Index Models for Pathway Based Personalized Prognostic Gene Expression Signatures

University of Wisconsin-Madison
Department of Biostatistics and Medical Informatics
Technical Report #216

Kevin H. Eng and Sijian Wang and Christina Kendziorski

Department of Biostatistics and Medical Informatics
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue
Madison, Wisconsin 53706
USA
e-mail: eng@stat.wisc.edu
e-mail: wangs@stat.wisc.edu
e-mail: kendzior@biostat.wisc.edu

1. Introduction

A main challenge of personalized medicine is to identify the individuating features of a patient and specify the way in which they can be used to determine an optimal treatment. Molecular profiling holds major promise to this end. Indeed, NIH Director Francis Collins suggests that the development of targeted therapeutics based on a molecular understanding of disease may be the "most profound consequence of the genome revolution" (Collins, 2010). Unfortunately, the information available in high-throughput genetic screens has yet to be fully utilized in routine patient care.

To date, the development of molecular based signatures to guide treatment is certainly notable with a number of high-throughput expression based assays now in clinical use for predicting recurrence of breast (Mook *et al.*, 2007; Sparano and Paik, 2008), colon, and prostate cancer, as well as transplant rejection (Clark-Langone *et al.*, 2010; Craver, 2010). Although invaluable in individual cases, the information provided by these signatures is limited. In particular, a patient is classified into one of a very few (two or three) groups, limiting the potential for personalized treatment. Furthermore, although current signatures provide guidance on whether or not to treat (e.g if recurrence is or is not predicted), they provide no information on *how* to treat.

We propose a simple statistical approach to address these limitations. The approach provides functionally relevant signatures that enable refined patient stratification and allow for the identification of patient specific aberrant pathways which may be useful in guiding treatment. In short, the proposed hierarchical framework models a survival related phenotype as attributable to known genes within known biological pathways. Given genes identified as conferring increased or decreased risk within a pathway, a pathway summary, or "index", is then constructed. The indices are used in a second model to identify important pathways. As we show in Section 3, the patient-specific index scores across important pathways (referred to as patient-specific risk profiles) are powerful and efficient characterizations useful in addressing a number of questions related to predicting survival and optimizing treatment.

Recently the variable selection problem with grouped predictors has been considered by several authors. Yuan and Lin (2006) and Zhao, Rocha and Yu (2009) introduced methods that penalize the L_2-norm and L_∞-norm, respectively, of coefficients within each group in a linear regression. Ma, Song and Huang (2007) applied the group lasso penalty to the Cox proportional hazards model. Based on the boosting technique, Luan and Li (2008) and Wei and Li (2007), respectively, developed a group additive regression model and a nonparametric pathway-based regression model to identify groups of genomic features that are related to several clinical phenotypes including the survival outcome. All these group variable selection methods have a common limitation: they select variables in an "all-in-all-out" fashion. In other words, when one variable in a group is selected, all other variables in the same group are also selected. Thus, these methods only conduct "group selection" but do not select important variables within the identified group. The reality, however, may be that some genes in a pathway are not related to the phenotype although the pathway as a whole is involved in the biological process.

In order to achieve sparsity within groups, Huang *et al.* (2009) imposed a bridge L_γ-norm penalty on coefficients within each group in linear regression. Wang *et al.* (2009) and Zhou and Zhu (2010) independently investigated the same criterion with $\gamma = 0.5$ (when groups overlap) for linear and Cox regression, respectively. A drawback of these methods is that the objective functions are no longer convex, which causes numerical problems in practice, especially when the number of predictors is large (e.g. > 500).

Details of our approach are given in the next section. Most of our analysis is focused on data from The Cancer Genome Atlas (TCGA) Project. Information on TCGA can be found at (`http://cancergenome.nih.gov`). Briefly, TCGA is a comprehensive and coordinated effort designed to improve the understanding and treatment of cancer through an increased understanding of the molecular basis of the disease gained by analysis and integration of multiple large-scale measurements of the phenome and genome collected on relatively large groups of patients. TCGA initially focused on three of the most disproportionately deadly cancers - lung, brain, and ovary. Considering the ovarian cancer data, Section 3 demonstrates that the proposed approach produces useful multi-category risk prediction, that are validated in independent datasets, and also provides functionally relevant information that may significantly impact ovarian cancer research as well as patient treatment.

2. Pathway Index Model

We observe survival information $(Y_i, \delta_i, x_{i1}, \dots, x_{ip})$, for $i = 1, \dots, n$ patients, where Y_i is a complete survival time when $\delta_i = 1$ and is a right censored time when $\delta_i = 0$. An individual's covariate vector $(x_{i1}, \dots, x_{ip})$ contains gene-level measurements where the number of genes, p, may be of ultra-high order. We will denote a single gene's measurements across individuals by the $(n \times 1)$ column vector X_j.

Pathway information. Through outside database information, e.g., KEGG (Kanehisa *et al.*, 2010) or the Gene Ontology, we define a set of pathways $\mathcal{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_m\}$ each of which is a set of genes:

$$\mathcal{P}_k = \{j : j\text{th gene is a member of the }k\text{th pathway}\} \tag{1}$$

It is expected that some genes are members of multiple pathways. Pathways are expected to contain order n or fewer genes which reduces estimation problem into informative and manageable pieces.

Marginal Pathway Index. For each pathway, we fit a lasso-penalized Cox model (Tibshirani, 1997) using all genes in the pathway. This is a log-linear model for the hazard in the kth pathway, $h_k(t, x)$, where $h_{0k}(t)$ is the baseline hazard:

$$\log h_k(t_i, x_i) = \log h_{0k}(t_i) + \sum_{j \in \mathcal{P}_k} x_{ij} \beta_j^{(k)}. \tag{2}$$

The signs of $\hat{\beta}_j^{(k)}$ indicate the risk direction of a gene's pathway effect. We call genes whose increased expression is associated with higher risk "susceptibility" genes and genes whose decreased expression is associated with higher risk "resistance genes." Accordingly, let $\bar{X}_k^S$ be the mean expression of all genes with positive coefficients (i.e., $\{i : \hat{\beta}_i^{(k)} > 0\}$) and $\bar{X}_k^R$ be the mean expression of all genes with negative coefficients.

We construct a summary measure of risk in pathway k for each patient i, by comparing the mean expression of the susceptibility genes to the resistance genes as follows:

$$Z_{ki} = \bar{X}_{ki}^S - \bar{X}_{ki}^R. \tag{3}$$

We denote the score for a pathway across all patients as $Z_k = (Z_{k1}, \dots, Z_{kn})'$. Note that genes with zero coefficients are not selected into the model, and as a result, do not influence the index value. Furthermore, we note that the X_j are centered and scaled to have unit standard deviation as is common in penalized estimation (Tibshirani, 1997).

Both for analysis and interpretability, we find it useful to consider dichotomization at zero which may be interpreted as the average expression of susceptibility genes outweighing the average expression of the resistance genes. These binary features are a useful, personalized description of the pathway-level characteristics of disease for each person. In particular, pathway k is said to be "risky" for patient i if $Z_{ki} > 0$, that is, the deleterious expression of the susceptibility genes outweighs the protective expression of the resistance genes ($\bar{X}^S_{ki} > \bar{X}^R_{ki}$).

Joint Prognostic Model. When a single prediction is warranted, we may use these constructed indexes Z_k jointly as features in a final prognostic Cox model. The Cox model minimizes a partial likelihood (Cox, 1972) with linear predictor

$$\eta = \sum_{k=1}^{m} Z_k \alpha_k \tag{4}$$

$$= \sum_{k=1}^{m} \alpha_k \bar{X}^S_k - \alpha_k \bar{X}^R_k \tag{5}$$

Predictions may be made by inputting new data X^*, constructing the pathway scores $Z^*_j = \bar{X}^{*S}_k - \alpha_k \bar{X}^{*R}_k$ (using the signs estimated previously) and computing the predicted relative risk $\eta^* = \sum_{k=1}^m Z^*_k \hat{\alpha}_k$. Note that non-pathway information such as clinical predictors may be incorporated easily into the pathway model by extending η to include non-genetic predictors.

When tuning parameters are required for the pathway models, we use 10-fold cross validation (Gui and Li, 2005). Different strategies for model fitting will be described in the following section.

3. Ovarian cancer patient-specific risk profiles (PSRPs)

A pilot arm of The Cancer Genome Atlas (TCGA) collected high-quality, high-dimensional, and multi-modal genetic data from women with ovarian cancer (OV). Using the gene expression feature set, we consider the use of pathway index models for predicting overall survival after treatment for primary disease. While a number of expression profiles for survival have been proposed, clinical markers like the quantity of residual tumor (cytoreduction) remain the most powerful prognostic markers (Bookman *et al.*, 2009).

As of September 11, 2010 there were 510 samples (all serous cystadenocarcinoma) with aligned survival, clinical and gene expression (Affymetrix HT-HGU133a) data in the public set. We consider the subset of these patients ($n = 503$) who received adjuvant chemotherapy (combined taxol and platinum-based) following surgery. Confounding the analysis is a high recurrence rate (283/503, 56.3%); women who recur receive extra rounds of chemotherapy (often of different types). Throughout this article, we adopt a clinically practical stance and consider overall survival unadjusted for women who experienced recurrences.

We separated the data into the training ($n = 234$) and test sets ($n = 269$), as suggested by the TCGA project, and we make use of two comparable, independent datasets ($n = 146$) available on the Gene Expression Omnibus. Bild *et al.* (2006) (GEO: GSE3149) has $n = 134$ samples after averaging 8 pairs of duplicated arrays and dropping 4 identifiers with missing survival information. Tothill *et al.* (2008) (GEO: GSE9899) was an Australian observational study with $n = 240$ arrays after dropping early stage and low malignant potential cancers.

In table 1, we list the 12 core cancer pathways identified by Jones *et al.* (2008) by their KEGG names (Kanehisa *et al.*, 2010) and subset the expression data into 984 genes which belong to these paths. Note that the KEGG database used four DNA repair pathways and we have included them separately (for 15 total pathways). We used the annotation package `hthgu133a.db` in R to match probes to pathways.

3.1. PSRPs predict survival

We subset the data to the 984 genes in the pathways listed in table 1 and fit our pathway index model to generate patient specific risk profiles (PSRPs). The number of genes selected for individual pathways varied; the top effects listed are heavily weighted towards known oncogenes and tumor suppressors. Pathways which had no genes selected are dropped from further analysis. Overall, 111 genes were selected (92 unique) and 12 pathways remain.

For comparison, we also fit a lasso penalized Cox model (Tibshirani, 1997) and SIS lasso Cox model (Fan and Song, 2010) to the data using the same set of 984 genes. These models selected 31 and 36 genes respectively, agreeing on 29 of them. Of these, 22 of the 29 were also found in our pathway model.

Table 1
Selected genes for core cancer pathways identified in Jones et al. (2008).

KEGG Pathway Name	# Genes Selected	Top 3 genes
MAPK signaling pathway	28	DUSP8, MYC, PDGFA
Wnt Signaling pathway	20	APC, PPP3CA, LEF1
TGF-beta signaling pathway	13	MYC, LTBP1, ROCK1
Base excision repair	12	SMUG1, FEN1, NTHL1
Jak-STAT signaling pathway	8	AKT2, MYC, IL2RG
Hedgehog signaling pathway	7	CSNK1G3, GAS1, WNT5B
Non-homologous end-joining	6	FEN1, XRCC4, DNTT
mTOR signaling pathway	6	AKT2, AKT1, TSC1
Nucleotide Excision Repair	4	DDB1, CETN2, XPA
Apoptosis	3	AKT2, PPP3CA, APAF1
Cell adhesion	2	CD6, HLA-DOB
Cell cycle	2	MYC, YWHAB
Notch Signaling pathway	0	–
Phosphatidylinositol	0	–
Mismatch repair	0	–

We compare the prognostic performance of the risk scores predicted by these models in figure 1. While they were tuned by cross-validation, note that the extra genes selected by Lasso and SIS Lasso lead to reduced performance. We show the results for one pathway involved in immune response FCϵ-RI signaling (KEGG: map04664) which selected 6 genes: AKT1, AKT2, PLCG, RAC1, SYK and PLA2G2D. Of these, AKT2, RAC1 and PLA2G2D were also selected by the Lasso and SIS-Lasso models.

Fig 1. *We compare performance of lasso (top row), SIS-Lasso (Middle) and a single pathway model (bottom) in the TCGA testing portion (left) and independent test data from Tothill et al. (2008) (center) and Bild et al. (2006) (right).*

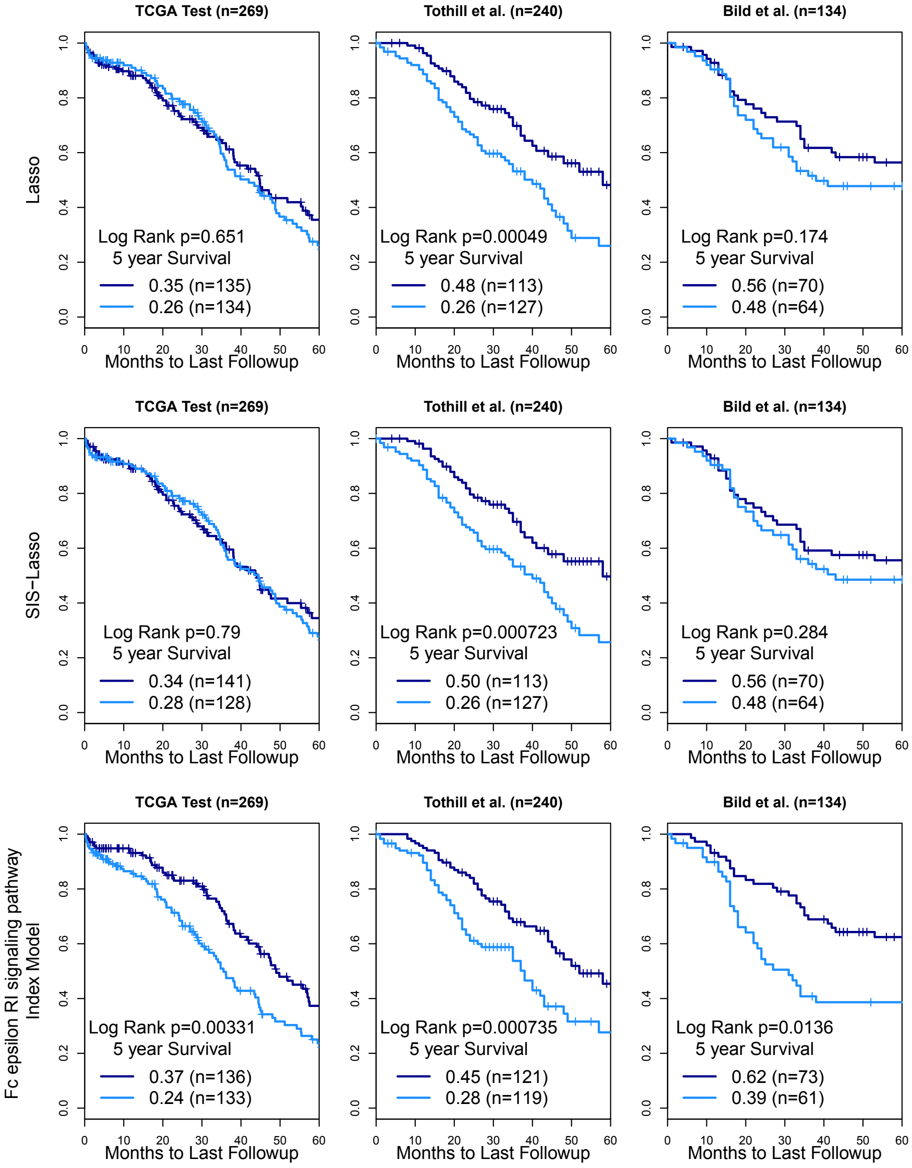
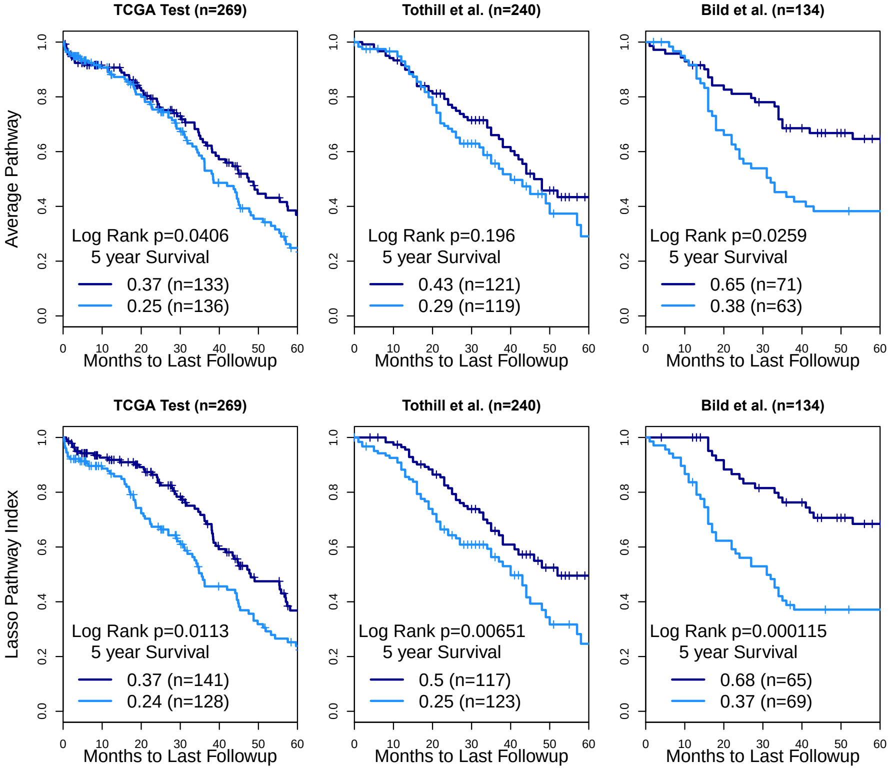


Fig 2. *We compare performance of the average expression within pathway (top row) and a lasso penalized Cox model across 12 single pathway models (bottom).*



3.2. Combining PSRPs improves prediction

Considering the core cancer pathways with non-zero models, each patient has 12 pathway-specific risk scores. We consider two ways of aggregating these to form a single prognostic marker. We use the 12 risk scores Z_k as predictors in a lasso penalized Cox model, the resulting fit is plotted in figure 2. This is similar to a method described by Rosenwald *et al.* (2002) who used the average expression of genes in clusters as a covariate in a multivariate Cox model. To illustrate the gain from estimating the signs of the effects, we adapt this model and include it in figure 2.

Recalling that each pathway was called "risky" if the average of the susceptibility genes outweighs the average of the resistance genes, a second method of aggregation is to simply compare the average of susceptibility signals across all pathways versus the average of all resistance signals. We define our aggregated continuous score to be

$$\left(\sum_{k=1}^{1} 2\bar{X}_k^S\right) - \left(\sum_{k=1}^{1} 2\bar{X}_k^R\right). \tag{6}$$

For the training data set, the top row of Figure 3 clearly shows a gradient of risk associated with the number of pathways involved regardless of their type. Taken further, if we simply count the number of risky pathways, we obtain a similar characterization of risk (bottom row of Figure 3).

The effects in the validation datasets are more varied but consistent: more pathways in the risk direction are associated with poorer prognosis and fewer pathways at risk imply better survival outcomes. It can be shown that the performance of our predictions in the independent studies is comparable to the original models trained in these data sets.

These effects are likely due to the non-uniform treatment for recurrences. A plurality of patients are expected to recur (Bhoola and Hoskins, 2008) after which the best choice among a number of chemotherapy agents is unclear. That is, in observational studies, disease management confounds our attempts to find a clear, singular underlying disease process. Accounting for differential treatment, in particular predicting optimal treatments will be the purview of future work.

3.3. PSRPs identify clinically meaningful pathway heterogeneity

A useful observation from the count-based model is the likely inadequacy of two group predictions for this cancer. We binarized the the pathway signatures into $Z_k \geq 0$ and $Z_k < 0$ and applied hierarchical clustering. Under a manhattan metric, there is a natural split into three clusters. The means of these clusters (table 2) are the joint frequencies of risk direction expression for the 12 pathways. When compared to the estimated 5-year survival times, we observe that the two high-risk categories have roughly the same prognosis but have mutually exclusive sets of pathways. We conjecture that, unlike the previous model, these two groups reflect two different subtypes with the same prognosis. Taken a bit further, high risk group A is made up of mostly DNA repair pathways and their predicate apoptosis pathway while high risk group B is made up of pathways which signal growth.

4. Discussion

The last decade of genome research has had a profound impact on scientific progress that, unfortunately, has not been matched by a similar impact on personalized genomic medicine. A major obstacle is the inability to easily and comprehensively characterize the genomic features affecting an individual's specific disease course and likelihood of response to treatment. There are a number of statistical approaches currently available for identifying genes useful for

Fig 3. *Two methods for aggregating risk across pathways are show in each row. Green curves have fewer pathways at risk, yellow are intermediate and red have many pathways at risk. The four columns correspond to the training, testing and two independent data sets (left to right).*

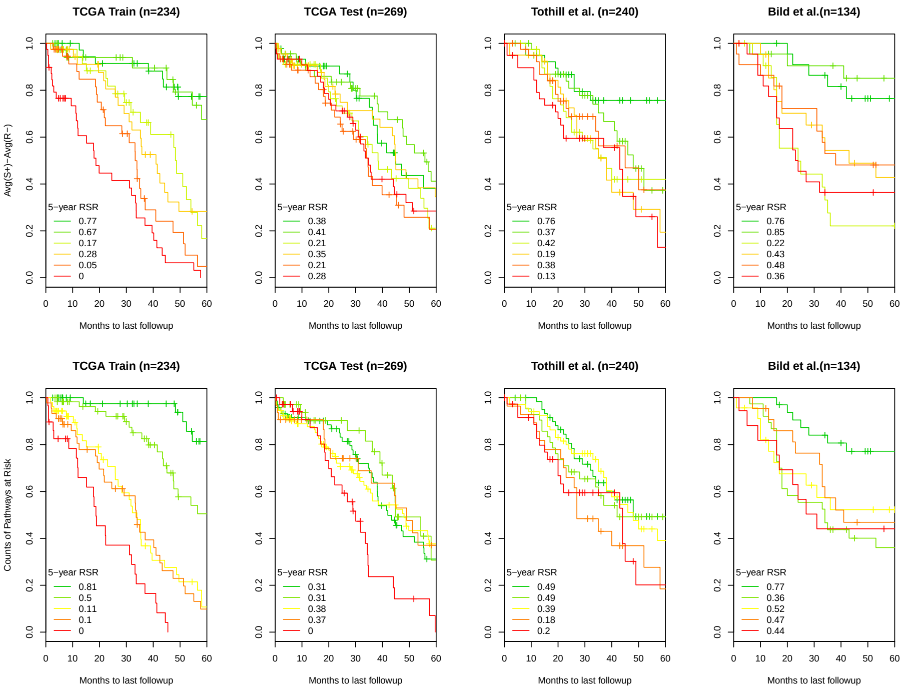


TABLE 2
Frequencies of risk direction expression by pathway for 3 clusters. Pathways with at least 0.70 *frequencies are bolded to emphasize pathway clustering. Note that the two high risk clusters have similar 5 year survival fractions.*

Frequency of Expression	Low Risk	High Risk A	High Risk B
Jak-STAT signaling	0.40	0.54	**0.94**
Wnt Signaling	0.30	0.53	**0.97**
MAPK signaling	0.27	0.62	**0.88**
Cell adhesion	0.52	0.54	**0.70**
Cell cycle	0.40	0.53	**0.79**
Non-homologous end-joining	0.38	**0.81**	0.24
mTOR signaling	0.25	**0.82**	0.67
Apoptosis	0.22	**0.80**	0.64
Base excision repair	0.26	**0.76**	0.64
Nucleotide Excision Repair	0.36	**0.71**	0.61
TGF-beta signaling	0.43	0.57	0.52
Hedgehog signaling	0.47	0.56	0.39
5 year survival (Kaplan Meier)			
TCGA Train	0.56	0.12	0.05
TCGA Test	0.38	0.22	0.26
Tothill *et al.* (2008)	0.48	0.29	0.29
Bild *et al.* (2006)	0.67	0.42	0.38

predicting survival related phenotypes, with utility demonstrated in studies of ovarian cancer similar to the TCGA ovarian cancer case study considered here (Berchuck *et al.*, 2005; Bild *et al.*, 2006; Crijns *et al.*, 2009; Denkert *et al.*, 2009; Dressman *et al.*, 2007; Tothill *et al.*, 2008; Yoshihara *et al.*, 2010). While the utility of these methods for prediction and novel discovery is considerable, it is often difficult to translate the derived gene lists into guidance on how to treat an individual patient. Furthermore, gene level approaches sacrifice considerable power in diseases such as cancer where genetic heterogeneity is proving to be more the rule than the exception. As noted in Jones *et al.* (2008) and a number of subsequent studies, individual tumors often show very little overlap at the gene level (in particular mutations, for example), but much consensus at the pathway level.

To address this we have proposed a pathway index model that relies on known pathways and provides patient-specific risk profiles (PSRPs) which specify the pathways in a patient conferring increased survival risk. The PSRPs provide for refined patient stratification that outperforms methods currently used in most ovarian cancer studies (Bild *et al.*, 2006; Tothill *et al.*, 2008). Perhaps more importantly, given that the pathways have some known relevance to disease and oftentimes known response to certain treatments, the PSRPs may prove to be directly useful in a clinical context. For example, knowing that a patient is at increased risk for a DNA repair pathway suggests platinum-based chemotherapy while aberrant mitotic processes suggest a taxane based treatment (Bhoola and Hoskins, 2008).

Screening important pathways using a *marginal model* is similar in spirit to sure independence screening (SIS) (Fan, Feng and Wu, 2010; Fan and Song, 2010), where there is reduced cost to adding a screening step prior to formal variable selection when there are few important predictors. Screening, therefore, brings the dimension of the problem into a feasible range. With the molecular profiling of twenty more cancers from the TCGA project underway, a number of applications and extensions are to be expected.

Acknowledgements

The authors thank Bret Hanlon and John Dawson for comments that helped improve the manuscript. This work is supported by T32-HL083806 (KE) and GM076274 (CK).

References

Berchuck, A., Iversen, E. S., Lancaster, J. M., Pittman, J., Luo, J., Lee, P., Murphy, S., Dressman, H. K., Febbo, P. G., West, M., Nevins, J. R. and Marks, J. R. (2005). Patterns of Gene Expression that Characterize Long-Term Survival in Advanced Stage Serous Ovarian Cancers. *Clinical Cancer Research* **11** 3686-3696.

Bhoola, S. and Hoskins, W. J. (2008). Diagnosis and Management of Epithelial Ovarian Cancer. *Obstetrics and Gynecology* **107** 1399-1410.

Bild, A., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J. M., Berchuck, A., Olson, J. A., Marks, J. R., K., D. H., West, M. and Nevins, J. R. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439** 353-357.

Bookman, M. A., Brady, M. F., McGuire, W. P., Harper, P. G., Alberts, D. S., Friedlander, M., Colombo, N., F., F. J., Argenta, P. A., De Geest, K., Mutch, D. G., Burger, R. A., Swart, A. M., Trimble, E. L., Accario-Winslow, C. and Roth, L. M. (2009). Evaluation of new platinum-based treatment regimens in advanced-stage ovarian cancer: a phase II trial of the gynecologic cancer intergroup. *Journal of Clinical Oncology* **27** 1419-1425.

Clark-Langone, K., Sangli, C., Krishnakumar, J. and Watson, D. (2010). Translating tumor biology into personalized treatment planning: analytical performance characteristics of the Oncotype DX (R) Colon Cancer Assay. *BMC cancer* **10** 691.

Collins, F. (2010). Has the revolution arrived? *Nature* **464** 674–675.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34** 187-220.

Craver, M. (2010). Genetic medicine finally hitting its stride. The Kiplinger Letter.

Crijns, A. P. G., Fehrmann, R. S. N., de Jong, S., Gerbens, F., Meersma, G. J., Klip, H. G., Hollema, H., Hofstra, R. M. W., te Meerman, G. J., de Vries, E. G. E. and van der Zee, A. G. J. (2009). Survival-Related Profile, Pathways, and Transcription Factors in Ovarian Cancer. *PLoS Medicine* **6** e1000024.

Denkert, C., Budczies, J., Darb-Esfahani, S., Györffy, B., Sehouli, J., Könsgen, D., Zeillinger, R., Weichert, W., Noske, A., Buckendahl, A.-C., Müller, B. M., Dietel, M. and Lage, H. (2009). A prognostic gene epxression index in ovarian cancer– validation across different independent data sets. *Journal of Pathology* **218** 273-280.

Dressman, H. K., Berchuck, A., Chan, G., Zhai, J., Bild, A., Sayer, R., Cragun, J., Clarke, J., Whitaker, R. S., Li, L., Gray, J., Marks, J., Ginsburg, G. S., Potti, A., West, M., Nevins, J. R. and Lancaster, J. M. (2007). An Integrated Genomic-Based Approach to Individualized Treatment of Patients with Advanced-Stage Ovarian Cancer. *Journal of Clinical Oncology* **25** 517-525.

Fan, J., Feng, Y. and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional Hazards model. arXiv:1002.3315v2.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality.

Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21** 3001.

Huang, J., Ma, S., Xie, H. and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S.-M., Fu, B., Lin, M.-T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D. R., Hidalgo, S. D. M. Leach, Klein, A. P., Jaffee, E. M., Goggins, M., Maitra, A., Iacobuzio-Donahue, J. R. C. amd Eshleman, Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E. and Kinzler, K. W. (2008). Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science* **321** 1801.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research* **38** D355-D360.

Luan, Y. and Li, H. (2008). Group additive regression models for genomic data analysis. *Biostatistics* **9** 100.

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

Mook, S., Veer, L., Emiel, J., Piccart-Gebhart, M. and Cardoso, F. (2007). Individualization of Therapy Using Mammaprint® ì: from Development to the MINDACT Trial. *Cancer Genomics-Proteomics* **4** 147.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B. and Staudt, L. M. (2002). The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large-B-Cell Lymphoma. *New England Journal of Medicine* **346** 1937-1947.

Sparano, J. and Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology* **26** 721.

Tibshirani, R. (1997). The Lasso method for variable selection in the cox model. *Statistics in Medicine* **16** 385-395.

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., Johnson, D. S., Trivett, M. K., Etemadmoghadam, D., Locandro, B., Traficante, N., Fereday, S., Hung, J. A., Chiew, Y.-E., Haviv, I., Group, A. O. C. S., Gertig, D., deFazio, A. and Bowtell, D. D. L. (2008). Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome. *Clinical Cancer Research* **14** 5198-5208.

Wang, S., Nan, B., Zhu, N. and Zhu, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika* **96** 307.

Wei, Z. and Li, H. (2007). Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics* **8** 265.

Yoshihara, K., Tajima, A., Yahata, T., Kodama, S., Fujiwara, H., Suzuki, M., Onishi, Y., Hatae, M., Sueyoshi, K., Fujiwara, H., Kudo, Y., Kotera, K., Masuzaki, H., Tashiro, H., Katabuchi, H., Inoue, I. and Tanaka, K. (2010). Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. *PLoS One* **5** e9615.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67.

Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37** 3468–3497.

Zhou, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *Arxiv preprint arXiv:1006.2871*.