Gong et al. BMC Systems Biology 2014, 8(Suppl 1):S3
http://www.biomedcentral.com/1752-0509/8/S1/S3

BMC Systems Biology

PROCEEDINGS Open Access

Pathway-gene identification for pancreatic cancer survival via doubly regularized Cox regression

Haijun Gong[1*†], Tong Tong Wu[2*†], Edmund M Clarke[3]

From The Twelfth Asia Pacific Bioinformatics Conference (APBC 2014)
Shanghai, China. 17-19 January 2014

Abstract

Background: Recent global genomic analyses identified 69 gene sets and 12 core signaling pathways genetically altered in pancreatic cancer, which is a highly malignant disease. A comprehensive understanding of the genetic signatures and signaling pathways that are directly correlated to pancreatic cancer survival will help cancer researchers to develop effective multi-gene targeted, personalized therapies for the pancreatic cancer patients at different stages. A previous work that applied a LASSO penalized regression method, which only considered individual genetic effects, identified 12 genes associated with pancreatic cancer survival.

Results: In this work, we integrate pathway information into pancreatic cancer survival analysis. We introduce and apply a doubly regularized Cox regression model to identify both genes and signaling pathways related to pancreatic cancer survival.

Conclusions: Four signaling pathways, including Ion transport, immune phagocytosis, TGFβ (spermatogenesis), regulation of DNA-dependent transcription pathways, and 15 genes within the four pathways are identified and verified to be directly correlated to pancreatic cancer survival. Our findings can help cancer researchers design new strategies for the early detection and diagnosis of pancreatic cancer.

Background

Pancreatic cancer [1] is a devastating disease with a very poor prognosis and a five-year survival rate around 3-5%. The most common form of pancreatic cancer is the pancreatic ductal adenocarcinoma (PDAC, a malignant exocrine cancer). In the past 30 years, no substantial progress has been made in PDAC diagnosis and treatment. New techniques and methods to investigate the dynamics of PDAC are urgently needed. Modern microarray technology has revolutionized the way that we study the complex biological systems, allowing pancreatic cancer researchers to make genome-wide expression profiling and measure other features for patients in a fast, precise, and cost-effective way. One aim of systems biologists is to correctly decipher and interpret the high-dimensional complex gene expression data, that is, to identify the key genetic signatures and signaling pathways implicated in the diseases.

Pancreatic cancer is characterized by rapid growth, early local and distant invasion, interactions with stromal cells (e.g., pancreatic stellate cells) [2] and fibrous tissue, and a high resistance to chemotherapy and radiotherapy. The evolution of pancreatic cancer is partially stimulated by the overexpression of several growth factors, cytokines, and genetic alterations [3,4] at different stages of PDAC. Recent global genomic analyses identified 69 gene sets and 12 core signaling pathways genetically altered in the pancreatic cancer [1]. Most of the previous genomic analyses and microarray studies focused on the identification of the differentially expressed and metastasis-associated genes at different stages of pancreatic cancer [3,5], ignoring an important clinical factor - survival time. Stratford et al.'s work identified six genetic signatures [6] associated with metastatic pancreatic cancer using a sequence of statistical

* Correspondence: hgong2@slu.edu; tongtong_wu@urmc.rochester.edu
† Contributed equally
[1]Department of Mathematics and Computer Science, Saint Louis University, Saint Louis, MO 63103, USA
[2]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA
Full list of author information is available at the end of the article

techniques, including the significance analysis of microarray (SAM) [7], centroid-based predictor [8], Pearson correlation, X-Tile [9], Kaplan-Meier estimator [10] and Cox model [11]. Though these genes could help discriminate high- and low-risk patients, the prediction was not based on survival time. A comprehensive understanding of the genetic signatures and signaling pathways that are directly correlated to pancreatic cancer survival will help cancer researchers to develop effective multi-gene targeted, personalized therapies for the pancreatic cancer patients at different stages and improve survival rate.

The Cox proportional hazards model [11] is the most popular survival model used to describe the relationship between the patient's survival time and predictor variables. When we have high-dimensional data (e.g. in a microarray study) where the number of predictors (genes) far exceeds the number of subjects (patients), the Cox model cannot be fitted directly unless the high-dimensionality is properly handled. The regularization approach has been widely used to select important variables from a large pool of candidate variables [12-14]. For example, a Lasso (least absolute shrinkage and selection operator) penalty can be imposed to individual variables to automatically remove unimportant ones by shrinking their regression coefficients to be exactly zero [15]. In our previous work [16], we applied a lasso penalized Cox regression method, for the first time, to investigate the signature genes that are correlated to the pancreatic cancer survival time. We identified 12 genes associated with the pancreatic cancer survival and eight of them have been confirmed to be genetically altered and differentially expressed in the cancer of gastric, colorectal, ovarian, breast, skin, kidney, colon, lung, and pancreatic in *in vivo* and *in vitro* experiments [17-25]. It has been shown that these survival-associated genes can also help to grade the stage and estimate the survival time of the PDAC patients.

However, the genes may perform as groups rather than individuals since some genes belong to the same pathways and get involved in the same biological processes. The pathway information is biologically important to our understanding of gene regulatory networks and cancer development [1]. The previous work [16] performs gene selection based on the strength of individual genes solely and ignores the information of signaling pathways. Recently, several variable selection methods have been introduced to consider group effects. For example, the group lasso method penalizes the L_2-norm (Euclidean norm) of the coefficients within each group in linear regression [26] and Cox proportional hazards model [12]. Based on the boosting technique, a group additive regression model [27] and a nonparametric pathway-based regression model [28] were developed to identify groups of genomic features that are related to several clinical phenotypes, including the survival outcome. However, those group selection methods only conduct "group selection" without "within-group selection", since they select variables in an "all-in-or-all-out" fashion. That is, if one variable in a group is selected, all the other variables in the same group will also get selected.

Although pathways as a whole are involved in the development of pancreatic cancer, according to the global genomic analyses, not all the genes in the same pathway are involved in the process. In this work, we employ a doubly regularized Cox (DrCox) regression model [29] that integrates both genes and signaling pathways for the pancreatic cancer survival analysis. Both non-overlap and overlap cases of DrCox are considered. Cyclic coordinate descent algorithms are derived for parameter estimation. We analyze the high-dimensional microarray data of pancreatic cancer patients with localized and resected PDAC collected between 1999 and 2007 [6] using DrCox. Four signaling pathways, including Ion transport, immune phagocytosis, TGFβ (spermatogenesis), regulation of DNA-dependent transcription pathways, and 15 genes within these four pathways are identified and verified to be directly correlated to pancreatic cancer survival. Compared with other methods, the DrCox model can provide more accurate and useful prediction of survival time [29]. These findings can help cancer researchers design new strategies for the early detection and diagnosis of pancreatic cancer at different stages.

Methods

In this section, we describe the doubly regularized Cox (DrCox) regression and derive the parameter estimates via cyclic coordinate descent algorithms. We first present the case where the groups do not overlap, i.e., each variable belongs to only one group. Then we discuss the overlap case, i.e., variables are allowed to belong to multiple groups.

Doubly regularized Cox (DrCox) regression for non-overlap cases

Assume that the p variables (genes) occur in K groups (pathways). We further assume the kth group has p_k variables and denote the p_k variables in the kth group by $\boldsymbol{X}_{(k)} = (X_{k1}, \ldots, X_{kpk})^T$, with the corresponding regression coefficients $\boldsymbol{\beta}_{(k)} = (\beta_{k1}, \ldots, \beta_{kpk})^T$. For a sample of n subjects, let T_i and C_i denote the survival time and the censoring time for subject $i = 1, \ldots, n$. The observed survival time is defined by $Y_i = \min\{T_i, C_i\}$ and the censoring indicator is $\delta_i = \boldsymbol{I}(T_i \le C_i)$. The p predictor variables of the ith subject is denoted by $X_i = (X_{i,(1)}^T, \ldots, X_{i,(K)}^T)^T$, where $X_{i,(k)} = (X_{i,k1}, \ldots, X_{i,kp_k})^T$. The survival time T_i and the censoring time C_i are conditionally independent given $\boldsymbol{X}_i$. The censoring mechanism is assumed to be noninformative. The observed data can be represented by the triplets $\{(Y_i, \delta_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$.

The Cox proportional hazards model [11] composed of p genes and K pathways is written by

$$h(t|X) = h_0(t)\exp\left(\sum_{k=1}^{K}\sum_{j=1}^{p_k}\beta_{kj}X_{kj}\right) = h_0(t)\exp\left(\sum_{k=1}^{K}X_{i,(k)}^T\beta_{(k)}\right),$$

where $\sum_{k=1}^{K} p_k = p$. The partial likelihood of the Cox model is

$$L_n(\beta) = \prod_{i\in D}\frac{\exp\left(\sum_{k=1}^{K}X_{i,(k)}^T\beta_{(k)}\right)}{\sum_{l\in R_i}\exp\left(\sum_{k=1}^{K}X_{l,(k)}\beta_{(k)}\right)},$$

where D is the set of indices of observed failures, and R_i is the set of indices of the subjects who are at risk at time Y_i.

To achieve the goal of both group and within-group variable selection and to overcome the non-convexity drawback, the doubly regularized Cox regression model imposes a mixture of lasso penalty and group lasso penalty to the log-partial likelihood $\ell_n(\boldsymbol{\beta}) = \log L_n(\boldsymbol{\beta})$

$$\begin{aligned} g(\beta) &= -\ell_n(\beta) + \lambda_1\sum_{k=1}^{K}\sum_{j=1}^{p_k}|\beta_{kj}| + \lambda_2\sum_{k=1}^{K}\sqrt{\sum_{j=1}^{p_k}\beta_{kj}^2} \\ &= -\ell_n(\beta) + \lambda_1\sum_{k=1}^{K}\|\beta_{(k)}\|_1 + \lambda_2\sum_{k=1}^{K}\|\beta_{(k)}\|_2, \end{aligned} \quad (1)$$

where $\|\beta_{(k)}\|_1 = \sum_{j=1}^{px}|\beta_{kj}|$ is the lasso penalty on individual parameters, $\|\beta_{(k)}\|_2 = \sqrt{\sum_{j=1}^{p_k}\beta_{kj}^2}$ is the group penalty on groups of parameters, and λ_1 and λ_2 are two nonnegative tuning constants controlling the strength of variables selection. The larger are the tuning constants, the fewer variables are retained in the model. In this paper, the value of the tuning constants are determined using k-fold cross validation (data-driven) technique to select a subset of relevant genes and signaling pathways for accurate and robust prediction.

Coordinate descent for non-overlap cases

Since there are more predictor variables than subjects ($p > n$), to tackle the high-dimensionality problem we use a cyclic coordinate descent algorithm, which has been shown to be computationally efficient [30-33]. The idea is to break a large optimization problem into a sequence of small ones. In other words, instead of estimating all the parameters at the same time, we can update each parameter one by one. Readers can refer to [31,32] for more details.

In the non-overlap case, where each variable belongs to only one group, estimation of parameters and selection of important variables can be conducted via the minimization of (1) iteratively w.r.t. one parameter by one parameter. The first step is to calculate the forward and backward directional derivatives of each parameter. If e_{kj} is the coordinate direction along which β_{kj} varies, then the forward and backward directional derivatives of β_{kj} are

$$\begin{aligned} d_{e_{kj}}g(\beta) &= \lim_{t\downarrow 0}\frac{g(\beta + te_{kj}) - g(\beta)}{t} \\ &= -\frac{\partial}{\partial\beta_{kj}}\ell_n(\beta) + \begin{cases} (\lambda_1+\lambda_2)(-1)^{I(\beta_{kj}<0)} & \text{if } \|\beta_{(k)}\|_2 = 0 \\ \lambda_1(-1)^{I(\beta_{kj}<0)} + \lambda_2\dfrac{\beta_{kj}}{\|\beta_{(k)}\|_2} & \text{if } \|\beta_{(k)}\|_2 > 0, \end{cases} \end{aligned}$$

and

$$\begin{aligned} d_{-e_{kj}}g(\beta) &= \lim_{t\downarrow 0}\frac{g(\beta - te_{kj}) - g(\beta)}{t} \\ &= \frac{\partial}{\partial\beta_{kj}}\ell_n(\beta) + \begin{cases} (\lambda_1+\lambda_2)(-1)^{I(\beta_{kj}>0)} & \text{if } \|\beta_{(k)}\|_2 = 0 \\ \lambda_1(-1)^{I(\beta_{kj}>0)} - \lambda_2\dfrac{\beta_{kj}}{\|\beta_{(k)}\|_2} & \text{if } \|\beta_{(k)}\|_2 > 0, \end{cases} \end{aligned}$$

where $I(\cdot)$ is an indicator function equal to 1 if the condition in the parentheses is satisfied and 0 otherwise, and

$$\frac{\partial}{\partial\beta_{kj}}\ell_n(\beta) = \sum_{i\in D}\left\{x_{i,kj} - \frac{\sum_{l\in R_i}\exp\left(\sum_{k=1}^{K}X_{l,(k)}^T\beta_{(k)}\right)x_{l,kj}}{\sum_{l\in R_i}\exp\left(\sum_{k=1}^{K}X_{l,(k)}^T\beta_{(k)}\right)}\right\}.$$

After obtaining the directional derivatives, we then need to decide which parameters to be updated and the direction for updating. If both of the directional derivatives $d_{e_{kj}}g(\beta)$ and $d_{-e_{kj}}g(\beta)$ are nonnegative, then the update for β_{kj} is skipped. If either directional derivative is negative, then we solve for the minimum along the corresponding direction. It is impossible for both directional derivatives to be negative due to the convexity of $g(\boldsymbol{\beta})$. After identifying the direction to update the parameter, one can use Newton's method to solve for the minimum. The update at iteration $m + 1$ is given by

$$\beta_{kj}^{m+1} = \beta_{kj}^{m} + \frac{-\frac{\partial}{\partial\beta_{kj}}\ell_n(\beta^m) + \lambda_1(-1)^{I(\beta_{kj}^m<0)}}{\frac{\partial^2}{\partial\beta_{kj}^2}\ell_n(\beta^m)} + \frac{\lambda_2\{(-1)^{I(\beta_{kj}^m<0)}I_1(\beta_{(k)}^m) + \frac{\beta_{kj}^m}{\|\beta_{(k)}^m\|_2}I_2(\beta_{(k)}^m)\}}{\frac{\partial^2}{\partial\beta_{kj}^2}\ell_n(\beta^m)}$$

where $\boldsymbol{\beta}m$ is the estimate at iteration m, $I_1(\cdot) = I(\|\cdot\|_2 = 0)$, and $I_2(\cdot) = I(\|\cdot\|_2 > 0)$.

DrCox regression via coordinate descent for overlap cases

However, in reality, one gene can get involved in different pathways. To consider overlapping, we modify the notation and objective function (1). We denote the p variables by $X_1, \ldots, X_p$ with the corresponding regression coefficients $\beta_1, \ldots, \beta_p$. Let $V_k \subseteq \{1, 2, \ldots, p\}$ be the set of indices of variables in the kth group. The objective function designed for the overlap case can be written as

$$g(\beta) = -\ell_n(\beta) + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{k=1}^{K}\sqrt{\sum_{j\in V_k}\beta_j^2}. \quad (2)$$

Note that predictor X_j can belong to several pathways but it is only associated with one coefficient β_j.

The parameter estimation needs to be modified accordingly. If we consider the coordinate direction e_j for β_j, the forward and backward directional derivatives of β_j are

$$d_{e_j}g(\beta) = \lim_{t\downarrow 0}\frac{g(\beta+te_j)-g(\beta)}{t}$$
$$= -d_{e_j}\ell_n(\beta)+\lambda_1(-1)^{I(\beta_j<0)}+\lambda_2\sum_{k\in G_j}\left\{(-1)^{I(\beta_j<0)}I_1(\beta_{(k)})+\frac{\beta_j}{\|\beta_{(k)}\|_2}I_2(\beta_{(k)})\right\}$$

and

$$d_{-e_j}g(\beta) = \lim_{t\downarrow 0}\frac{g(\beta-te_j)-g(\beta)}{t}$$
$$= -d_{-e_j}\ell_n(\beta)+\lambda_1(-1)^{I(\beta_j>0)}+\lambda_2\sum_{k\in G_j}\left\{(-1)^{I(\beta_j>0)}I_1(\beta_{(k)})-\frac{\beta_j}{\|\beta_{(k)}\|_2}I_2(\beta_{(k)})\right\}.$$

where $G_j \subseteq \{1, 2, \ldots, K\}$ are the indices of groups that X_j belongs to. After determining the direction for updating, the coefficient can be updated by

$$\beta_j^{m+1} = \beta_j^m + \frac{-\frac{\partial}{\partial\beta_j}\ell_n(\beta^m)+\lambda_1(-1)^{I(\beta_j^m<0)}}{\frac{\partial^2}{\partial\beta_j^2}\ell_n(\beta^m)} + \frac{\lambda_2\sum_{k\in G_j}\left\{(-1)^{I(\beta_j^m<0)}I_1(\beta_{(k)}^m)+\frac{\beta_j^m}{\|\beta_{(k)}^m\|_2}I_2(\beta_{(k)}^m)\right\}}{\frac{\partial^2}{\partial\beta_j^2}\ell_n(\beta^m)}.$$

Results and discussion

The DrCox model with the cyclic coordinate descent algorithm is applied to analyze the PDAC data collected between 1999 and 2007. The aim of this work is to identify core signaling pathway sets and genetic signatures within those pathways related to pancreatic cancer survival. The microarray data of pancreatic cancer include 102 samples [6], which are publicly available at Gene Expression Omnibus (access code 21501). According to [6], among these 102 PDAC patients, 66 died at the end of the study (censoring rate 35%). The survival time ranges from 1 month to 5 years. The Kaplan-Meier curve is plotted in Figure 1 to show the probability of survival in 5 years for the 102 PDAC patients. Each step means an actual event happens, i.e. a pancreatic cancer patient dies. A short vertical line without a drop means a patient gets censored for different reasons, drops off the study or the study ends. Additionally, two stage variables, T stage and N stage, are given to describe the stages of pancreatic cancer, where T stage describes the size of the primary tumor ranging from 1 to 4 and N stage describes the spread to nearby (regional) lymph nodes with values 0 or 1.

The whole dataset is randomly split into the training, validation, and testing sets with equal sizes. The training set is used for model fitting, and the validation set is used for tuning constants selection. Using the 3-fold cross-validation, we got the optimal values of $\lambda_1 = 0.3$ and $\lambda_2 = 0.1$, which minimize the log-partial likelihood function. Figure 2 shows the 3-D plots of the log-partial likelihood function and the number of selected genes vs. (λ_1, λ_2), respectively. Under the optimal tuning constants, 4 pathways and 15 genes are selected from the pool of 12660 probes of 6910 genes in 130 pathway sets organized in [1], which belong to 15 core groups in the pancreatic cancer studies. The selected pathways include the pathways of "regulation of DNA-dependent transcription" (6 out of 2096 genes are selected), "Ion transport" (7 out of 555 genes are selected), "immune phagocytosis" (1 out of 215 genes is selected), and "TGFβ (spermatogenesis)" (1 out of 268 genes is selected) pathways. These identified pathways and genes are biologically meaningful and consistent with the existing scientific findings. In particular, three genes - ZNF233, SLC22A8, and PCYT1B - were identified in the previous work [16] using a Lasso penalized Cox model when considering gene signatures only.

• **Regulation of DNA-dependent transcription pathway** is well-known to be related to the development of cancer. It regulates the frequency and rate of cellular DNA-dependent transcription. This work identified three families of six genes that are related to pancreatic cancer survival. The six genes are DENND4A, KLF13, ZNF229, ZNF233, ZNF395, and ZNF432.

- **DENND4A** is a c-myc promoter-binding protein [34], which mediates signal transduction in the nucleus and regulate the DNA replication and transcription. DENND4A can also activate the RAB10 protein, which is a key regulator of polarized sorting in epithelial cells, from an inactive GDP-bound form to an active GTP-bound form through promoting GDP $\rightarrow$ GTP exchange.
- **KLF13** belongs to the KLF family of transcription factors for several oncogenes and tumor suppressor genes [35,36] and it plays an important role in the tumor progression [36]. Recent study shows that KLF13 is overexpressed in the oral cancer cells. Inhibiting KLF13's expression can decrease the proliferation of cancer cell and increase its sensitivity to ionizing radiation [36]. In pancreatic cancer, KLF13 can suppress the cell growth and neoplastic transformation mediated by K-RAS, which is mutated in more than 90% of pancreatic tumors [35]. Our work suggests that KLF13 may be a useful biomarker for early detection and possible targets for the pancreatic cancer therapy.
- **Zinc finger protein family members: ZNF229, ZNF233, ZNF395 and ZNF432** are DNA-binding protein domains consisting of zinc fingers. Many of these zinc finger proteins, including ZNF233 (also identified in the previous work [16]), have been found to be associated with the abnormality of chromosome 19 in the studies of kidney [23] and pancreatic

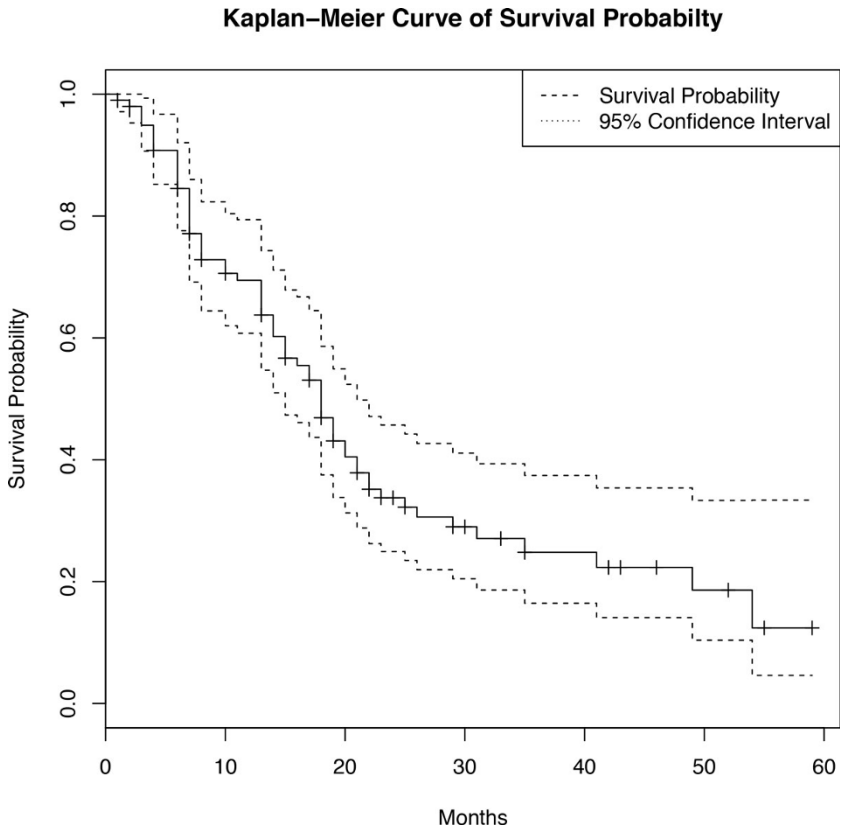

Figure 1 Kaplan-Meier curve of survival probability and 95% confidence interval for 102 PDAC patients The Kaplan-Meier survival curve (solid line) describes the probability of survival for the 102 PDAC patients. The dashed lines represent the 95% confidence interval. The horizontal axis represents the survival time (in months). Each step means an actual event happens, i.e. a pancreatic cancer patient dies. A short vertical line without a drop means a patient get censored for different reasons, e.g. drops off the study or the study ends.

cancers [1]. Our analysis reveals that Zinc finger proteins and the corresponding pathway might be associated with the survival of pancreatic cancer.

• **Ion transport pathway** plays integral roles in the development of cancer. Since the plasma membrane ion channels contribute to all basic cellular process [37,38], many ion channels are implicated in the uncontrolled proliferation, decreased apoptosis, and unorganized angiogenesis. According to [37], the ion channels also contribute to the six hallmarks of cancer [39]: "1) self-sufficiency in growth signals, 2) insensitivity to antigrowth signals, 3) evasion of programmed cell death (apoptosis), 4) limitless replicative potential, 5) sustained angiogenesis and 6) tissue invasion and metastasis."

We identified seven genes from three different channels or families, including the TRP channel (TRPV5 and TRPM6) regulating the transcellular Ca^{2+} transport, KCNK channel (KCNK3 and KCNK18) regulating the K^{+} transport, and solute carrier (SLC) family (SLC22A8, SLC8A3, and SLC24A6). Recent experimental studies have indicated that these three families play important roles in the cancer development.

- **TRP (Ca^{2}+) channel and TRPV5, TRPM6 genes** regulate the Calcium-mediated signal transduction that is frequently altered in cancer [40]. Several genes in TRPV channel have been detected to be up-regulated in prostate, colon, and breast cancer cells [40-42]. Particularly, TRPV5 and TRPV6 genes exhibit unusually high levels of single nucleotide polymorphisms (SNPs) in African populations as compared to other populations [41]. Moreover, the genes TRPM6 and TRPM7 in the TRPM channel can enhance the secretion of angiogenic factors, for example VEGF [40], resulting in a sustained unorganized angiogenesis process. The TRP channel and TRPV5, TRPM6 genes identified in pancreatic cancer survival data could be possible targets for the future cancer diagnosis and treatment.

- **KCNK (K+) channel and KCNK3, KCNK18 genes** regulate the potassium (K+) transport and

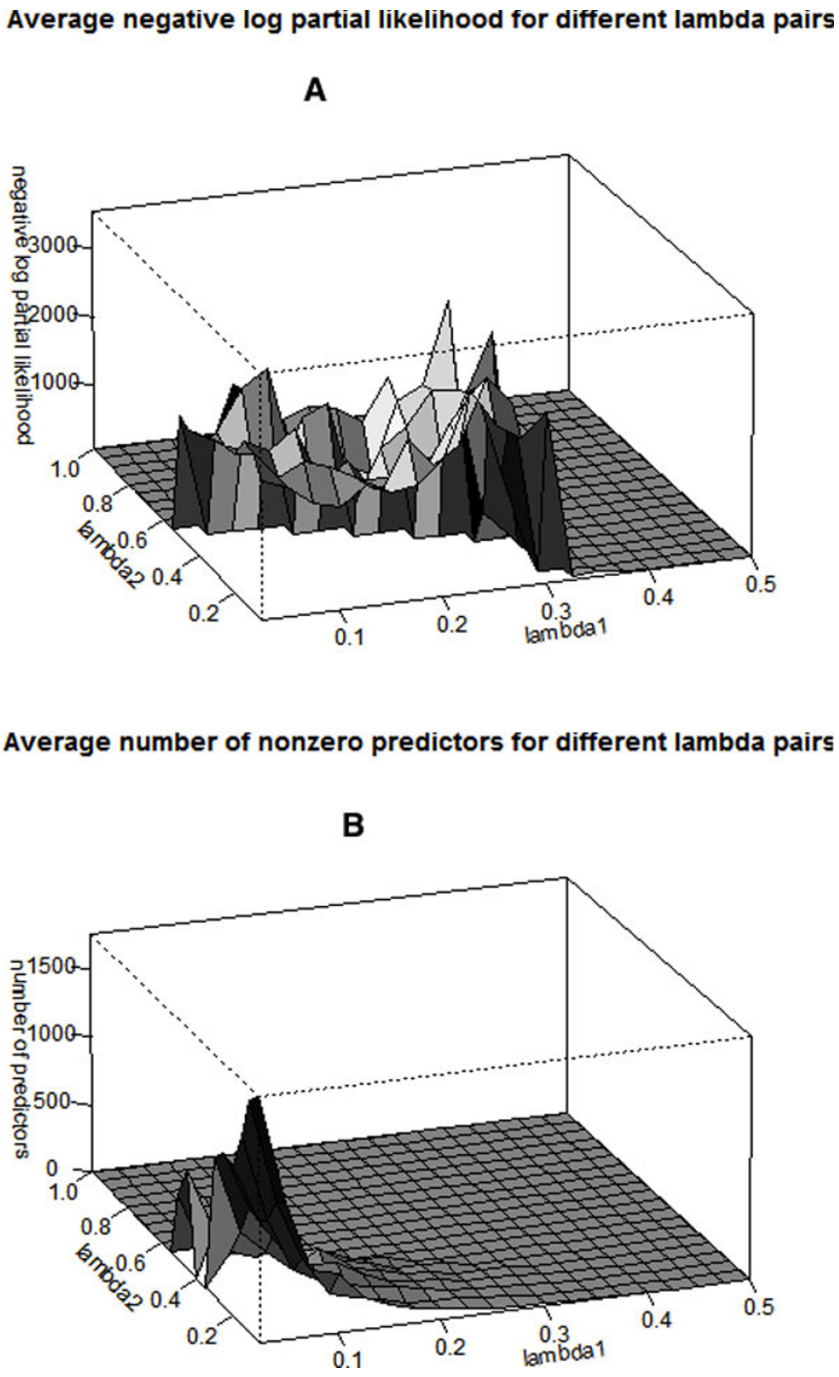

Figure 2 3-fold cross validation for tuning constants (λ_1, λ_2). (A) shows the 3-D plot of the log-partial likelihood function vs. (λ_1, λ_2), and the optimal values are $\lambda_1 = 0.3$ and $\lambda_2 = 0.1$; (B) shows the 3-D plot of the number of selected genes with nonzero regression coefficients vs. (λ_1, λ_2).

membrane potential (Vm) in response to different physical and chemical factors [38,40]. Several KCNK channel genes, for example, KCNK9 [43], are overexpressed in breast and lung cancers, and the gene KCNK2 can promote prostate cancer cell's growth [40,44].

- **SLC family:** SLC22A8, SLC8A3, SLC24A6 are membrane transport proteins that are involved in the transport and excretion of many organic ions, drugs and toxicants. Some genes in SLC family are cancer-related, for example, SLC43A2 whose overexpression is associated with the adenocarcinomas and

squamous cell carcinoma [45], which was identified in the previous work [16].

• **Immune phagocytosis pathway and CYBA gene:** One prominent hallmark feature of cancer is the evasion of immune destruction [39]. The immune system is important in preventing tumor initiation and controlling tumor growth through identifying and eliminating the cancer cells (i.e., tumor immune surveillance) [46]. Macrophages and other phagocytic cells are important players in the innate immune system whose functions include phagocytosis (homeostatic cell clearance), antigen presentation (pathogen defense), and cytokine production (inflammatory responses). Recent evidence [46-48] revealed that the active immune phagocytosis pathway could inhibit tumor growth through phagocytic clearance, i.e., programmed cell removal in clearing damaged and foreign cells. The CYBA gene is a tumor suppressor [49], which regulates the immune system cells - phagocytes, involved in autophagy. The phagocytosis and superoxide production is primary regulated by the cytochrome b- 245, (light) alpha subunit (also known as $p22^{phox}$), which is encoded by the gene CYBA. CYBA's mutation will cause the failure of phagocytosis and immune defects [50]. This observation supports our prediction that the immune phagocytosis and tumor suppressor gene CYBA might be associated with pancreatic cancer survival and tumor immune evasion. Targeting this pathway might lead to effective cancer immunotherapies.

• **TGFβ core pathway (spermatogenesis signaling set) and PCYT1B gene**: The transforming growth factor beta (TGFβ) signaling pathway is critical in regulating many cellular processes, including the cell growth, differentiation and apoptosis. It has genetic alterations in 100% of pancreatic cancers [1]. The gene PCYT1B (phosphate cytidylyl transferase 1 choline β) was identified to be associated with pancreatic cancer survival, which is consistent with the previous work [16]. The expression of PCYT1B is frequently deregulated in cancer cells of epithelial ovarian [21], high grade gliomas [51], and pancreatic ductal adenocarcinoma [22]. Moreover, PCYT1B is a key regulator in the choline phospholipid metabolism, which is altered in the cancers of breast [19], colon [20], ovarian [21], and gliomas [51]. These observations support our prediction that PCYT1B and TGFβ pathway are correlated with pancreatic cancer survival and they might help to grade the stage of pancreatic cancer patients.

Compared with the previous work [16], which selected 12 survival-relevant genes using a Lasso penalized Cox model, the DrCox model identified 4 pathways and 15 genes related to pancreatic cancer survival. We divide the patients into long- and short-survival groups based on the selected pathways and genes and conduct the logrank test to compare the two groups. The survival probabilities of these two groups are plotted in the Figure 3. The logrank test gives a p-value of 0.0179, which means the two groups can be well separated and our finding of 4 pathways and 15 genes is significant.

Conclusions

In this work, we employed the doubly regularized Cox (DrCox) regression coupled with the coordinate descent algorithm to analyze the high-dimensional gene expression data of patients with localized and resected PDAC. Different from the previous work [16], this DrCox model can incorporate both gene and pathway information and simultaneously infer genetic signatures and important signaling pathways that are related to the pancreatic cancer survival. The proposed cyclic coordinate descent algorithm can quickly remove irrelevant genes and signaling pathways, so the prediction of survival time is more accurate and robust than other methods. Other group selection models select variables in an "all-in-or-all-out" fashion with no within-group selection, that is, if one variable in a group (pathway) is selected, all the other variables in the same group will get selected. For example, if gene PCYT1B in the TGFβ pathway is selected, all the rest of genes in the TGFβ pathway will be selected as well. However, not all the genes in the TGFβ pathway are involved in the development of pancreatic cancer. The advantage of our DrCox method is that it can conduct both group selection and within-group selection simultaneously and eliminate the irrelevant.

This work identified four signaling pathways, including Ion transport, immune phagocytosis, TGFβ (spermatogenesis), regulation of DNA-dependent transcription pathways, and 15 genes within these four pathways, which are directly correlated to pancreatic cancer survival. Pancreatic cancer patients with these deregulated signaling pathways and mutated genes might have a shorter survival time. Several inferred signaling components have been confirmed to be altered frequently in the cancer of pancreatic, oral, prostate, colon, breast and lung in the *in vivo* or *in vitro* experiments. Our finding predicts that, the TRP (Ca^2+) channel-related genes (TRPV5 and TRPM6) and KCNK (K+) channel-related genes in the ion transport pathway are possible biomarkers of pancreatic cancer survival. The Immune phagocytosis pathway with the tumor suppressor CYBA gene, which regulates the immune system cells and autophagy through phagocytic clearance, have not received enough attention in the existing pancreatic cancer research literature. The gene PCYT1B in the TGFβ pathway is frequently deregulated in cancer cells compared with normal cells, which might help to grade the stage of pancreatic cancer patients. The KLF13 in the regulation of

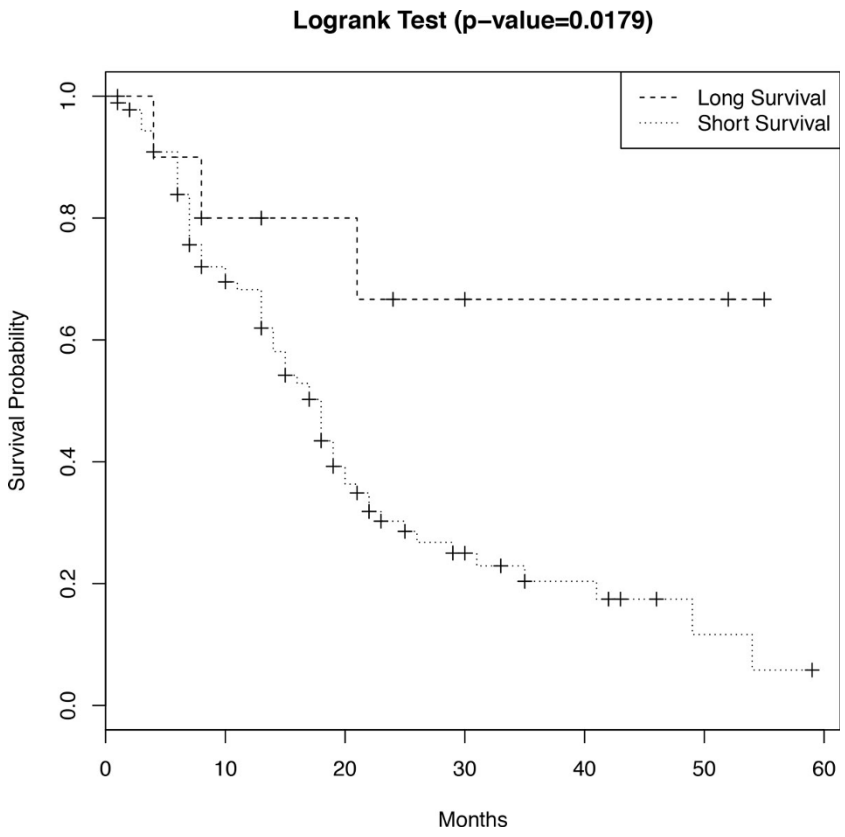

Figure 3 Logrank test of the long- and short-survival groups based on the 4 pathways and 15 genes (p-value = 0.0179). The 102 PDAC patients are divided into long- and short-survival groups based on the 4 pathways and 15 genes. The survival probabilities of these two groups are compared using the logrank test. The p-value of 0.0179 means the two groups are well separated and our finding of 4 pathways and 15 genes is significant.

DNA-dependent transcription pathway could regulate the cell growth through regulating KRAS pathway. These findings demonstrate that these survival-associated genetic signatures and pathways could be useful biomarkers for early cancer detection and diagnosis and help pancreatic cancer researchers to grade the cancer stage and select appropriate therapies to prolong the patient's survival time at different stages.

This work is the first attempt to infer the pancreatic cancer survival-associated signaling pathway sets and genetic signatures within those pathways using statistical techniques. However, any statistical findings need to be tested by the further clinical and wet lab experiments of pancreatic cancer. We are unable to test our results with other independent datasets in this paper due to the data source limitation. We do expect our results can get verified or falsified by further investigation. We hope the genetic signatures and pathways found in this paper could help cancer researchers design new strategies for the early detection and diagnosis and lead to effective treatments and immunotherapies for pancreatic cancer.

Competing interests
The authors declare that there are no competing interests.

Authors' contributions
HG and TTW are joint first authors and both authors contributed equally. HG, TTW, and EMC proposed the study, TTW prepared the computational code, HG and TTW analyzed the results and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
HG was supported by the new faculty start-up grant from the Saint Louis University, TTW and EMC were supported by NSF grant CCF-0926181.

Declarations
Publication of this article was funded by the Saint Louis University to HG and NSF grant CCF-0926181 to TTW.
This article has been published as part of *BMC Systems Biology* Volume 8 Supplement 1, 2014: Selected articles from the Twelfth Asia Pacific Bioinformatics Conference (APBC 2014): Systems Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/8/S1.

Authors' details
[1]Department of Mathematics and Computer Science, Saint Louis University, Saint Louis, MO 63103, USA. [2]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA. [3]Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Published: 24 January 2014

References

1. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW: **Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses.** *Science* 2008, **321**:1801-1806.
2. Vonlaufen A, Joshi S, et al.: **Pancreatic Stellate Cells: Partners in Crime with Pancreatic Cancer Cells.** *Cancer Res* 2008, **68**:2085-2093.
3. Bardeesy N, DePinho RA: **Pancreatic cancer biology and genetics.** *Nature Reviews Cancer* 2002, **2(12)**:897-909.
4. Maitra A, Hruban R: **Pancreatic Cancer.** *Annu Rev Pathol Mech Dis* 2008, **3**:157-188.
5. Friess H, Ding J, Kleeff J, *et al*: **Microarray-based identification of differentially expressed growth- and metastasis-associated genes in pancreatic cancer.** *Cell Mol Life Science* 2003, **60**:1180-1199.
6. Stratford JK, Bentrem DJ, Anderson JM, *et al*: **A Six-Gene Signature Predicts Survival of Patients with Localized Pancreatic Ductal Adenocarcinoma.** *PLoS Med* 2010, **7**:e1000307.
7. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci* 2001, **98**:5116-5121.
8. Hu Z, Fan C, Oh DS, Marron JS, He X, *et al*: **The molecular portraits of breast tumors are conserved across microarray platforms.** *BMC Genomics* 2006, **7**:96.
9. Camp RL, Dolled-Filhart M, Rimm DL: **X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization.** *Clin Cancer Res* 2004, **10**:7252-7259.
10. Kaplan EL, Meier P: **Nonparametric estimation from incomplete observations.** *Journal of the American Statistical Association* 1958, **53**:457-481.
11. Cox DR: **Regression models and life-tables (with discussion).** *J R Statist Soc B* 1972, **34**:187-220.
12. Ma S, Song X, Huang J: **Supervised group Lasso with applications to microarray data analysis.** *BMC Bioinformatics* 2007, **8**:60-76.
13. Sohn I, Kim J, Jung SH, Park C: **Gradient lasso for Cox proportional hazards model.** *Bioinformatics* 2009, **25**:1775-1781.
14. Tibshirani R: **Univariate Shrinkage in the Cox Model for High Dimensional Data.** *Stat Appl in Genet and Mol Biol* 2009, **8**:21.
15. Tibshirani R: **The lasso method for variable selection in the Cox model.** *Statist Med* 1997, **16**:385-395.
16. Wu TT, Gong H, Clarke EM: **A Transcriptome Analysis by Lasso Penalized Cox Regression for Pancreatic Cancer Survival.** *Journal of Bioinformatics and Computational Biology* 2011, **9**:63.
17. Shi Y, Zhai H, Wang X, *et al*: **Ribosomal proteins S13 and L23 promote multidrug resistance in gastric cancer cells by suppressing drug-induced apoptosis.** *Exp Cell Res* 2004, **296**:337-46.
18. Denis M, Chadeneau C, *et al*: **Over-expression of the S13 ribosomal protein in actively growing cells.** *Int J Cancer* 1993, **55**:275-280.
19. Katz-Brull R, Seger D, Rivenson-Segal D, Rushkin E, Degani H: **Metabolic markers of breast cancer: enhanced choline metabolism and reduced choline-ether-phospholipid synthesis.** *Cancer Res* 2002, **62**:1966-70.
20. Mori N, Delsite R, Natarajan K, Kulawiec M, Bhujwalla Z, Singh K: **Loss of p53 function in colon cancer cells results in increased phosphocholine and total choline.** *Mol Imaging* 2004, **3**:319-23.
21. Iorio E, Ricci A, Bagnoli M, *et al*: **Activation of phosphatidylcholine-cycle enzymes in human epithelial ovarian cancer cells.** *Cancer Res* 2010, **70**:2126-2135.
22. von Forstner C, Egberts JH, Ammerpohl O, *et al*: **Gene Expression Patterns and Tumor Uptake of 18F- FDG, 18F-FLT, and 18F-FEC in PET/MRI of an Orthotopic Mouse Xenotransplantation Model of Pancreatic Cancer.** *J Nucl Med August* 2008, **49**:1362-1370.
23. Koeman J, Russell R, *et al*: **Somatic Pairing of Chromosome 19 in Renal Oncocytoma is Associated with Deregulated ELGN2-Mediated Oxygen-Sensing Response.** *PLoS Genetics* 2008, **4**:e1000176.
24. Cuezva J, Chen G, Alonso A, Isidoro A, Misek D, Hanash S, Beer D: **The bioenergetic signature of lung adenocarcinomas is a molecular marker of cancer diagnosis and prognosis.** *Carcinogenesis* 2004, **25**:1157-63.
25. Hernlund E, Hjerpe E, Avall-Lundqvist E, Shoshan M: **Ovarian carcinoma cells with low levels of beta-F1-ATPase are sensitive to combined platinum and 2- deoxy-D-glucose treatment.** *Mol Cancer Ther* 2009, **8**:1916-23.
26. Yuan M, Lin Y: **Model selection and estimation in regression with grouped variables.** *J R Statist Soc B* 2006, **68**:49-67.
27. Luan Y, Li H: **Group additive regression models for genomic data analysis.** *Biostatistics* 2008, **9**:100-113.
28. Wei Z, Li H: **Nonparametric pathway-based regression models for analysis of genomic data.** *Biostatistics* 2007, **8**:265-284.
29. Wu TT, Wang S: **Doubly Regularized Cox Regression for High-dimensional Survival Data with Group Structures.** *Statisstics and Its Interface* 2013, **6**:175-186.
30. Fu WJ: **Penalized regressions: the bridge versus the lasso.** *J Comp and Graph Stat* 1998, **7**:397-416.
31. Friedman J, Hastie T, Hoefling H, Tibshirani R: **Pathwise coordinate optimization.** *Ann Appl Stat* 2007, **1**:302-332.
32. Wu TT, Lange K: **Coordinate descent algorithms for lasso penalized regression.** *Ann Appl Stat* 2008, **2**:224-244.
33. Wu TT, Chen YF, Hastie T, Sobel E, Lange K: **Genomewide association analysis by lasso penalized logistic regression.** *Bioinformatics* 2009, **25**:714-721.
34. Ray R, Miller D: **Cloning and characterization of a Human c-myc promoter-binding protein.** *Molecular and Cellular Biology* 1991, 2154-2161.
35. Fernandez-Zapico ME, Billadeau DD, Urrutia R: **Klf13 Suppresses the Transforming Activity of K-Ras By Direct Downregulation the Cyclin B Gene.** *Pancreas* 2004, **29**:360.
36. Henson B, Gollin S: **Overexpression of KLF13 and FGFR3 in oral cancer cells.** *Cytogenet Genome Res* 2010, **128**:192-198.
37. Prevarskaya N, Skryma R, Shuba Y: **Ion channels and the hallmarks of cancer.** *Trends in Molecular Medicine* 2010, **16**:107-121.
38. Goldstein S, Bockenhauer D, O'Kelly I, Zilberberg N: **Potassium leak channels and the KCNK family of two-p-domain subunits.** *Nature Reviews Neuroscience* 2001, **2**:175-184.
39. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100**:57-70.
40. Bodding M: **TRP proteins and cancer.** *Cell Signal* 2007, **19**:617-624.
41. Peng J: **TRPV5 and TRPV6 in transcellular Ca(2+) transport: regulation, gene duplication, and polymorphisms in African populations.** *Adv Exp Med Biol* 2011, **704**:239-75.
42. Dong H, *et al*: **Molecular mechanisms underlying Ca2+-mediated motility of human pancreatic duct cells.** *Am J Physiol Cell Physiol* 2010, **299(6)**: C1493-503.
43. Mu D, *et al*: **Genomic amplification and oncogenic properties of the KCNK9 potassium channel gene.** *Cancer Cell* 2003, **3**:297-302.
44. Voloshyna I, *et al*: **TREK-1 is a novel molecular target in prostate cancer.** *Cancer Res* 2008, **68**:1197-1203.
45. Haase C, Bergmann R, Fuechtner F, Hoepping A, Pietzsch J: **L-type amino acid transporters LAT1 and LAT4 in cancer: uptake of 3-O-methyl-6-18F-fluoro-L-dopa in human adenocarcinoma and squamous cell carcinoma in vitro and in vivo.** *J Nucl Med* 2007, **48**:2063-2071.
46. Jaiswal S, MP Chao RM, Weissman I: **Macrophages as mediators of tumor immunosurveillance.** *Trends Immunol* 2010, **31**:212-219.
47. Chao MP, W I, Majeti R: **Programmed cell removal: A new obstacle in the road to developing cancer.** *Nat Rev Cancer* 2011, **12(1)**:58-67.
48. Chao MP, Weissman IL, Majeti R: **The CD47-SIRPalpha Pathway in Cancer Immune Evasion and Potential Therapeutic Implications.** *Curr Opin Immunol* 2012, **24**:225-232.
49. Powell JA, *et al*: **Sequencing, Transcript Identification, and Quantitative Gene Expression Profiling in the Breast Cancer Loss of Heterozygosity Region 16q24.3 Reveal Three Potential Tumor-Suppressor Genes.** *Genomics* 2002, **80**:303-310.
50. Nakano Y, Longo-Guess CM, Bergstrom DE, *et al*: **Mutation of the Cyba gene encoding p22phox causes vestibular and immune defects in mice.** *J Clin Invest* 2008, **118**:1176-1185.
51. Righi V, Roda J, Paz J, *et al*: **1H HR-MAS and genomic analysis of human tumor biopsies discriminate between high and low grade astrocytomas.** *NMR in Biomedicine* 2009, **22**:629-637.

doi:10.1186/1752-0509-8-S1-S3
Cite this article as: Gong *et al.*: Pathway-gene identification for pancreatic cancer survival via doubly regularized Cox regression. *BMC Systems Biology* 2014 **8**(Suppl 1):S3.