathway-based	0.2823	0.2857	0.3118	0.3140
Loi's data	Gene-based	0.3194	0.3286	0.3473	0.3464
	Pathway-based	0.2881	0.3167	0.3241	0.3279

Stat Probab Lett. Author manuscript; available in PMC 2011 September 7.

Table 2

Factors contributing to performance of pathway-based RSF prediction models. Shown are test set errors using 1-C, averaged over 10 independent experiments.

Method	Miller et al. (2005) data	Loi et al. (2008) data
Pathway-based RSF	0.2823	0.2881
(1) No gene screening in gene sets	0.2992	0.3014
(2) No K-means clusters	0.2903	0.2934
(3) No clinical variables	0.2895	0.2925

Table 3

The top 10 predictors with largest standardized VIMP for the Miller et al. (2005) dataset

Name	Description	Set size	VIMP	Gene Symbols of selected genes used to estimate supergenes
GO:0001935	Endothelial cell proliferation	4	1.921	DLG1, HMOX1
GO:0031175	Neuron projection development	13	1.459	LAMB1, CDK5, CHL1, CDK5R1, STX3, EFHD1, NRTN, GALR2, PTEN, RASGRF1, STMN3, GDNF
GO:0000059	Protein import into nucleus, docking	16	1.347	CSE1L, RANBP5, IPO4
GO:0000281	Cytokinesis after mitosis	3	1.344	MYH10, NUSAP1
GO:0006537	Glutamate biosynthetic process	3	1.202	PRODH, LOC440792
Lymphnode			1.201	
GO:0009168	Purine ribonucleoside monophosphate biosynthetic process	5	1.160	AMPD1, AMPD3, CECR1
GO:0000077	DNA damage checkpoint	11	1.160	RAD1, RAD9A, FOXN3, CHEK1, RAD17, ATR, CHEK2, HUS1, ZAK, BRIP1
GO:0015711	Organic anion transport	14	1.139	SLC16A1, SLC16A3, SLC16A5
GO:0018206	Peptidyl-methionine modification	3	1.119	METAP1, PDF