Network-based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment

Wei Zhang[1], Takayo Ota[2], Viji Shridhar[2], Jeremy Chien[2], Baolin Wu[3], Rui Kuang[1]*

1 Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, Minnesota, United States of America, **2** Department of Laboratory Medicine and Experimental Pathology, Mayo Clinic College of Medicine, Rochester, Minnesota, United States of America, **3** Division of Biostatistics, School of Public Health, University of Minnesota Twin Cities, Minneapolis, Minnesota, United States of America

Abstract

Cox regression is commonly used to predict the outcome by the time to an event of interest and in addition, identify relevant features for survival analysis in cancer genomics. Due to the high-dimensionality of high-throughput genomic data, existing Cox models trained on any particular dataset usually generalize poorly to other independent datasets. In this paper, we propose a network-based Cox regression model called Net-Cox and applied Net-Cox for a large-scale survival analysis across multiple ovarian cancer datasets. Net-Cox integrates gene network information into the Cox's proportional hazard model to explore the co-expression or functional relation among high-dimensional gene expression features in the gene network. Net-Cox was applied to analyze three independent gene expression datasets including the TCGA ovarian cancer dataset and two other public ovarian cancer datasets. Net-Cox with the network information from gene co-expression or functional relations identified highly consistent signature genes across the three datasets, and because of the better generalization across the datasets, Net-Cox also consistently improved the accuracy of survival prediction over the Cox models regularized by $L_2-\mathrm{norm}$ or $L_1-\mathrm{norm}$. This study focused on analyzing the death and recurrence outcomes in the treatment of ovarian carcinoma to identify signature genes that can more reliably predict the events. The signature genes comprise dense protein-protein interaction subnetworks, enriched by extracellular matrix receptors and modulators or by nuclear signaling components downstream of extracellular signal-regulated kinases. In the laboratory validation of the signature genes, a tumor array experiment by protein staining on an independent patient cohort from Mayo Clinic showed that the protein expression of the signature gene FBN1 is a biomarker significantly associated with the early recurrence after 12 months of the treatment in the ovarian cancer patients who are initially sensitive to chemotherapy. Net-Cox toolbox is available at http://compbio.cs.umn.edu/Net-Cox/.

Citation: Zhang W, Ota T, Shridhar V, Chien J, Wu B, et al. (2013) Network-based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment. PLoS Comput Biol 9(3): e1002975. doi:10.1371/journal.pcbi.1002975

Editor: Satoru Miyano, University of Tokyo, Japan

Received May 10, 2012; **Accepted** January 23, 2013; **Published** March 21, 2013

Funding: This work was supported by NSF grant III1117153 (http://www.nsf.gov/) and NIH grants GM083345, CA134848, CA136393 (http://www.nih.gov/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kuang@cs.umn.edu

Introduction

Survival analysis is routinely applied to analyzing microarray gene expressions to assess cancer outcomes by the time to an event of interest [1–3]. By uncovering the relationship between gene expression profiles and time to an event such as recurrence or death, a good survival model is expected to achieve more accurate prognoses or diagnoses, and in addition, to identify genes that are relevant to or predictive of the events [4,5]. The Cox proportional hazard model [6] is widely used in survival analysis because of its intuitive likelihood modeling with both uncensored patient samples and censored patient samples who are event-free by the last follow-up. Due to the high dimensionality of typical microarray gene expressions, the Cox regression model is usually regularized with penalties such as L_2 penalty in ridge regression [7–10], L_1 Lasso regularization [11–16] and L_2 regularization in Hilbert space [17]. While those penalties were designed as a statistical or algorithmic treatment for the high-dimensionality problem, these Cox models are still prone to noise and overfitting to the low sample size. An important prior information that has been largely ignored in survival analysis is the modular relations among gene expressions. Groups of genes are co-expressed under certain conditions or their protein products interact with each other to carry out a biological function. It has been shown that protein-protein interaction network or co-expressions can provide useful prior knowledge to remove statistical randomness and confounding factors from high-dimensional data for several classification and regression models [18–21]. The major advantage of these network-based models is the better generalization across independent studies since the network information is consistent with the conserved patterns in the gene expression data. For example, previous studies in [18,20] discovered that more consistent signature genes of breast cancer metastasis can be identified from independent gene expression datasets by network-based classification models. The observations also motivated several graph algorithms for

Author Summary

Network-based computational models are attracting increasing attention in studying cancer genomics because molecular networks provide valuable information on the functional organizations of molecules in cells. Survival analysis mostly with the Cox proportional hazard model is widely used to predict or correlate gene expressions with time to an event of interest (outcome) in cancer genomics. Surprisingly, network-based survival analysis has not received enough attention. In this paper, we studied resistance to chemotherapy in ovarian cancer with a network-based Cox model, called Net-Cox. The experiments confirm that networks representing gene co-expression or functional relations can be used to improve the accuracy and the robustness of survival prediction of outcome in ovarian cancer treatment. The study also revealed subnetwork signatures that are enriched by extracellular matrix receptors and modulators and the downstream nuclear signaling components of extracellular signal-regulators, respectively. In particular, FBN1, which was detected as a signature gene of high confidence by Net-Cox with network information, was validated as a biomarker for predicting early recurrence in platinum-sensitive ovarian cancer patients in laboratory.

detecting cancer causal genes in protein-protein interaction network [22,23].

In this article, we propose a network-based Cox proportional hazard model called Net-Cox to explore the co-expression or functional relation among gene expression features for survival analysis. The relation between gene expressions are modeled by a gene relation network constructed by co-expression analysis or prior knowledge of gene functional relations. In the Net-Cox model, a graph Laplacian constraint is introduced as a smoothness requirement on the gene features linked in the gene relation network. Figure 1 illustrates the general framework of Net-Cox for utilizing gene network information in survival analysis. In the framework, the cost function of Net-Cox, shown in the box, combines the total likelihood of Cox regression with a network regularization. The total log-likelihood is a function of the linear regression coefficients $\boldsymbol{\beta}$ and the base hazard $h_0(t)$ on each followup time $\{t_1,t_2,...,t_{10}\}$, represented by the likelihood ratios with the patient gene expression data and the survival information specified by followup times and event indicators. The gene network is either constructed with gene co-expression information or a given gene functional linkage network. The gene network is modeled as a constraint to encourage smoothness among correlated genes, for example gene i and j in the network, such that the coefficients of the genes connected with edges of large weights are similarly weighted. The cost function of Net-Cox can be solved by alternating optimization of $\boldsymbol{\beta}$ and $h_0(t)$ by iterations. An algorithm that solves the Net-Cox model in its dual representation is also introduced to improve the efficiency. The complete model is explained in detail in Section **Materials and Methods**.

In this study, we applied Net-Cox to identify gene expression signatures associated with the outcomes of death and recurrence in the treatment of ovarian carcinoma. Ovarian cancer is the fifth-leading cause of cancer death in US women [3]. Identifying molecular signatures for patient survival or tumor recurrence can potentially improve diagnosis and prognosis of ovarian cancer. Net-Cox was applied on three large-scale ovarian cancer gene expression datasets [3,24,25] to predict survivals or recurrences and to identify the genes that may be relevant to the events. Our study is fundamentally different from previous survival analysis on ovarian cancer [3,24–26], which are based on univariate Cox regression. For example, in [3], gene expression profiles from 215 stage II–IV ovarian tumors from TCGA were used to identify a prognostic gene signature (univariate Cox $p-\mathrm{value}<0.01$) for overall survival, including 108 genes correlated with poor (worse) prognosis and 85 genes correlated with good (better) prognosis. In [24], a Cox score is defined to measure the correlation between gene expression and survival. The genes with a Cox score that exceeds an empirically optimized threshold in leave-one-out cross-validation were reported as signature genes. Similarly, in [25] and [26], a univariate Cox model was applied to identify association between gene expressions and survival (univariate Cox $p-\mathrm{value}<0.01$). Our study is based on gene networks enriched by co-expression and functional information and thus identifies subnetwork signatures for predicting survival or recurrence in ovarian cancer treatment.

Results

In the experiments, Net-Cox was applied to analyze three ovarian cancer gene expression datasets listed in Table 1. Net-Cox (equation (9)) was compared with $L_2-\mathrm{Cox}$ (equation (6)) and $L_1-\mathrm{Cox}$ (equation (7)) with performance evaluation in survival prediction and gene signature identification for the analysis of patient survival and tumor recurrence. First, for evaluation with a better focus on cancer-relevant genes, the expressions of a list of 2647 genes that are previously known to be related to cancer (Sloan-Kettering cancer genes) are used. On the data of these 2647 genes, Net-Cox, $L_2-\mathrm{Cox}$ and $L_1-\mathrm{Cox}$ were evaluated by consistency of signature gene selection across the three datasets, accuracy of survival prediction and assessment of statistical significance. Next, more comprehensive experiments on all 7562 mappable genes were conducted to identify novel signature genes associated with ovarian cancer. Finally, we further analyzed and validated ovarian cancer signatures by an additional tumor array experiment and literature survey. In all the experiments, gene co-expression networks and a gene functional linkage network were used to derive the network constraints for Net-Cox. The details of data preparation and the algorithms are described in Section **Materials and Methods**.

Net-Cox identifies consistent signature genes across independent datasets

To evaluate the generalization of the models, we first measured the consistency among the signature genes selected from the three independent datasets by each method. Specifically, we report the percentage of common genes in the three rank lists identified by a method. This measurement assumes that even under the presence of biological variability in gene expressions and patient heterogeneities in each dataset, genes that are selected in multiple datasets are more likely to be true signature genes. Thus, higher consistency across the datasets might indicate higher quality in gene selection.

In Figure 2, we plot the number of common genes among the first k (up to 300) genes in the gene ranking lists from all of the three datasets for the death event and two datasets (TCGA and Tothill) for the recurrence event. For the parameter setting of Net-Cox, we fixed λ to be the optimal parameter in the five-fold cross-validation (see Section **Materials and Methods** and report the results with $\alpha=0.01$ and 0.5. Since the ranking lists of Net-Cox with $\alpha=0.95$ are nearly identical to those of $L_2-\mathrm{Cox}$, they are not reported for better clarity in the figure. The first observation is

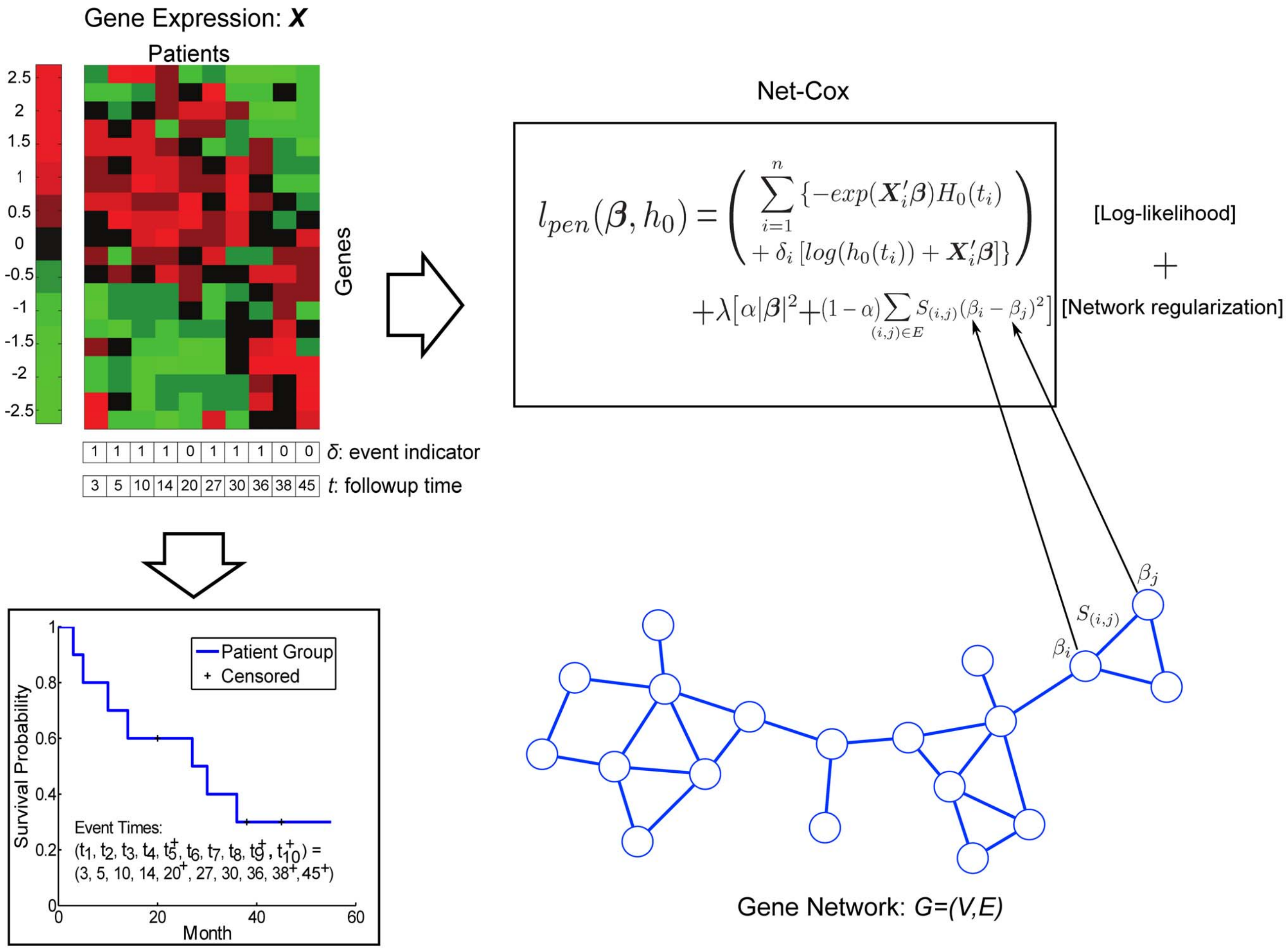

Figure 1. Overview of Net-Cox. The patient gene expression data X and the survival information specified by followup times t and event indicators δ are illustrated on the left. The cost function of Net-Cox given in the box combines the total likelihood of Cox regression with a network regularization. The gene network shown is used as a constraint to encourage smoothness among correlated genes, i.e. the coefficients of the genes connected with edges of large weights are similarly weighted.
doi:10.1371/journal.pcbi.1002975.g001

that the gene rankings by Net-Cox are more consistent than those by L_2-Cox and L_1-Cox at all the cutoffs. Moreover, Net-Cox with $\alpha=0.01$ identified more common signature genes than Net-Cox with $\alpha=0.5$. For example, for the tumor recurrence outcome, Net-Cox (Co-expression) with $\alpha=0.01$ and 0.5 identified 36 and 29 common genes among the first 100 genes in the gene ranking lists, Net-Cox (Functional linkage) with $\alpha=0.01$ and 0.5 identified 49 and 23 common genes, and L_2-Cox and L_1-Cox only identified 19 and 6 common genes, respectively. In general, variable selection by L_1-Cox is not stable from high-dimensional gene expression data, and thus, the overlaps in the gene lists by L_1-Cox are significantly lower than the other methods. It is also interesting to see the gradient of the overlap ratio from $\alpha=0.01$ to $\alpha=0.5$, and then to $\alpha=1$ (L_2-Cox), which indicates that, when a gene network plays more an important role in gene selection, the gene rankings tend to be more consistent. This observation is consistent with previous studies with protein-protein interaction network or gene co-expression network [18,20,21]. Note that since

Table 1. Patient samples in the ovarian cancer datasets.

	Dataset (GEO ID)	TCGA (N/A)	Tothill (GSE9899)	Bonome (GSE26712)
Death	**# of Censored**	227	160	24
	# of Uncensored	277	111	129
Recurrence	**# of Censored**	241	86	N/A
	# of Uncensored	263	185	N/A

The number of patients categorized by censoring and uncensoring for the death and recurrent events is reported in each dataset. Note that the Bonome dataset does not provide information on recurrence.
doi:10.1371/journal.pcbi.1002975.t001

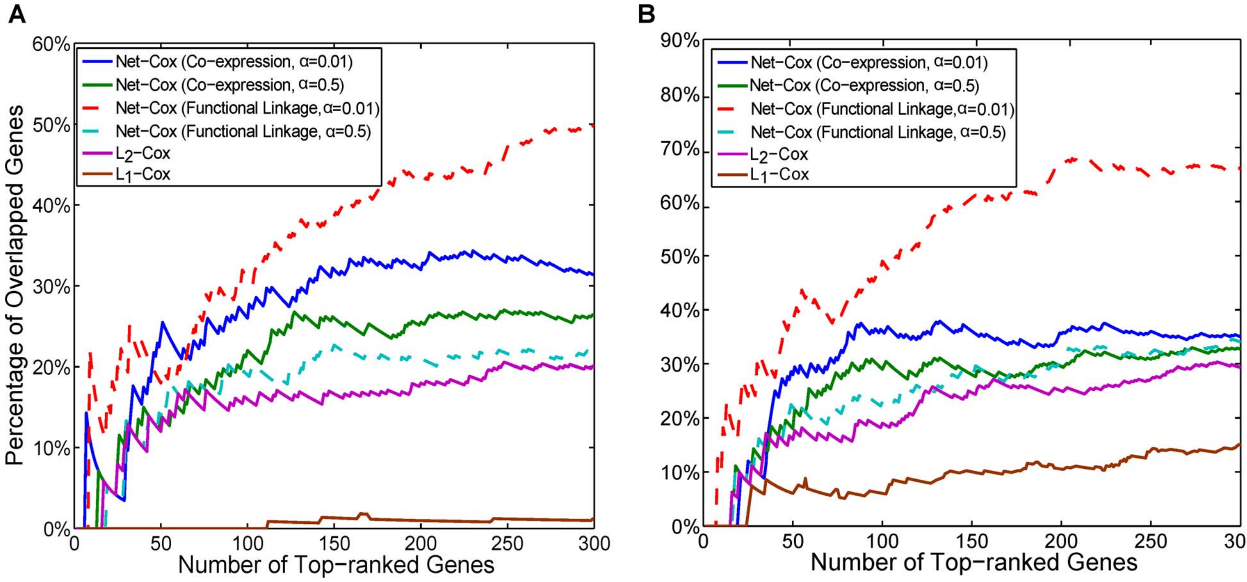

Figure 2. Consistency of signature genes (Sloan-Kettering cancer genes). The x-axis is the number of selected signature genes ranked by each method. The y-axis is the percentage of the overlapped genes between the selected genes across the ovarian cancer datasets. The plots show the results for the death outcome (A) and the tumor recurrence outcome (B).
doi:10.1371/journal.pcbi.1002975.g002

the overlaps are across three datasets for the death event and across two datasets for the recurrence event, the overlaps for the death event is expected to be lower than those for the recurrence event. Another important difference is that the same functional linkage network is always used while the co-expression network is dataset-specific. Thus, it is also expected that the overlaps by Net-Cox with the functional linkage network is higher than those by Net-Cox with the co-expression network. Together, the results demonstrate that Net-Cox effectively utilized the network information to improve gene selection and accordingly, the generalization of the model to independent data.

Net-Cox improves survival prediction across independent datasets

Five-fold cross-validation was first conducted for parameter tuning for Net-Cox, $L_2-\mathrm{Cox}$ and $L_1-\mathrm{Cox}$ on each dataset. The optimal parameters of Net-Cox are reported in Table S1. To test how well the models generalize across the datasets, we trained Net-Cox model, $L_2-\mathrm{Cox}$ model, and $L_1-\mathrm{Cox}$ model with the TCGA dataset, and then predicted the survival of the patients in the other two datasets with the TCGA-trained models. In training, we used the optimal λ and α from the five-fold cross-validation to train the models with the whole TCGA dataset. The results are given in Table 2. In all the cases, Net-Cox obtained more significant $p-\mathrm{values}$ in the log-rank test than $L_2-\mathrm{Cox}$ and $L_1-\mathrm{Cox}$. To further compare the results, we show the Kaplan-Meier survival curves and the ROC curves in Figure 3. The first four columns of plots in the figure show the Kaplan-Meier survival curves for the two risk groups defined by Net-Cox with co-expression network and functional linkage network, $L_2-\mathrm{Cox}$, and $L_1-\mathrm{Cox}$. The fifth column of plots compare the time-dependent area under the ROC curves based on the estimated risk scores (*PIs*). In Figure 3, in many regions, Net-Cox achieved large improvement over both $L_2-\mathrm{Cox}$ and $L_1-\mathrm{Cox}$ while the improvement is less obvious in several other regions. Overall, Net-Cox achieved better or similar AUCs in all the time points in the three plots. To evaluate the statistical significance of the differences between the time-dependent AUCs generated by Net-Cox and the other two methods, in Table S2 we report $p-\mathrm{values}$ at each event time with the null hypothesis that the two time-dependent AUCs estimated by two models are equal. At many points of the event time, the time-dependent AUCs generated from Net-Cox are significant higher.

The cross-validation log-partial likelihood (CVPLs) for the combinations of (λ,α) in the five-fold cross-validation are also reported in Table S3. In all the cases, the optimal CVPLs of Net-Cox are higher than those of $L_2-\mathrm{Cox}$. $L_1-\mathrm{Cox}$ was fine-tuned with 1000 choices of parameters with a very small bin size. In one of the cases (TCGA: Recurrence), the optimal CVPL of $L_1-\mathrm{Cox}$ is higher but in the other cases, the optimal CVPLs of Net-Cox are higher. Interestingly, the optimal α is often 0.1 or 0.5, indicating the optimal CVPL is a balance of the information from gene expressions and the network. The observations prove that the network information is useful for improving survival analysis. The left column of Figure S1 shows the average time-dependent area under the ROC curves based on the estimated risk scores (*PI*) of the patients in the fifth fold of the five repeats, and Table S4A and S4B show log-rank $p-\mathrm{values}$ of the fifth fold of the five repeats. Net-Cox achieved the best overall survival prediction although the results are less obvious than those of the cross-dataset analysis.

Statistical assessment

To understand the role of the gene network on the consistency in gene selection and the contribution to the log-partial likelihood, we tested Net-Cox with randomized co-expression networks. In each randomization, the weighted edges between genes were shuffled. We report the mean and the standard deviation of the percentage of overlapping genes of 50 randomizations in Figure 4. Compared with the consistency plots with the true networks, the overlaps by Net-Cox on the randomized networks are much lower. We also report the boxplot of the log-partial likelihood in the same 50 randomized co-expression network with $\alpha = 0.01$ in Figure 5. Compare with the log-partial likelihood with the real co-expression network, the range of the likelihood generated with

Table 2. Log-rank test $p-$values in cross-dataset evaluation (Sloan-Kettering cancer genes).

	Test Dataset	Net-Cox (Co-exp)	Net-Cox (FL)	$L_2-\mathrm{Cox}$	$L_1-\mathrm{Cox}$
Death	**Tothill**	1.1178E-06	2.5938E-07	2.9932E-06	0.0011
	Bonome	7.6088E-07	3.6039E-06	5.2590E-06	0.1165
Recurrence	**Tothill**	0.0567	0.0786	0.1115	0.4219

The survival prediction performance on Tothill and Bonome datasets using the Cox models trained with TCGA dataset are reported.
doi:10.1371/journal.pcbi.1002975.t002

the randomized networks is again lower by a large margin, which provides clear evidence that the co-expression network is informative for survival analysis.

To further understand the role of the network information in cross-validation, we fixed the optimal parameter λ and conducted the same five-fold cross-validation with randomized co-expression networks to compute the CVPL with different α in $\{0.01, 0.1, 0.5. 0.95\}$. We repeated the process on 20 random networks for each α. The boxplots of CVPLs with different αs are shown in Figure 6. In all measures, the CVPL with the true gene network is well above the mean of the 20 random cases. Another important observation is that, in both plots, when the randomized network information is more trusted with a smaller α, the variance of the CVPLs is also getting larger; and the case with $\alpha=0.01$ gives the worst CVPL mean and the largest variance. The result indicates that the randomized networks did not provide any valuable information in survival prediction. In contrast, with the true gene network, CVPLs generated from $\alpha=0.01$ and $\alpha=0.1$ are much higher than the ones from $\alpha=0.95$ and $L_2-\mathrm{Cox}$ ($\alpha=1$). Again, these results convincingly support the importance of using the network information in survival prediction.

Evaluation by whole gene expression data

Besides the 2647 Sloan-Kettering genes, all the 7562 mappable genes were also tested to evaluate Net-Cox, $L_2-\mathrm{Cox}$ and $L_1-\mathrm{Cox}$ by consistency of signature gene selection across the three datasets and accuracy of survival prediction in similar experiments. For the signature gene consistency, Figure S2 reports

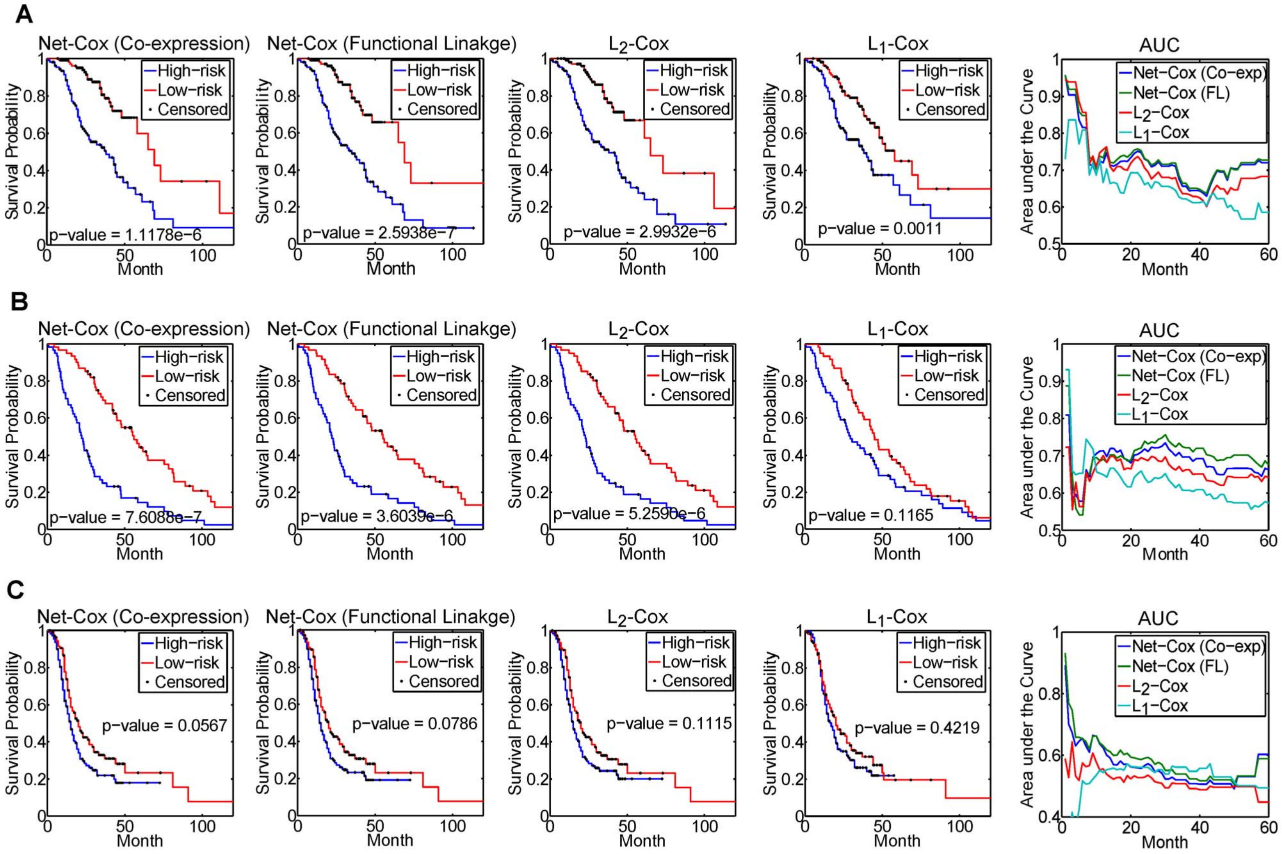

Figure 3. Cross-dataset survival prediction (Sloan Kettering cancer genes). The first four columns of plots show the Kaplan-Meier survival curves for the two risk groups defined by Net-Cox (co-expression network), Net-Cox (functional linkage network), $L_2-\mathrm{Cox}$ and $L_1-\mathrm{Cox}$. The fifth column of plots compare the time-dependent area under the ROC curves based on the estimated risk scores (PIs). The plots show the results for the death outcome by training with TCGA dataset and test on Tothill Dataset (A), the death outcome by training with TCGA dataset and test on Bonome Dataset (B), the tumor recurrence outcome by training with TCGA dataset and test on Tothill Dataset (C).
doi:10.1371/journal.pcbi.1002975.g003

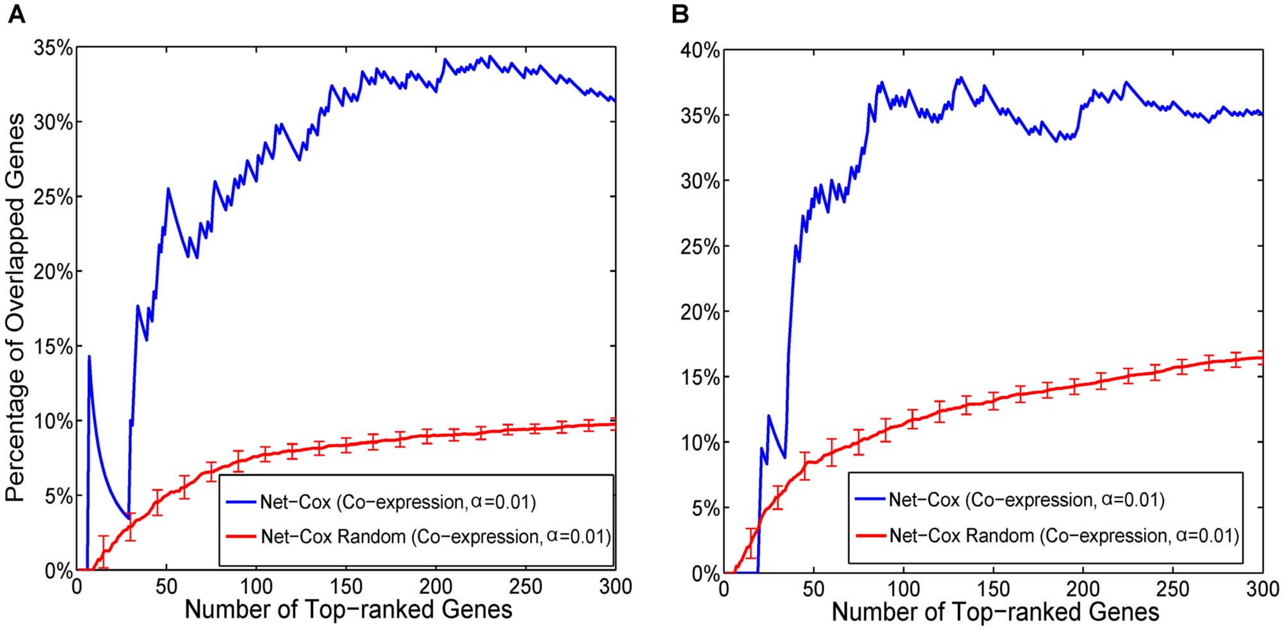

Figure 4. Consistency of signature genes on randomized co-expression networks. The x-axis is the number of selected signature genes ranked by each method. The y-axis is the percentage of the overlapped genes between the selected genes across the ovarian cancer datasets. The red curve reports the mean and the standard deviation of the percentages averaged over the experiments of 50 randomized networks. The plots show the results for the death outcome (A) and the tumor recurrence outcome (B).
doi:10.1371/journal.pcbi.1002975.g004

the percentage of common genes identified by each method in the ranking lists from the datasets. For the cross-dataset validation, Table S5 shows the log-rank test $p-\mathrm{values}$ by training the TCGA datasets and test on the other two datasets, and Figure S3 shows the Kaplan-Meier survival curves for the two risk groups defined by Net-Cox, $L_2-\mathrm{Cox}$ and $L_1-\mathrm{Cox}$ and compares the time-dependent area under the ROC curves. For the five-fold cross-validation, the right column of Figure S1 shows the average time-dependent area under the ROC curves based on the estimated risk scores (PI) of the patients in the fifth fold of the five repeats, and Table S4C and S4D report log-rank test $p-\mathrm{values}$ of the fifth fold of the five repeats. Overall, similar observations are made in experimenting with all the genes, though the improvements are less significant compared with the results by experimenting with the Sloan-Kettering cancer genes. One possible explanation is that, since the genes in the Sloan-Kettering gene list are more cancer relevant, the gene expressions may be more readily integrated with the network information.

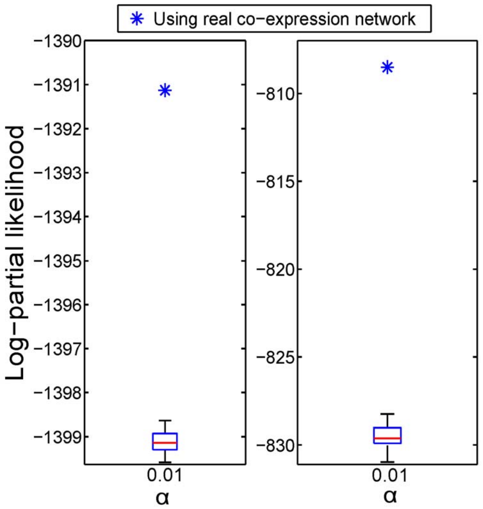

Figure 5. Statistical analysis of log-partial likelihood. The optimal λ was fixed and $\alpha = 0.01$ is set to allow better evaluation of the network information. The log-partial likelihood computed by Net-Cox on the real co-expression network and on the randomized co-expression network are reported against tumor recurrence in the TCGA and Tothill datasets. The stars represent the results with the real co-expression networks, and the boxplots represent the results with the randomized networks.
doi:10.1371/journal.pcbi.1002975.g005

Signature genes are ECM components or modulators

To analyze the signature genes identified by Net-Cox and $L_2-\mathrm{Cox}$, we created consensus rankings across the three datasets by re-ranking the genes with the lowest rank by Net-Cox and $L_2-\mathrm{Cox}$ in the three datasets. Specifically, for each gene, a new ranking score is assigned as the lowest of its ranks in the three datasets, and then, all the genes were re-ranked by the new ranking score. The top-15 genes selected by Net-Cox and $L_2-\mathrm{Cox}$ in the consensus rankings are shown in Table 3. For the death outcome, nine signature genes, FBN1, VCAN, SPARC, ADIPOQ, CNN1, DCN, LOX, EDNRA, LPL, known to be related to ovarian cancer [27–35] are only discovered by Net-Cox. Among the ten common genes highly ranked by both Net-Cox and $L_2-\mathrm{Cox}$, three are collagen genes, and MFAP5, TIMP3, THBS2, and CXCL12 are previously known to be relevant to ovarian cancer [36–39]. For the recurrence outcome, there are eleven common signature genes detected by both Net-Cox and $L_2-\mathrm{Cox}$. Net-Cox identified six additional ovarian cancer related signature genes [27–29,40–42].

The intersection of the 60 genes identified by Net-Cox in Table 3 contains 41 unique genes. We performed a literature survey of the 41 genes, out of which eighteen are supported by

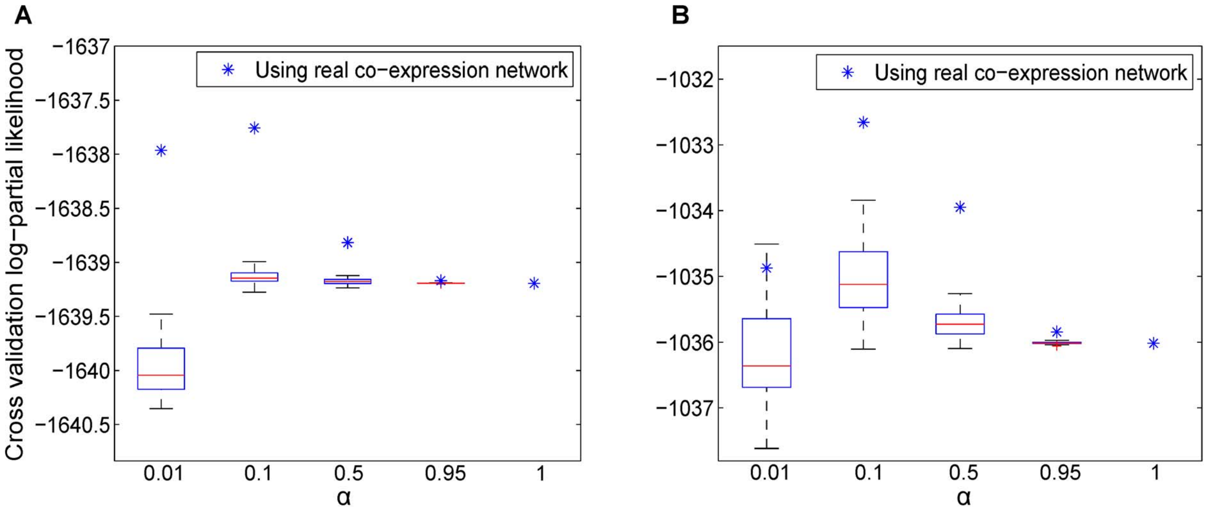

Figure 6. Statistical analysis of cross-validation log-partial likelihood (CVPL). The optimal λ was fixed and α is varied from 0.01 to 1. The CVPL of five-fold cross-validation on the real co-expression network and on the randomized co-expression network are reported against tumor recurrence in TCGA dataset (A) and Tothill dataset (B). The stars represent the results with the real co-expression networks, and the boxplots represent the results with the randomized networks.
doi:10.1371/journal.pcbi.1002975.g006

literature to be related to ovarian cancer shown in Table 4. Most of the genes whose over-expression is associated with poor outcome are stromal or extracellular-related proteins. The genes such as VCAN, TIMP3, THBS2, ADIPOQ, PARC, NPY, MFAP5, DCN, LOX, FBN1, EDNRA, and CXCL12 are either components or modulators of extracellular matrix. In particular, LOX protein is involved in extracellular matrix remodeling by cross-linking collagens. Extracellular matrix remodeling through over-expression of collagens has been shown to contribute to platinum resistance, and platinum resistance is the main factor in chemotherapy failure and poor survival of ovarian cancer patients. Therefore, the identification of these extracellular matrix proteins as biomarkers of early recurrence and poor survival outcome in patients with ovarian cancer is consistent with the suggested pathobiological role of some of these proteins in platinum resistance.

Enriched PPI subnetworks and GO terms

The top-100 signature genes with the largest regression coefficients by Net-Cox and $L_2-\mathrm{Cox}$ learned from the TCGA dataset were mapped to the human protein-protein interaction (PPI) network obtained from HPRD [43] and also analyzed with

Table 3. Top-15 signature genes.

Death			**Recurrence**		
Net-Cox (Co-exp)	**Net-Cox (FL)**	$L_2-\mathrm{Cox}$	**Net-Cox (Co-exp)**	**Net-Cox (FL)**	$L_2-\mathrm{Cox}$
FBN1	COL11A1	COL11A1	COL5A2	COL11A1	COL11A1
COL5A2	MFAP4	FABP4	COL1A1	COL10A1	NLRP2
VCAN	TIMP3	MFAP4	COL5A1	CRYAB	CRYAB
SPARC	MFAP5	COMP	THBS2	NPY	PTX3
AEBP1	COL5A2	BCHE	FAP	IGF1	COL10A1
AOC3	THBS2	FAP	COL3A1	COMP	CXCL12
COL3A1	FAP	COL5A2	COL11A1	KLK5	THBS2
THBS2	CXCL12	MFAP5	FBN1	THBS2	NPY
PLN	AEBP1	TIMP3	VCAN	PI3	KLK5
ADIPOQ	RYR3	THBS2	INHBA	CXCL12	COMP
COL5A1	LOX	HOXA5	CTSK	MFAP5	FAP
CNN1	COL5A1	NUAK1	COL1A2	VGLL1	MFAP5
COL6A2	EDNRA	COL5A1	SPARC	CCL11	PI3
COL1A2	NUAK1	SLIT2	AEBP1	EPHB1	PDGFD
DCN	LPL	CXCL12	SERPINE1	OXTR	CHRDL1

The table lists the genes with over-expression indicating higher hazard of death or recurrence, identified by Net-Cox and $L_2-\mathrm{Cox}$ in the consensus ranking across the three datasets.
doi:10.1371/journal.pcbi.1002975.t003

Table 4. Literature review of the candidate ovarian cancer genes.

Gene Sym	Reference	Description
ADIPOQ	[30]	ADIPOQ 45T/G and 276G/T polymorphisms is associated with susceptibility to polycystic ovary syndrome(PCOS).
CCL11	[42]	CCL11 signaling plays an important role in proliferation and invasion of ovarian carcinoma cells.
CNN1	[31]	CCN1 plays a role in ovarian carcinogenesis by stimulating survival and antiapoptotic signaling pathways.
CRYAB	[71]	Low expression of lens crystallin CRYAB is significantly associated with adverse ovarian patient survival.
CXCL12	[29]	CXCL12 and vascular endothelial growth factor synergistically induce neoangiogenesis in human ovarian cancers.
DCN	[42]	Ovarian DCN is an ECM-associated component, which acts as a multifunctional regulator of GF signaling in the primate ovary.
EDNRA	[32]	Endothelin peptide is produced before ovulation and the contractile action of EDN2 within the ovary is facilitated via EDNRA.
FBN1	[27]	FBN1 controls the bioactivity of TGFβs and associate with polycystic ovary syndrome (PCOS).
IGF1	[41]	Ovarian follicular growth is controlled by the production of intraovarian growth regulatory factors such as IGF1.
INHBA	[40]	INHBA is the promoter of TAF4B; TAF4B in the ovary is essential for proper follicle development.
LOX	[33]	Inhibition of LOX expression portends worse clinical parameters for ovarian cancer.
LPL	[35]	LPL is differentially expressed between preoperative samples of ovarian cancer patients and those of healthy controls.
MFAP5	[36]	MAGP2 is an independent predictor of survival in advanced serous ovarian cancer.
NPY	[72]	NPY receptor is expressed in human primary ovarian neoplasms.
SPARC	[29]	SPARC expression in ovarian cancer cells is inversely correlated with the degree of malignancy.
THBS2	[38]	In ovarian cancer an aberrant methylation process is responsible for down-regulation of THBS2.
TIMP3	[37]	TIMP2 and TIMP3 play functional role in LPA-induced invasion as negative regulators.
VCAN	[28]	VCAN V1 isoform is overexpressed in ovarian cancer stroma compared with normal ovarian stroma and ovarian cancer cells.

This table reports the citations that describe relevance of the signature genes with over-expression indicating higher hazard of death or recurrence, identified by Net-Cox across the three datasets.
doi:10.1371/journal.pcbi.1002975.t004

DAVID functional annotation tool [44]. We report the densely connected PPI subnetworks constructed from the 100 genes selected by Net-Cox in Figure 7. Compared with the PPI subnetworks generated from the 100 genes selected by $L_2-\mathrm{Cox}$, which contain 10 genes in the death subnetwork and 6 genes in the recurrence subnetworks (shown in Figure S4), the subnetworks are both larger and denser. The subnetworks identified from the co-expression networks in Figure 7(A) are also larger than the subnetworks identified by the functional linkage network in Figure 7(B) although many genes are shared. In the recurrence subnetworks, DCN, THBS1, and THBS2 are members of the $\mathrm{TGF}-\beta$ signaling KEGG pathway, and FBN1 controls the bioactivity of TGFβs and relates to polycystic ovary syndrome [27]. In addition, ten genes are members of the focal adhesion KEGG pathway. These results point to a possibility that extracellular matrix signaling through focal adhesion complexes may constitute a pathway by which tumor cells escape chemotherapy and produce recurrence in chemotherapy [45]. Nine genes in the death subnetworks are members of the extracellular matrix(ECM)-receptor interaction KEGG pathway, and eighteen genes are annotated as ECM component. It was shown that ECM acts as a model substratum for the preferential attachment of human ovarian tumor cells in vitro [46]. FOS and JUN constitutes a nuclear signaling components downstream of extracellular signal-regulated kinases (ERK1/2) that are mediators of growth factor and adhesion-related signaling pathways [47]. In addition, the genes are also enriched by regulation of gene expression, positive regulation of cellular process, developmental process, transcription regulator activity, and growth factor binding, all of which are well-known cancer relevant functions. The significantly enriched GO functions are listed in Table S6 and Table S7. Extracellular matrix, extracellular region, and extracellular structure organization are consistently the most significantly enriched in the analysis.

Laboratory experiment validates FBN1's role in chemo-resistance

FBN1 was ranked 1st and 8th by Net-Cox with co-expression network in death and recurrence outcomes while $L_2-\mathrm{Cox}$ only ranked FBN1 at 27th and 42nd, respectively. It is interesting to note that in the PPI subnetworks in Figure 7(A), FBN1 is connected with VCAN and DCN, both of which bear the annotation of extracellular matrix. The dense subnetwork boosted the ranking of FBN1 when Net-Cox was applied. We further validated the role of FBN1 in ovarian cancer recurrence using tumor microarrays (TMAs) consisting of a cohort of 78 independent patients (see Section **Materials and Methods**). The expression level of FBN1 in ovarian cancer was scored by one observer who is blinded to the clinical outcome and described as: absent (0), moderate (1), and high (2) as illustrated by Figure 8.

In Figure S5A, the Kaplan-Meier survival curve shows the recurrence for groups by the FBN1 staining scores. At the initial 12 month, there is no difference in the recurrence rate between the groups with high and low FBN1 staining. After 12 month, the recurrence rate is lower in the low staining group. The similar patterns are also observed in the re-examination of the gene expression datasets in Figure S5B–E. Except the TCGA dataset on the Affymetrix platform (Figure S5E), the pattern is clearly observed on the other two platforms, exon arrays and Agilent arrays. The discrepancy in the Affymetrix data could be related to data pre-processing or experimental noise. The plots suggest that FBN1 plays a role on platinum-sensitive ovarian cancer, and it could be developed as a target for platinum-sensitive patients with high FBN1 expression after about 12 months of the treatment.

In the context of ovarian cancer treatment, a platinum-sensitive patient group can be defined as the group of patients who was free of recurrence or developed a recurrence after k month of the treatment, where $k \geq 14$ depends on the treatment plan and the

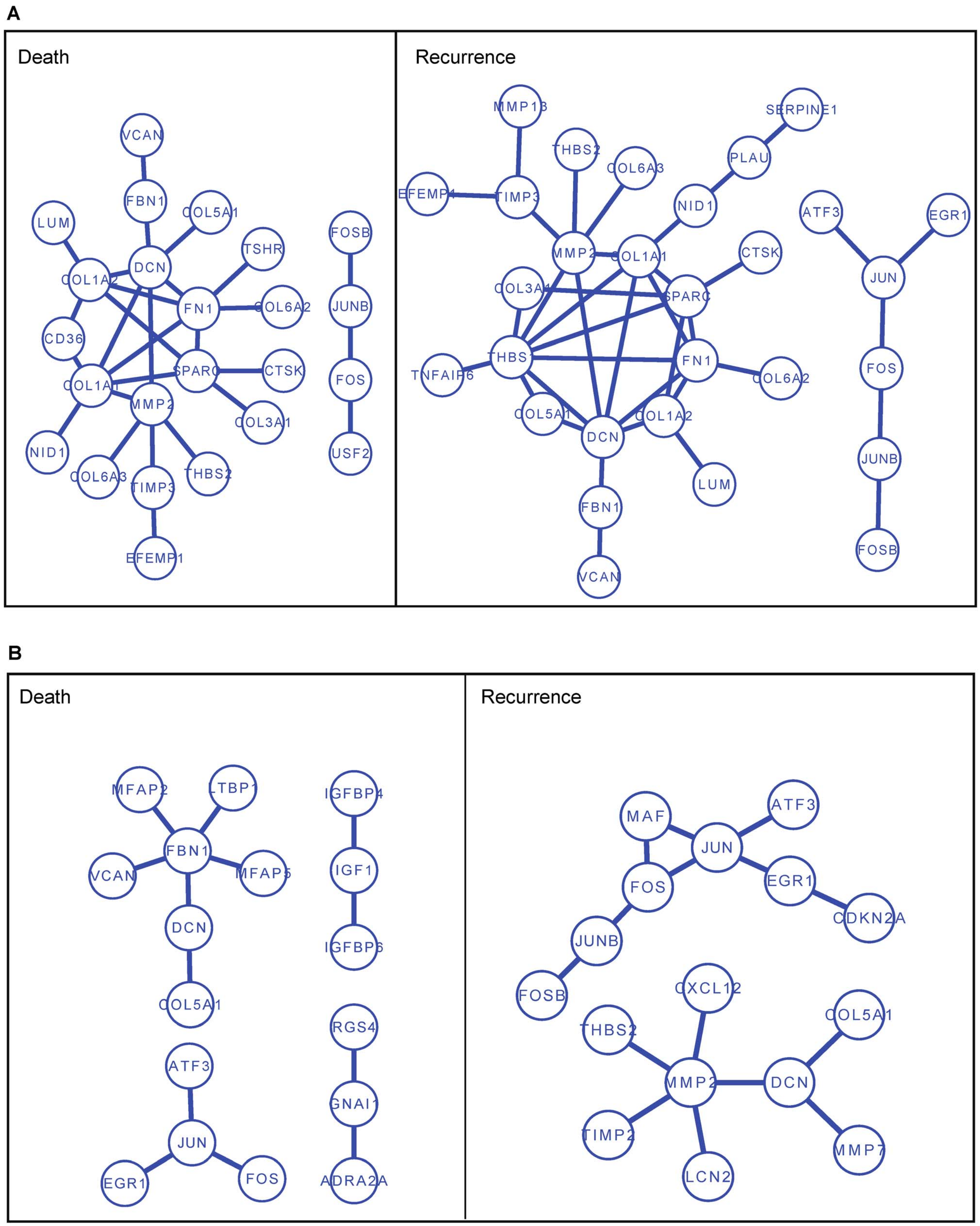

Figure 7. Protein-Protein interaction subnetworks of signature genes identified by Net-Cox on the TCGA dataset. (A) The PPI subnetworks identified by Net-Cox on the co-expression network. (B) The PPI subnetworks identified by Net-Cox on the functional linkage network.
doi:10.1371/journal.pcbi.1002975.g007

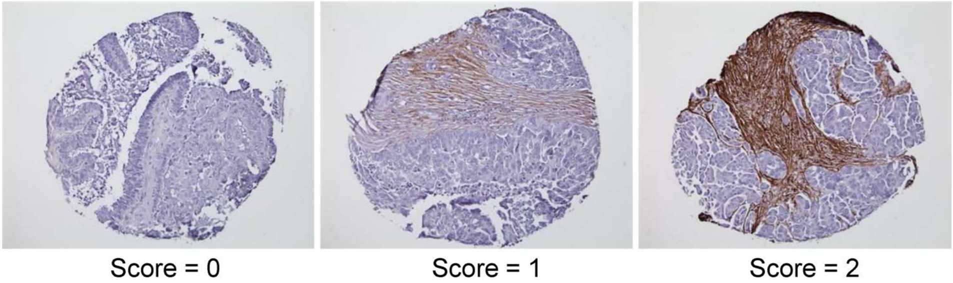

Figure 8. Representative photomicrographs showing various levels of FBN1 expression in ovarian tumor arrays. The brown regions are stromal area showing expression of FBN1.
doi:10.1371/journal.pcbi.1002975.g008

follow-up. To better evaluate the role of FBN1, we plot the Kaplan-Meier survival curve only for the platinum-sensitive patients in Figure 9, i.e. we removed all the patients who developed recurrence before k month and considered the follow-ups up to 72 months after the treatment. Due to the small sample size of the Mayo Clinic data, we set $k=14$ while $k=20$ for the gene expression datasets. In Figure 9A, the difference between the survival curves of low FBN1 staining and high staining patient groups is more significant. Similarly, Figure 9B–E show the survival curves for the platinum-sensitive patients for groups by the expression value of FBN1 in gene expression datasets. Compare to the matched curves in Figure S5, the log-rank test $p-\text{values}$ are more significant except the TCGA dataset on the Affymetrix platform. Overall, the observations strongly support the hypothesized role of FBN1 in platinum-sensitive ovarian cancer patients.

Discussion

Many methods were proposed for survival analysis on high-dimensional gene expression data with highly correlated variates [4,5]. In this paper, we propose Net-Cox, a network-based survival model, which to our knowledge is among the first models that directly incorporate network information in survival analysis. The graph Laplacian constraint introduced in Net-Cox is positive definite and thus, the Net-Cox model can be solved as efficiently as solving the $L_2-\text{Cox}$ model. In the dual form of Net-Cox, the model is scalable to genomic data with $p \gg n$. Net-Cox not only makes survival predictions but also generate densely connected subnetworks enriched by genes with large regression coefficients.

Net-Cox is most related to the L_p shrinkage-based Cox models typically with L_1 (Lasso) and L_2 (ridge) penalties [5]. The purpose of applying L_1 regularization is to obtain a sparse estimate of the linear coefficients for solving the high-dimensionality problem. A Ridge penalty results in small regression coefficients to avoid overfitting problem with the small sample size. Compared with Net-Cox, neither Lasso nor ridge regularized Cox regression models are designed to incorporate any prior information among genes in the objective function for survival analysis. Another alternative solution in the literature is to apply dimension reduction methods to obtain a small number of features for subsequent survival analysis such as principal components analysis (PCA) [48–50] and partial least squares (PLS) [51–54]. These methods first compute the principle components to capture the maximal covariance with the outcomes or the maximal variance in the gene expression data, and then project the original high-dimensional gene expressions into a space of the directions of the principle components. Typically, these methods do not utilize any prior information. It is also usually difficult to interpret the results since the features in the project space are not directly mappable to any particular gene expression. There are also tree-based ensemble methods for survival analysis such as bagging of survival trees and random forests [55,56]. The tree-based methods usually also require a variable selection step to reduce the dimensionality. Multiple trees are then built from different samplings of training data and the results of the individual trees are aggregated for making predictions. Since the trees are built from random sampling, the resulted forests consist of different trees. Thus, the interpretation of the trees can be very difficult [4].

In [57], a supervised group Lasso approach (SGLasso) is proposed to account for the cluster structure in gene expression data as prior information in survival analysis. In this approach, gene clusters are first identified with clustering. Important genes are then identified with Lasso model within each cluster and finally, the clusters are selected with group Lasso. More recently, the method in [58] combined a group Lasso constraint with Lasso Cox regression (sparse-group Lasso). An additional parameter is introduced to balance between Lasso and group Lasso constraints. There are two major discrepancies between Net-Cox and the graph Lasso methods. First, while group Lasso assumes non-overlapping cluster structures among gene expressions, the gene network introduced in Net-Cox captures more global relation among all the genes. Specifically, beyond the cluster partition of genes into co-expression groups, a gene network represents pair-wise relationships between genes, which contain information of modularities, subgraph structures and other global properties such as centralities and closenesses. Second, while SGLasso adopts an unsupervised strategy to cluster genes as predefined groups for selection, Net-Cox identifies subnetwork signatures in a supervised manner, in which the selected subnetworks are enriched by genes with large regression coefficients by the design of the network constraint. In Table S3(g), we reported the results of group Lasso and sparse-group Lasso in the five-fold cross-validation with the R package "SGL" [58]. Compared with the CVPLs by the other methods in Table S3(a)–(f), the CVPLs in Table S3(g) for group Lasso and sparse-group Lasso are consistently lowest when 25 or 100 gene clusters are used as groups. Thus, we did not further compare and analyze other results by the group Lasso models.

The experiments in this paper clearly demonstrated that the network information is useful for improving the accuracy of survival prediction as well as increasing the consistency in discovering signature genes across independent datasets. Since the signature genes were discovered based on their relation in the networks, they enrich dense PPI subnetworks, which are useful for pathway analysis. It is also interesting to note that the PPI subnetworks of signature genes identified by Net-Cox on the TCGA dataset is enriched by extracellular matrix proteins such as collagens, fibronectin, and decorin. Previous gene expression studies had identified stromal gene signatures in ovarian tumors to

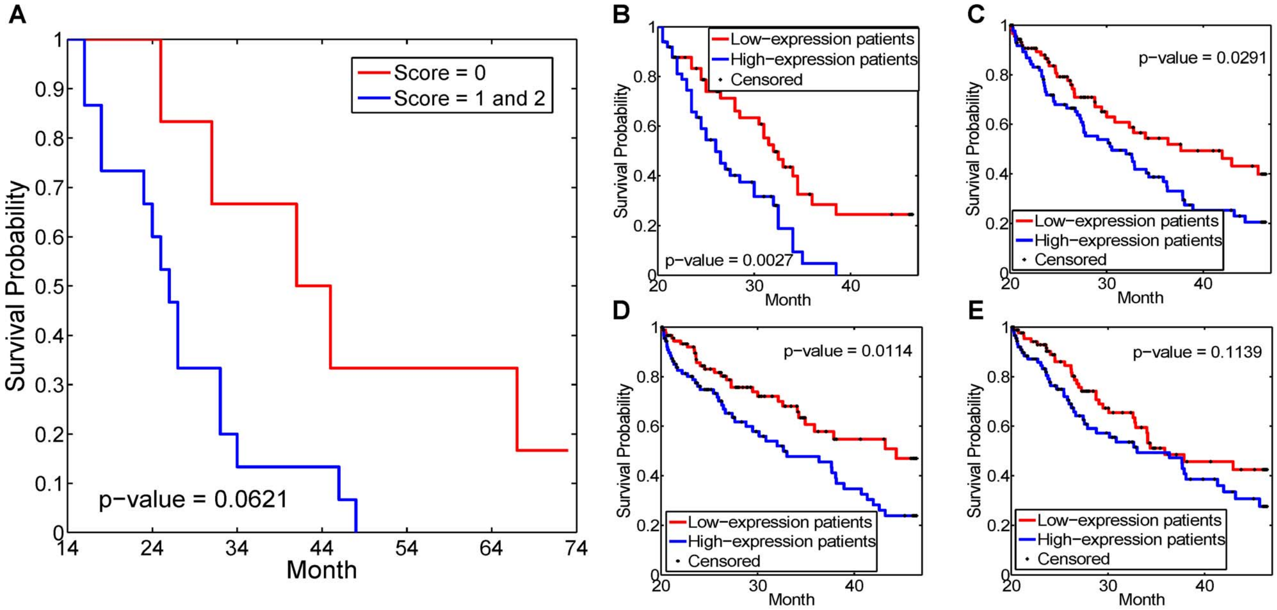

Figure 9. Kaplan-Meier survival plots on FBN1 expression groups. (A) Kaplan-Meier survival curve of recurrence between 14 to 72 month by FBN1 staining groups on Mayo Clinic dataset. (B) Kaplan-Meier survival curve of recurrence between 20 to 72 month by the expression of FBN1 on Tothill dataset. (C)–(E) Kaplan-Meier survival curves of recurrence between 20 to 72 month by the expression of FBN1 on TCGA dataset with AgilentG4502A platform, HuEx-1_0-st-v2 platform, and Affymetrix HG-U133A platform, respectively. In plot(A), the groups with FBN1 staining score 1 and 2 are combined into the high-expression group. In plots(B)–(E), the patients are divided into two groups of the same size by the expression of FBN1.
doi:10.1371/journal.pcbi.1002975.g009

be associated with poor survival outcome [24]. Therefore, our observation that the stromal subnetwork enriched by extracellular matrix proteins and stromal-related proteins is consistent with the role of stromal gene signature in poor prognosis. Finally, collagen matrix remodelling has been linked to platinum resistance, and ovarian cancer cells grown on collagens are more resistant to platinum agents than their counterpart grown on non-collagen substratum [59]. The tumor array validation indicates that FBN1 can serve as a biomarker for predicting recurrence of platinum-sensitive ovarian cancer.

Materials and Methods

This section describes the data preparation, the Cox models and the experimental setup. We first describe the construction of the gene relation networks and the processing of the microarray gene expression datasets. We then review the Cox regression models and introduce the regularization framework of Net-Cox by adding a network constraint to the Cox model. The algorithms to efficiently estimate the optimal solution for Net-Cox are outlined. We also describe the procedures for cross-validation and parameter tuning, and the evaluation measures. At last, tumor array preparation is explained.

Gene relation network construction

We denote gene relation network by $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{W})$, where $\boldsymbol{V}$ is the vertex set, each element of which represents a gene, and $\boldsymbol{W}$ is a $|\boldsymbol{V}|\times|\boldsymbol{V}|$ positively weighted adjacency matrix. $\boldsymbol{D}$ is a diagonal matrix with $D_{ii}=\sum_j W_{ij}$ and $\boldsymbol{S}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$ is the normalized weighted adjacency matrix by dividing the square root of the column sum and the row sum. Two gene relation networks were used with Net-Cox, the gene co-expression network and the gene functional linkage network.

Gene co-expression network. A gene co-expression network was generated from a gene correlation graph model. In the weighted adjacency matrix $\boldsymbol{W}$, each W_{ij} is the reliability score [60] based on the absolute value of the Pearson's correlation coefficients between genes v_i and v_j, calculated as $W_{ij}=\frac{1}{R_{i,j}\times R_{j,i}}$, where $R_{i,j}$ is gene v_i's rank among all the genes with respect to the correlation with gene v_j and $R_{j,i}$ is gene v_j's rank with respect to the correlation with gene v_i. Note that the gene co-expression network is directly inferred from the gene expression dataset. Thus, a gene co-expression network is specific to the dataset used for computing the co-expression network.

Gene functional linkage network. A human gene functional linkage network was constructed by a regularized Bayesian integration system [61]. The network contains maps of functional activity and interaction networks in over 200 areas of human cellular biology with information from 30,000 genome-scale experiments. The functional linkage network summarizes information from a variety of biologically informative perspectives: prediction of protein function and functional modules, cross-talk among biological processes, and association of novel genes and pathways with known genetic disorders [61]. Each edge in the network is weighted between [0,1] to quantify the functional relation between two genes. Thus, the functional linkage network provides much more comprehensive information than Human protein-protein interaction network, which was more frequently used as the network prior knowledge.

Gene expression dataset preparation

Three independent microarray gene expression datasets for studying ovarian carcinoma were used in the experiments [3,24,25]. The information of patient samples in each dataset is given in Table 1. All the three datasets were generated by the Affymetrix HG-U133A platform. The raw .CEL files of two datasets were downloaded from GEO website (Tothill: GSE9899) and (Bonome: GSE26712) [24,25]. The TCGA dataset was downloaded from The Cancer Genome Atlas data portal [3]. The raw files were normalized by RMA [62]. After merging probes by gene symbols and removing probes with no gene symbol, a total of 7562 unique genes were derived from the 22,283 probes and overlapped with the functional linkage network for this study. Note that the Bonome dataset does not provide information on recurrence. Thus, only TCGA and Tothill datasets were used for studying recurrence while all the three datasets were used for studying death. In cross-dataset validation, the batch effects among the three datasets were removed by applying ComBat [63]. Besides testing all the genes, for a better focus on genes that are more likely to be cancer relevant, we derived a set of 2647 genes from the cancer gene list compiled by Sloan-Kettering Cancer Center (SKCC) [64].

The TCGA datasets with AgilentG4502A platform (gene expression array) and HuEx-1_0-st-v2 (exon expression array) were used to evaluate the signature gene FBN1 in Figure 9. The processed level 3 data with expression calls for gene/exon were downloaded from the TCGA data portal.

Cox proportional hazard model

Consider the Cox regression model proposed in [6]. Given $\boldsymbol{X}$, the gene expression profile of n patients over p genes, the instantaneous risk of an event at time t for the i^{th} patient with gene expressions $\boldsymbol{X}_i=(X_{i1},\ldots,X_{ip})'$ is given by

$$h(t|\boldsymbol{X}_i)=h_0(t)exp(\boldsymbol{X}'_i\boldsymbol{\beta}), \qquad (1)$$

where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)'$ is a vector of regression coefficients, and $h_0(t)$ is an unspecified baseline hazard function. In the classical setting with $n>p$, the regression coefficients are estimated by maximizing the Cox's log-partial likelihood:

$$pl(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i\left\{\boldsymbol{X}'_i\boldsymbol{\beta}-log\left[\sum_{j\in R(t_i)}exp(\boldsymbol{X}'_j\boldsymbol{\beta})\right]\right\}, \qquad (2)$$

where t_i is the observed or censored survival time for the i^{th} patient, and δ_i is an indicator of whether the survival time is observed ($\delta_i=1$) or censored ($\delta_i=0$). $\boldsymbol{R}(t_i)$ is the risk set at time t_i, i.e. the set of all patients who still survived prior to time t_i. The commonly used Breslow estimator [65] to estimate the baseline hazard $h_0(t)$ is given by

$$\hat{h}_0(t_i)=1/\sum_{j\in R(t_i)}exp(\boldsymbol{X}'_j\hat{\boldsymbol{\beta}}). \qquad (3)$$

The partial likelihood and the Breslow estimator are induced by the total log-likelihood

$$l(\boldsymbol{\beta},h_0)=\sum_{i=1}^{n}\{-exp(\boldsymbol{X}'_i\boldsymbol{\beta})H_0(t_i)+\delta_i[log(h_0(t_i))+\boldsymbol{X}'_i\boldsymbol{\beta}]\}, \qquad (4)$$

with

$$H_0(t_i)=\sum_{t_k\le t_i}h_0(t_k). \qquad (5)$$

The optimal regression coefficients $\boldsymbol{\beta}$ is estimated based on the maximization of the total log-likelihood by alternating between maximization with respect to $\boldsymbol{\beta}$ (with Newton-Raphson) and $h_0(t)$ (by equation (3)).

In the analysis of microarray gene expressions, the number of gene features p is larger than the number of subjects n by several magnitudes ($p\gg n$). Fitting the Cox regression model will lead to large regression coefficients, which are not reliable. One possible solution is to introduce a $L_2-\text{norm}$ constraint to shrink regression coefficients estimates towards zero [7,10]. In the $L_2-\text{Cox}$ model, the regression coefficients are estimated by maximizing the penalized total log-likelihood:

$$\begin{aligned} l_{pen}(\beta,h_0)=&\sum_{i=1}^{n}\{-exp(\mathrm{X}'_i\beta)H_0(t_i)+\delta_i[log(h_0(t_i))+\mathrm{X}'_i\beta]\}\\ &-\frac{1}{2}\lambda\sum_{j=1}^{p}\beta_j^2, \end{aligned} \qquad (6)$$

where $\lambda\sum_{j=1}^{p}\beta_j^2$ is the penalty term and λ is the parameter controlling the amount of shrinkage. Another possibility is to introduce a $L_1-\text{norm}$ constraint for variable selection [11,13]. The $L_1-\text{Cox}$ model penalizes the log-partial likelihood (equation (2)) by $\lambda\sum_{j=1}^{p}|\beta_j|$ leading to:

$$pl_{pen}(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i\left\{\boldsymbol{X}'_i\boldsymbol{\beta}-log\left[\sum_{j\in R(t_i)}exp(\boldsymbol{X}'_j\boldsymbol{\beta})\right]\right\}-\lambda\sum_{j=1}^{p}|\beta_j|. \qquad (7)$$

In our experiments, R package "glmnet" [66] was used in the implementation of $L_1-\text{Cox}$.

Network-constrained Cox regression (Net-Cox)

We introduce a network-constraint to the Cox model as follows,

$$l_{pen}(\boldsymbol{\beta},h_0)=l(\boldsymbol{\beta},h_0)-\frac{1}{2}\lambda\boldsymbol{\beta}'[(1-\alpha)\boldsymbol{L}+\alpha\boldsymbol{I}]\boldsymbol{\beta}, \qquad (8)$$

where $\boldsymbol{L}$ is a positive semidefinite matrix derived from network information, $\boldsymbol{I}$ is an identity matrix, and λ is the parameter controlling the weighting between the total likelihood and the network constraint. $\alpha\in(0,1]$ is another parameter weighting the network matrix and the identity matrix in the network constraint. For convenience, we define $\boldsymbol{\Gamma}=(1-\alpha)\boldsymbol{L}+\alpha\boldsymbol{I}$ and rewrite the object function as

$$\begin{aligned} l_{pen}(\boldsymbol{\beta},h_0)=&\sum_{i=1}^{n}\{-exp(\mathrm{X}'_i\boldsymbol{\beta})H_0(t_i)+\delta_i[log(h_0(t_i))+\mathrm{X}'_i\beta]\}\\ &-\frac{1}{2}\lambda\beta'\boldsymbol{\Gamma}\boldsymbol{\beta}. \end{aligned} \qquad (9)$$

The term $\lambda\boldsymbol{\beta}'[(1-\alpha)\boldsymbol{L}+\alpha\boldsymbol{I}]\boldsymbol{\beta}$ in equation (8) is a network Laplacian constraint to encode prior knowledge from a network. Given a normalized graph weight matrix $\boldsymbol{S}$, we assume that co-expressed (related) genes should be assigned similar coefficients by defining the following cost term over the coefficients,

$$\begin{aligned} \Psi(\boldsymbol{\beta})=&\frac{1}{2}\sum_{i,j=1}^{p}S_{i,j}(\beta_i-\beta_j)^2\\ =&\boldsymbol{\beta}'(\mathrm{I}-\mathrm{S})\boldsymbol{\beta}=\boldsymbol{\beta}'\mathrm{L}\boldsymbol{\beta}. \end{aligned} \qquad (10)$$

As illustrated in Figure 1, the Laplacian constraint encourages a smoothness among the regression coefficients in the network. Specifically, for any pair of genes connected by an edge, there is a cost proportional to both the difference in the coefficients and the edge weight. Large difference between coefficients on two genes connected with a highly weighted edge will result in a large cost in the objective function. Thus, the objective function encourages assigning similar weights to genes connected by edges of larger weights. By adding an additional $L_2-\mathrm{norm}$ constraint to $\Psi(\boldsymbol{\beta})$ weighted by α, we obtain the network constraint $(1-\alpha)\boldsymbol{\beta}'\boldsymbol{L}\boldsymbol{\beta}+\alpha|\boldsymbol{\beta}|^2=\boldsymbol{\beta}'\boldsymbol{\Gamma}\boldsymbol{\beta}$ in equation (8) and (9). The $L_2-\mathrm{norm}$ of $\boldsymbol{\beta}$ similarly regularizes the uncertainty in the network constraint, which could have a singular Hessian matrix, and the α parameter balances between the $L_2-\mathrm{norm}$ and the "Laplacian-norm". The smaller the α parameter, the more importance put on the network information.

Alternating optimization algorithm

The objective function defined by equation (9) can be solved by alternating optimization of $\boldsymbol{\beta}$ and $h_0(t)$. The maximization with respect to $\boldsymbol{\beta}$ is done by Newton-Raphson method. The derivative of equation (9) is

$$\begin{aligned}\frac{\partial l_{pen}(\boldsymbol{\beta},h_0)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}[\delta_i - exp(\mathrm{X}'_i\boldsymbol{\beta})H_0(t_i)]\mathrm{X}_i - \lambda\boldsymbol{\Gamma}\boldsymbol{\beta} \\ &= \mathrm{X}'\boldsymbol{\Delta} - \lambda\boldsymbol{\Gamma}\boldsymbol{\beta},\end{aligned} \tag{11}$$

where $\Delta_i=\delta_i-exp(\boldsymbol{X}'_i\boldsymbol{\beta})H_0(t_i)$, and the second derivative is

$$\begin{aligned}\frac{\partial^2 l_{pen}(\boldsymbol{\beta},h_0)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= -\left[\sum_{i=1}^{n} exp(\mathrm{X}'_i\boldsymbol{\beta})H_0(t_i)\mathrm{X}_i\mathrm{X}'_i\right] - \lambda\boldsymbol{\Gamma} \\ &= -\mathrm{X}'\mathrm{D}\mathrm{X} - \lambda\boldsymbol{\Gamma},\end{aligned} \tag{12}$$

where $\boldsymbol{D}$ is the diagonal matrix with $D_{ii}=exp(\boldsymbol{X}'_i\boldsymbol{\beta})H_0(t_i)$. Thus, the full algorithm to solve the Net-Cox model is given below.

```
1. **Initialization:** $\boldsymbol{\beta}=0$; Compute $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{S}$.
2. **Do** until convergence
   (a) **Do** Newton-Raphson iteration
       i.   Compute the first derivative $l_{pen}{}'(\boldsymbol{\beta},h_0)=\frac{\partial l_{pen}(\boldsymbol{\beta},h_0)}{\partial \boldsymbol{\beta}}$
       ii.  Compute the second derivative $l''_{pen}(\boldsymbol{\beta},h_0)=\frac{\partial^2 l_{pen}(\boldsymbol{\beta},h_0)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}$
       iii. Update $\boldsymbol{\beta}=\boldsymbol{\beta}-\{l''_{pen}(\boldsymbol{\beta},h_0)\}^{-1}l'_{pen}(\boldsymbol{\beta},h_0)$
   (b) Update $\hat{h}_0(t_i)=1/\sum_{j\in R(t_i)} exp(\boldsymbol{X}'_j\boldsymbol{\beta})$
3. **Return** $\boldsymbol{\beta}$
```

Using Newton-Raphson method to update $\boldsymbol{\beta}$ requires inverting the Hessian matrix, which is time consuming and often inaccurate. An alternative approach is to reduce the covariant space from p to n, which relates to singular value decomposition that exploits the low rank of the gene expression matrix $\boldsymbol{X}$ [10]. The equation

$$\begin{aligned}\frac{\partial l_{pen}(\boldsymbol{\beta},h_0)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}[\delta_i - exp(\mathrm{X}'_i\boldsymbol{\beta})H_0(t_i)]\mathrm{X}_i - \lambda\boldsymbol{\Gamma}\boldsymbol{\beta} \\ &= \mathrm{X}'\boldsymbol{\Delta} - \lambda\boldsymbol{\Gamma}\boldsymbol{\beta} = \boldsymbol{0}\end{aligned} \tag{13}$$

implies that $\boldsymbol{\beta}=\boldsymbol{\Gamma}^{-1}\boldsymbol{X}'\boldsymbol{\eta}$ for some $\boldsymbol{\eta}$. Thus, the dual form of equation (9) with respect to $\boldsymbol{\eta}$ is

$$\begin{aligned}l_{pen}(\boldsymbol{\eta},h_0) =& \sum_{i=1}^{n}\{-exp(\mathrm{Z}'_i\boldsymbol{\eta})H_0(t_i)+\delta_i[log(h_0(t_i))+\mathrm{Z}'_i\boldsymbol{\eta}]\} \\ & -\frac{1}{2}\lambda\boldsymbol{\eta}'\mathrm{Z}\boldsymbol{\eta}\end{aligned} \tag{14}$$

with $\boldsymbol{Z}_i=\boldsymbol{X}\boldsymbol{\Gamma}^{-1}\boldsymbol{X}_i$ and $\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{\Gamma}^{-1}\boldsymbol{X}'$. In its dual form, it is clear that the new object function (14) is equivalent to equation (9) but the problem dimension is reduced from p to n.

Cross validation and parameter tuning

To determine the optimal tuning parameters λ and α, we performed five-fold cross-validation following the procedure proposed by [10] on each of the three datasets. In the cross-validation, four folds of data are used to build a model for validation on the fifth fold, cycling through each of the five folds in turn, and then the (λ,α) pair that maximizes the cross-validation log-partial likelihood (CVPL) are chosen as the optimal parameters. CVPL is defined as

$$CVPL(\lambda,\alpha)=\sum_{i=1}^{5}\left[pl(\hat{\boldsymbol{\beta}}^{(-i)}_{(\lambda,\alpha)})-pl^{(-i)}(\hat{\boldsymbol{\beta}}^{(-i)}_{(\lambda,\alpha)})\right](15)$$

where $\hat{\boldsymbol{\beta}}^{(-i)}$ is the optimal $\boldsymbol{\beta}$ learned from the data without the ith fold. In the equation, $pl()$ denotes the log-partial likelihood on all the samples and $pl^{(-i)}()$ denotes the log-partial likelihood on samples excluding the ith fold. We performed a grid search for the optimal (λ,α) maximizing the sum of the contributions of each fold to the log-partial likelihood in CVPL. In particular, λ was chosen from {1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1} (λs larger than 1 do not change the ranking of $\boldsymbol{\beta}$ anymore), and α was chosen from {0.01,0.1,0.5,0.95}. Note that, when $\alpha=1$, Net-Cox ignores the network information and is reduced to $L_2-\mathrm{Cox}$. For $L_1-\mathrm{Cox}$, the optimal λ was chosen from 1000 λs by the "glmnet" parameter setting with the largest CVPL.

Evaluation measures

The Log-rank test [67] and time-dependent ROC [68] were used to evaluate measurements of the prediction performance by a survival model. For the gene expression profile X in the test set, the prognostic indexes $PI=\boldsymbol{X}'\hat{\boldsymbol{\beta}}$ is computed, where $\hat{\boldsymbol{\beta}}$ is the regression coefficients of the survival model, to rank the patients by descending order. We assigned the top 40% of the patients as the $high-\ risk$ group and the bottom 40% as the $low-risk$ group.

The Log-rank test is a statistical hypothesis test for comparison of two Kaplan-Meier survival curves with the null hypothesis that there is no difference between the population survival curves, i.e. the probability of an event occurring at any time point is the same for each population. The test statistic is compared with a χ^2 distribution with one degree of freedom to derive the significance $p-\mathrm{value}$ reflecting the difference between two survival curves. The log-rank test only evaluates whether the patients are assigned to the "right group" but not how well the patients are ranked within the group by examining the PI. A more refined approach is afforded by the time-dependent ROC curves [68,69]. Time-dependent ROC curves evaluate how well the PI classifies the patients into $high-risk$ and $low-risk$ prognosis groups. Letting $f(\boldsymbol{X})=\boldsymbol{X}'\hat{\boldsymbol{\beta}}$, we can define time-dependent sensitivity and specificity functions at a cutoff point c as

$$sensitivity[c,t|f(\boldsymbol{X})] = Pr\{f(\boldsymbol{X}) > c|\delta(t) = 1\},$$
$$specificity[c,t|f(\boldsymbol{X})] = Pr\{f(\boldsymbol{X}) \leq c|\delta(t) = 0\}$$

with $\delta(t)$ being the event indicator at time t [69]. The corresponding ROC curve for any time t, $ROC[t|f(\boldsymbol{X})]$, is the plot of $sensitivity[c,t|f(\boldsymbol{X})]$ versus $1-specificity[c,t|f(\boldsymbol{X})]$ with different cutoff point c. $AUC[t|f(\boldsymbol{X})]$ is denoted as the area under the $ROC[t|f(\boldsymbol{X})]$ curve. A larger $AUC[t|f(\boldsymbol{X})]$ indicates better prediction of time to event at time t, as measured by sensitivity and specificity evaluated at time t. We plot the AUCs at each time t to compare the methods.

To select gene variables in the multi-variate scenario by Net-Cox and $L_2-\mathrm{Cox}$, we ranked the genes by the magnitude of the coefficients $\boldsymbol{\beta}$. To justify this simple ranking method, we examined the relation between the magnitude of the coefficients for each gene and the contribution of the gene to the log-partial likelihood in Figure S6. It is clear in the plot that the genes towards the two tails of the ranking list contributes most of the likelihood, and the proportion of the contributions are consistent with the ranking. For $L_1-\mathrm{Cox}$, we ranked the genes by the first-time jump into the active set when decreasing the tuning parameter λ in the solution path.

Tumor array preparation

With approval by the Mayo Clinic Institutional Review Board, archived ovarian epithelial tumor specimens from patients with advanced-stage, high-grade serous, or endometrioid tumors obtained prior to exposure to any chemotherapy were utilized to construct the TMA array. The array was constructed using a custom-fabricated device that utilizes a 0.6-mm tissue corer and a 240-capacity recipient block. Triplicate cores from each tumor were included, as were cores of liver as fiducial markers and controls for immunohistochemistry reactions. Five-micrometer-thick sections were cut from the TMA blocks. Immunohistochemistry was performed essentially as described in [70]. Sections of tissue arrays were deparaffinized, rehydrated, and submitted to antigen retrieval by a steamer for 25 minutes in target retrieval solution (Dako, Carpinteria, CA, USA). Endogenous peroxide was diminished with 3% H_2O_2 for 30 min. Slides were blocked in protein block solution for 30 min and then blocked with avidin and biotin for 10 min each, followed by overnight incubation with 1:1000 diluted Anti-FBN1 antibody (HPA021057, Sigma-Aldrich) at 4°C. The sections were then incubated with biotinylated universal link for 15 min and streptavidin for 25 min at 25°C. Slides were developed in diaminobenzine and counterstained with hematoxylin.

Supporting Information

Figure S1 Time-dependent AUCs averaged across the five test folds in five-fold cross-validation. The plots report the results of using Sloan-Kettering cancer genes (left column) and all mappables genes (right column). The plots show the results for the death outcome of TCGA dataset (A), the death outcome of Tothill dataset (B), the death outcome of Bonome dataset (C), the tumor recurrence outcome of TCGA dataset (D) and the tumor recurrence outcome of Tothill dataset (E).
(PDF)

Figure S2 Marker gene consistency (all mappable genes). The x-axis is the number of selected signature genes ranked by each method. The y-axis is the percentage of the overlapped genes between the selected genes across the ovarian cancer datasets. The results are shown for the death outcome (A) and the tumor recurrence outcome (B).
(PDF)

Figure S3 Cross-dataset survival prediction (all mappable genes). The first four columns of plots show the Kaplan-Meier survival curves for the two risk groups defined by Net-Cox (co-expression network), Net-Cox (functional linkage network), $L_2-\mathrm{Cox}$ and $L_1-\mathrm{Cox}$. The fifth column of plots compare the time-dependent area under the ROC curves based on the estimated risk scores (PIs). The results are shown for the death outcome by training with TCGA dataset and test on Tothill Dataset (A), for the death outcome by training with TCGA dataset and test on Bonome Dataset (B) and for the tumor recurrence outcome by training with TCGA dataset and test on Tothill Dataset (C).
(PDF)

Figure S4 Protein-Protein interaction sub-networks of marker genes identified by Net-Cox and $L_2-\mathrm{Cox}$ on the TCGA dataset. (A) The PPI subnetworks identified by Net-Cox on the co-expression network. (B) The PPI subnetworks identified by Net-Cox on the functional linkage network. (C) The PPI subnetwrks identified by $L_2-\mathrm{Cox}$.
(PDF)

Figure S5 Kaplan-Meier survival plots on FBN1 expression groups. (A) Kaplan-Meier survival curve of recurrence by FBN1 staining groups. The group with low FBN1 expression has a lower recurrence rate compared with the groups with high expression after 12 month of treatment. (B) Kaplan-Meier survival curve of recurrence by the expression of FBN1 on Tothill dataset. (C)–(E) Kaplan-Meier survival curves of recurrence by the expression of FBN1 on TCGA dataset with AgilentG4502A platform, HuEx-1_0-st-v2 platform, and Affymetrix HG-U133A platform, respectively. In plots(B)–(E), the patients are divided into two groups of the same size by the expression of FBN1.
(PDF)

Figure S6 Contributions to the log-partial likelihood by each individual gene by Net-Cox on the Tothill dataset (Sloan-Kettering cancer genes). The x-axis is the index of the genes sorted by coefficients.
(PDF)

Table S1 Optimal parameters of Net-Cox. The parameters are selected by CVPLs in five-fold cross-validation. (a) Sloan-Kettering cancer genes. (b) All mappable genes.
(PDF)

Table S2 Statistical significance of the improvement in time-dependent AUCs in cross-dataset evaluation (Sloan-Kettering cancer genes). The R package "timeROC" (the algorithm was described in the paper "Estimating and Comparing time-dependent areas under ROC curves for censored event times with competing risks") was used to compute the $p-\mathrm{values}$. The null hypothesis asserts that two time-dependent AUCs estimated by two models are equal. The significant $p-\mathrm{values}$ smaller than 0.1 are bold. The tables show the results for the death outcome by training with TCGA dataset and test on Tothill Dataset (a), for the death outcome by training with TCGA dataset and test on Bonome Dataset (b), for the tumor recurrence outcome by training with TCGA dataset and test on Tothill Dataset (c).
(PDF)

Table S3 Cross validation partial likelihood (CVPL) in five-fold cross-validation (Sloan-Kettering cancer

genes). (a) The death outcome of TCGA dataset. (b) The tumor recurrence outcome of TCGA dataset. (c) The death outcome of Tothill dataset. (d) The tumor recurrence outcome of Tothill dataset. (e) The death outcome of Bonome dataset. (f) L_1-Cox. (g) Group Lasso and Sparse-Group Lasso.
(PDF)

Table S4 Log-rank test $p-$values of the test folds on five-fold cross-validation. The most significant $p-$values across four models with cut-off 0.05 are bold. (a) Sloan-Kettering cancer genes and the death outcome. (b) Sloan-Kettering cancer genes and the tumor recurrence outcome. (c) All mappable genes and the death outcome. (d) All mappable genes and the tumor recurrence outcome.
(PDF)

Table S5 Log-rank test $p-$values in cross-dataset evaluation (all mappable genes). The survival prediction performance on Tothill and Bonome datasets using the Cox models trained with TCGA dataset are reported.
(PDF)

Table S6 Enriched GO terms by the signature genes of death outcome. The $p-$values in $-log10$ scale are shown for the enriched GO terms.
(PDF)

Table S7 Enriched GO terms by the signature genes of recurrence. The $p-$values in $-log10$ scale are shown for the enriched GO terms. A "X" denotes a $p-$value larger than 0.01.
(PDF)

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions for improving the experimental design and the presentation of the paper.

Author Contributions

Conceived and designed the experiments: WZ VS JC BW RK. Performed the experiments: WZ TO. Analyzed the data: WZ JC BW RK. Contributed reagents/materials/analysis tools: TO VS JC. Wrote the paper: WZ JC BW RK.

References

1. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, et al. (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.
2. Bøvelstad HM, Nygård S, Størvold HL, Aldrin M, Borgan Ø, et al. (2007) Predicting survival from microarray data-a comparative study. Bioinformatics 23: 2080–2087.
3. Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615.
4. Van Wieringen W, Kun D, Hampel R, Boulesteix A (2009) Survival prediction using gene expression data: a review and comparison. Computational statistics & data analysis 53: 1590–1603.
5. Witten D, Tibshirani R (2010) Survival analysis with high-dimensional covariates. Stat Methods Med Res 19: 29–51.
6. Cox DR (1972) Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological) 34: 187–220.
7. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12: 55–67.
8. Pawitan Y, Bjohle J, Wedren S, Humphreys K, Skoog L, et al. (2004) Gene expression profiling for prognosis using cox regression. Stat Med 23: 1767–1780.
9. Hastie T, Tibshirani R (2004) Efficient quadratic regularization for expression arrays. Biostatistics 5: 329–340.
10. Van Houwelingen HC, Bruinsma T, Hart AAM, Van't Veer LJ, Wessels LFA (2006) Cross-validated cox regression on microarray gene expression data. Statistics in Medicine 25: 3201–3216.
11. Tibshirani R (1997) The lasso method for variable selection in the cox model. Stat Med 16: 385–395.
12. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. The Annals of statistics 32: 407–499.
13. Gui J, 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: 3001–3008.
14. Segal MR (2006) Microarray gene expression data with linked survival phenotypes: diffuse large-b-cell lymphoma revisited. Biostatistics 7: 268–285.
15. Park M, Hastie T (2007) L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69: 659–677.
16. Sohn I, Kim J, Jung SH, Park C (2009) Gradient lasso for cox proportional hazards model. Bioinformatics 25: 1775–1781.
17. Li H, Luan Y (2003) Kernel cox regression models for linking gene expression profiles to censored survival data. Pac Symp Biocomput : 65–76.
18. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140.
19. Li C, Li H (2008) Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24: 1175–1182.
20. Hwang T, Sicotte H, Tian Z, Wu B, Kocher J, et al. (2008) Robust and efficient identification of biomarkers by classifying features on graphs. Bioinformatics 24: 2023–2029.
21. Tian Z, Hwang T, Kuang R (2009) A hypergraph-based learning algorithm for classifying gene expression and arraycgh data with prior knowledge. Bioinformatics 25: 2831–2838.
22. Vandin F, Upfal E, Raphael B (2011) Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol 18: 507–522.
23. Kim Y, Wuchty S, Przytycka T (2011) Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput Biol 7: e1001095.
24. Tothill RW, Tinker AV, George J, Brown R, Fox SB, et al. (2008) Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clinical Cancer Research 14: 5198–5208.
25. Bonome T, Levine DA, Shih J, Randonovich M, Pise-Masison CA, et al. (2008) A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer research 68: 5478–5486.
26. Crijns APG, Fehrmann RSN, Jong SD, Gerbens F, Meersma GJ, et al. (2009) Survival-related profile, pathways, and transcription factors in ovarian cancer. PLoS Med 6: e1000024.
27. Hatzirodos N, Bayne RA, Irving-Rodgers HF, Hummitzsch K, Sabatier L, et al. (2011) Linkage of regulators of tgf-β activity in the fetal ovary to polycystic ovary syndrome. FASEB J 25: 2256–2265.
28. Ghosh S, Albitar L, LeBaron R, Welch WR, Samimi G, et al. (2010) Up-regulation of stromal versican expression in advanced stage serous ovarian cancer. Gynecologic oncology 119: 114–120.
29. Yiu GK, Chan WY, Ng SW, Chan PS, Cheung KK, et al. (2001) Sparc (secreted protein acidic and rich in cysteine) induces apoptosis in ovarian cancer cells. The American journal of pathology 159: 609–622.
30. Xian L, He W, Pang F, Hu Y (2012) Adipoq gene polymorphisms and susceptibility to polycystic ovary syndrome: a huge survey and meta-analysis. European Journal of Obstetrics & Gynecology and Reproductive Biology 161: 117–124.
31. Gery S, Xie D, Yin D, Gabra H, Miller C, et al. (2005) Ovarian carcinomas: Ccn genes are aberrantly expressed and ccn1 promotes proliferation of these cells. Clinical Cancer Research 11: 7243–7254.
32. Adam M, Saller S, Ströbl S, Hennebold J, Dissen G, et al. (2012) Decorin is a part of the ovarian extracellular matrix in primates and may act as a signaling molecule. Human Reproduction 27: 3249–3258.
33. Rocconi RP, Kirby TO, Seitz RS, Beck R, Straughn Jr JM, et al. (2008) Lipoxygenase pathway receptor expression in ovarian cancer. Reproductive Sciences 15: 321–326.
34. Bridges PJ, Jo M, Al Alem L, Na G, Su W, et al. (2010) Production and binding of endothelin-2 (edn2) in the rat ovary: endothelin receptor subtype a (ednra)-mediated contraction. Reproduction, Fertility and Development 22: 780–787.
35. Sutphen R, Xu Y, Wilbanks GD, Fiorica J, Grendys Jr EC, et al. (2004) Lysophospholipids are potential biomarkers of ovarian cancer. Cancer Epidemiology Biomarkers & Prevention 13: 1185–1191.
36. Mok SC, Bonome T, Vathipadiekal V, Bell A, Johnson ME, et al. (2009) A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2. Cancer cell 16: 521–532.
37. Sengupta S, Kim KS, Berk MP, Oates R, Escobar P, et al. (2006) Lysophosphatidic acid downregulates tissue inhibitor of metalloproteinases, which are negatively involved in lysophosphatidic acid-induced cell invasion. Oncogene 26: 2894–2901.
38. Czekierdowski A, Czekierdowska S, Danilos J, Czuba B, Sodowski K, et al. (2008) Microvessel density and cpg island methylation of thbs2 gene in malignant ovarian tumors. J Physiol Pharmacol 59: 53–65.
39. Kryczek I, Lange A, Mottram P, Alvarez X, Cheng P, et al. (2005) Cxcl12 and vascular endothelial growth factor synergistically induce neoangiogenesis in human ovarian cancers. Cancer research 65: 465–472.
40. Geles KG, Freiman RN, Liu WL, Zheng S, Voronina E, et al. (2006) Cell-type-selective induction of c-jun by taf4b directs ovarian-specific transcription networks. Proc Natl Acad Sci U S A 103: 2594–2599.

41. Richards JS, Russell DL, Ochsner S, Hsieh M, Doyle KH, et al. (2002) Novel signaling pathways that control ovarian follicular development, ovulation, and luteinization. Recent Prog Horm Res 57: 195–220.
42. Levina V, Nolen BM, Marrangoni AM, Cheng P, Marks JR, et al. (2009) Role of eotaxin-1 signaling in ovarian cancer. Clinical Cancer Research 15: 2647–2656.
43. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13: 2363–2371.
44. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature Protocols 4: 44–57.
45. Sood AK, Coffin JE, Schneider GB, Fletcher MS, DeYoung BR, et al. (2004) Biological significance of focal adhesion kinase in ovarian cancer: role in migration and invasion. Am J Pathol 165: 1087–1095.
46. Allen HJ, Sucato D, Woynarowska B, Gottstine S, Sharma A, et al. (1990) Role of galaptin in ovarian carcinoma adhesion to extracellular matrix in vitro. J Cell Biochem 43: 43–57.
47. Yang X, Kovalenko OV, Tang W, Claas C, Stipp CS, et al. (2004) Palmitoylation supports assembly and function of integrintetraspanin complexes. The Journal of Cell Biology 167: 1231–1240.
48. Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. PLoS Biology 2: e108.
49. Bair E, Hastie T, Paul D, Tibshirani R (2006) Prediction by supervised principal components. Journal of the American Statistical Association 101: 119–137.
50. Li L, Li H (2004) Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics 20: 3406–3412.
51. Nguyen D, Rocke D (2002) Partial least squares proportional hazard regression for application to dna microarray survival data. Bioinformatics 18: 1625–1632.
52. Bastien P (2004) Pls-cox model: application to gene expression. Proceedings in Computational Statistics: 655–662.
53. Bastien P, Vinzi V, Tenenhaus M (2005) Pls generalised linear regression. Computational Statistics & Data Analysis 48: 17–46.
54. Boulesteix A, Strimmer K (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in bioinformatics 8: 32–44.
55. Hothorn T, Lausen B, Benner A, Radespiel-Troger M (2004) Bagging survival trees. Stat Med 23: 77–91.
56. Hothorn T, Buhlmann P, Dudoit S, Molinaro A, van der Laan M (2006) Survival ensembles. Biostatistics 7: 355–373.
57. Ma S, Song X, Huang J (2007) Supervised group lasso with applications to microarray data analysis. BMC Bioinformatics 8: 60.
58. Simon N, Friedman J, Hastie T, Tibshirani R (2012) A sparse-group lasso. Journal of Computational and Graphical Statistics DOI 10: 681250
59. Sherman-Baust C, Weeraratna A, Rangel L, Pizer E, Cho K, et al. (2003) Remodeling of the extracellular matrix through overexpression of collagen vi contributes to cisplatin resistance in ovarian cancer cells. Cancer Cell 3: 377–386.
60. Ucar D, Neuhaus I, Ross-MacDonald P, Tilford C, Parthasarathy S, et al. (2007) Construction of a reference gene association network from multiple profiling data: application to data analysis. Bioinformatics 23: 2716–2724.
61. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, et al. (2009) Exploring the human genome with functional maps. Genome Research 19: 1093–1106.
62. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264.
63. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics 8: 118–127.
64. Higgins ME, Claremont M, Major JE, Sander C, Lash AE (2007) Cancergenes: a gene selection resource for cancer genome projects. Nucleic Acids Research 35: D721–D726.
65. Breslow NE (1972) Discussion of professor cox's paper. J R Statist Soc : 216–217.
66. Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization paths for cox's proportional hazards model via coordinate descent. Journal of Statistical Software 39: 1–13.
67. Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer chemotherapy reports 50: 163–170.
68. Heagerty PJ, Lumley T, Pepe MS (2000) Time-dependent roc curves for censored survival data and a diagnostic marker. Biometrics 56: 337–344.
69. Li H, Gui J (2004) Partial cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics 20: i208–i215.
70. Chien J, Aletti G, Baldi A, Catalano V, Muretto P, et al. (2006) Serine protease htra1 modulates chemotherapy-induced cytotoxicity. J Clin Invest 116: 1994–2004.
71. Stronach EA, Sellar GC, Blenkiron C, Rabiasz GJ, Taylor KJ, et al. (2003) Identification of clinically relevant genes on chromosome 11 in a functional model of ovarian cancer tumor suppression. Cancer research 63: 8648–8655.
72. Körner M, Waser B, Reubi JC (2003) Neuropeptide y receptor expression in human primary ovarian neoplasms. Laboratory investigation 84: 71–80.