Genomics 97 (2011) 350–357

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Prediction of cancer prognosis with the genetic basis of transcriptional variations

Hyojung Paik [a,b], Eunjung Lee [c], Inho Park [d], Junho Kim [a], Doheon Lee [a,*]

[a] Department of Bio and Brain Engineering, KAIST, 335 Gwahangno, Yuseong-gu, Daejeon, 305–701, Republic of Korea
[b] Plant Systems Engineering Center, KRIBB, 111 Gwahangno, Yuseong-gu, Daejeon, 305–806, Republic of Korea
[c] Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[d] Healthcare Service Group, Incubation Center, Samsung SDS, Seoul 135–798, Republic of Korea

ARTICLE INFO

Article history:
Received 6 October 2010
Accepted 9 March 2011
Available online 16 March 2011

Keywords:
Genetic architecture
Genotype
Transcriptional variation
Survival prediction

ABSTRACT

Phenotypes of diseases, including prognosis, are likely to have complex etiologies and be derived from interactive mechanisms, including genetic and protein interactions. Many computational methods have been used to predict survival outcomes without explicitly identifying interactive effects, such as the genetic basis for transcriptional variations. We have therefore proposed a classification method based on the interaction between genotype and transcriptional expression features (CORE-F). This method considers the overall "genetic architecture," referring to genetically based transcriptional alterations that influence prognosis.

In comparing the performance of CORE-F with the ensemble tree, the best-performing method predicting patient survival, we found that CORE-F outperformed the ensemble tree (mean AUC, 0.85 vs. 0.72). Moreover, the trained associations in the CORE-F successfully identified the genetic mechanisms underlying survival outcomes at the interaction-network level. Details of the learning algorithm are available in the online supplementary materials located at http://www.biosoft.kaist.ac.kr/coref.

1. Introduction

The architecture that underlies complex phenotypes consists of interactions among genetic, transcriptomic, and proteomic features, such as SNPs, and the expression of genes and proteins. Exploration of epigenetic and eQTL (expression quantitative loci) interaction networks is increasingly recognized as an important approach to understanding disease phenotypes. The enumeration of network components by identifying molecular-level interactions may enhance predictions of disease outcome [1,2] and cancer metastasis [3]. A computational method that can model the interactions between genetic and transcriptional signatures may help uncover genetic mechanisms and may assist in more precisely diagnosing and predicting prognosis of patients.

Various types of possible predictors of disease phenotypes have been suggested, including expression of proteins [4] and alternative splicing events [5]. Particularly in genetics, the diagnosis of disease and the prediction of its progression have been based on genome-wide association studies. These studies have identified genetic variations associated with various diseases, including heart disease [6] and diabetes [7], although experimental support for the functional roles of the identified loci is incomplete in many cases. The genetic architecture modeling approach has demonstrated that genetic variations regulate gene expression, both directly and indirectly, lead to phenotype variations associated with disease [8–10], including response to drugs [11]. These results have shown that a classification algorithm that incorporates a model of genetic architecture (i.e. a map of genetic causality of transcriptional signatures and phenotype variations) may assist in predicting patient prognosis.

Generally, tree-based approaches have been utilized to predict patient survival outcomes; for example, the CART (classification and regression tree) [12], random survival forest [13], and ensemble tree [14] methods. Since the hierarchical structure of trees assumes recursive splits at each node, this method may be inefficient in detecting interactions within training data [15]. Tree-based approaches have therefore had limited success in training interaction models that include epistatic effects in SNP–SNP interactions [16]. Although analysis of genetic architecture is a promising route to understanding disease phenotypes and underlying mechanisms [17,18], tree-based survival prediction methods have been insufficient for the learning of genetic interactions that contribute to complex traits [19], such as the prognosis of cancer patients. Thus, the development of a classification method for disease survival outcome that includes consideration of the underlying genetic architecture remains a challenge in genetics and bioinformatics.

We have developed a novel classification method, based on the interaction between genotype and transcriptional expression features (CORE-F), to predict the prognosis of cancer patients. This method learns the associations among genotypes and transcriptional signatures in models of "genetic architecture," that is, maps of the genetic variations that drive phenotypic variations via interactions with transcriptional alterations [8]. The developed kernel function in the

[*] Corresponding author. Fax: +82 42 350 8680.
E-mail address: dhlee@kaist.ac.kr (D. Lee).

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.03.005

CORE-F algorithm was designed to predict the class label of survival phenotypes based on genetic architecture consisting of interactions between SNPs and expression signatures. The present study consists of three parts: 1) the continuous labeling of survival outcomes, 2) selection of survival associated features, genotypes and expression signatures via statistical analyses, and 3) building and validating the training results of the CORE-F algorithm. We utilized a set of tissue samples from patients with ovarian cancer, obtained from The Cancer Genome Atlas (TCGA) consortium [20], to validate the performance of CORE-F relative to that of the ensemble tree [14] method, which has outperformed other tree-based approaches without overfitting. To identify loci and expression signatures associated with survival, we prepared four sets of expression signatures and SNPs in prognosis-related pathways (part 2). Supervised learning of the CORE-F algorithm (part 3) was accomplished by dividing the continuous label of survival outcomes into two classes, true (poor outcomes) and false (good outcomes), using a cut-off parameter. For the modeling of genetic architectures, genetic associations among gene expression levels and survival classes were determined by training (part 3) and were introduced into the designed kernel function of CORE-F. CORE-F outperformed the ensemble tree, with a 13% performance improvement of AUC, indicating that, by focusing on genetic architectures, we successfully developed novel frameworks for predicting survival outcomes.

2. Results

We developed a method for the supervised classification of survival outcomes based on "genetic architecture," consisting of SNPs and expression signatures associated with trait variations (survival outcomes) via their interactions. Survival outcomes of cancer samples might be presented with unbounded scale depending on the heterogeneity and censoring of survivorship. In this regard, we utilized a developed membership function (μ), which graded survival times on a bounded scale [0, 1] (Fig. 1a). Sets of genotypes and expression signatures belonging to various prognosis-related pathways were analyzed for survival-associated feature selections, respectively. After the statistical analysis (Fig. 1b and c), sets of genotypes at the j-th locus (g_i^j) and transcriptional signatures in the k-th gene (e_i^k) were prepared for each of the samples i. Using a splitted learning set, interactions between genetic variations and transcriptional signatures for the true class of survival time (poor survival outcomes) were trained (Fig. 1d). Using the trained interactions between genotypes and expression signatures, we utilized the kernel function of the CORE-F algorithm (σ), a designed fractional function, to measure the scores that predicted a true class (poor survival outcome) while considering the model of genetic architecture (Fig. 1e). The CORE-F was then used to determine the true class using the kernel function-derived value and the decision parameter, θ ($\sigma > \theta \rightarrow 1$) (Fig. 1e). In summary, statistical analyses were used to prepare different sets of genotypes and expression signatures for each of the corresponding pathways, and trained interactions between SNPs and expressions within learning sets were routinely tested for the building of CORE-F.

2.1. Class labeling for disease survival outcomes

Because typical approaches utilize proportional hazard ratios to determine the relative risk for patients [7], classification of samples according to individual survival time is widely desired. The membership function we developed transformed survival times and censorings of samples into graded class labels bounded by [0, 1], allowing classification of samples by individual prognoses. The observed survival outcomes were classified by two cut-off parameters: P_b for poor survival outcome groups and P_g for good survival outcome groups ($0 < P_b < P_g$) (Fig. 2a). However, censoring of the survival times and intermediate survival outcomes ($P_b < \text{survival time} < P_g$) hampered efforts to label the survival times with scalar values (Fig. 2a, b). Therefore, the membership function from fuzzy set theory [21] was used to transform the survival outcomes into class labels. In particular, a cosine function was used to grade the class labels for survival times with respect to survival time distribution curves (Fig. 2c). The uncertainty in survival time for early censored cases, in which a short period of survivorship was determined relative to the censoring event, was also modeled by a membership function. The forked line in Fig. 2d denotes the labeling uncertainty of the early censored survival times, with a larger value of the survival class comparing early deceased cases without censoring. Based on the fuzzy set theory prediction of the class label degrees, the correlation coefficient (0.95) between survival times and class labels indicated that the membership function developed here appropriately labeled the survival times (Fig. 2d). Due to the cutoff period for studies of survival in patients with ovarian cancer, the thresholds for poor and good survival outcomes were 1 and 5 years, respectively ($P_b = 1$ year, $P_g = 5$ year) [22]. The purpose of the membership function was to provide a gradient labeling scale for survivorship bounded by 0 and 1 in a continuous manner. Therefore, the labels for prognoses were introduced to select survivorship associated features for the training of CORE-F, and were also utilized to determine true class (poor survival outcomes) for the supervised learning of CORE-F.

2.2. Selecting SNPs and expression signatures in prognosis-related pathways

Before the training of interactions between SNPs and expression signatures, we selected survival associated SNPs and expression signatures due to the lack of prognostic loci and expression signatures with replicated results. Expression signatures and SNPs in four drug-related pathways were analyzed as markers associated with survival outcomes in patients receiving therapy for ovarian cancer, such as drug treatment [23,24]. The drug-related pathways analyzed were platinum-related, taxane-related, membrane transport ABC, and cytochrome P450 pathways (Table 1). Member genes and corresponding SNPs in the selected pathways were identified by the method of Huang et al [25]. Using the resources of PharmGKB [26] and KEGG [27], we identified 139 genes and 2746 corresponding SNPs. Table 1 presents results of our analyses using two-step statistical approaches (Fig. 1b and c). Since the object of this statistical approach was to identify features associated with survival and CORE-F profiles of the interactions between SNPs and gene expression in the training set, we explored several SNPs and expression signatures associated with survival in the entire set of samples. We identified 72 genetic variations (SNPs) and 10 expression signatures as features significantly associated with the class label of survival times (Table 1 in bold characters). In the subsequent training step, four sets of interactions between features in different pathways were trained independently to build the CORE-F and ensemble tree routine depending on the corresponding pathways. For example, using a set of SNPs and alterations in expression in the membrane transport ABC pathway, CORE-F learned interactions among 10 SNPs and two sets of expression. In summary, our use of 1) a small number of survival associated features in different pathways, 2) independent selection of survival associated features without considering inter-relationships, and 3) learning of interactions between SNPs and gene expression for each training set indicate that CORE-F trained without overfitting.

2.3. Building and validation of CORE-F

CORE-F was constructed using the trained interactions between SNPs and expression signatures in the selected pathways, as described in Methods section. The power of CORE-F was tested by comparing its performance with that of the ensemble tree, which guarantees effective

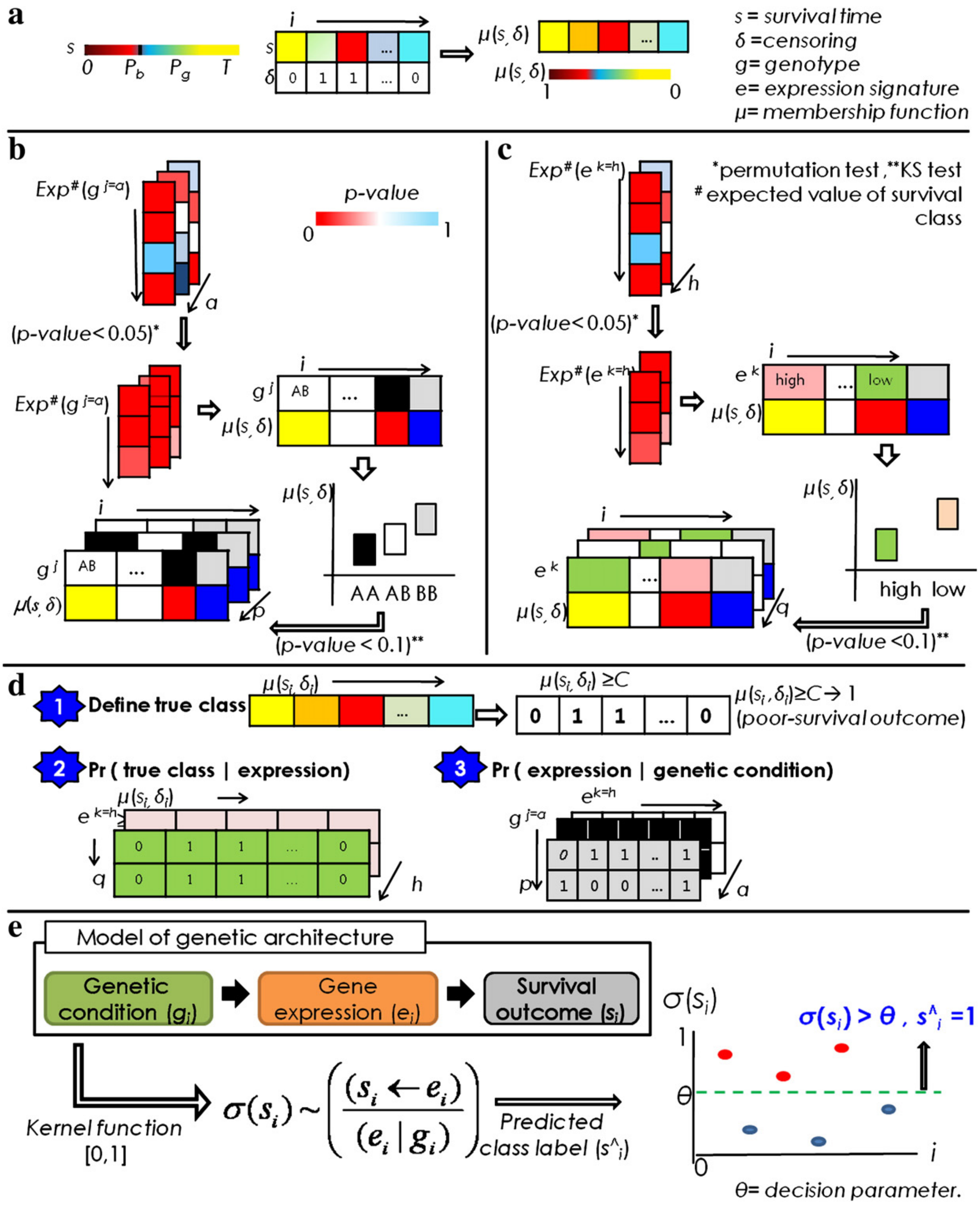

Fig. 1. Overview of feature selection and classifier construction. a) Continuous class labeling of survival outcomes. b) Selection of genotype features using statistical analysis. c) Selection of expression features by statistical analysis. d) Training process. e) Classification of the survival class using the kernel function of CORE-F.

training of multiple features without overfitting [14]. Although we tested trained interactions among four sets of SNPs and expression signatures in selected pathways, only the best-performing interactions between member SNPs and expression signatures in each single pathway were utilized in further analyses. As shown in Table 2, the best results of CORE-F were achieved using a training regimen that involved interactions among 12 prognostic features of the membrane transport ABC pathway (10 SNPs and two expression signatures). In comparison, the ensemble tree performed optimally after a training regimen involving 23 features of the platinum-related pathway (19 SNPs and 4 expression signatures). We therefore validated the performance of CORE-F by comparing its best-performing results with the membrane transport ABC pathway and the best-performing results of the ensemble tree with the platinum-related pathway.

The optimized classifiers that yielded the best performances (listed in Table 2) were used to evaluate the performances of CORE-F and the ensemble tree using 10-fold cross-validation and Kaplan–Meier (KM) survival curves (Fig. 3a and b). We found that CORE-F outperformed the ensemble tree with respect to cross-validation and ROCs (mean AUC, 0.85 vs. 0.72), using sets of samples with genotype

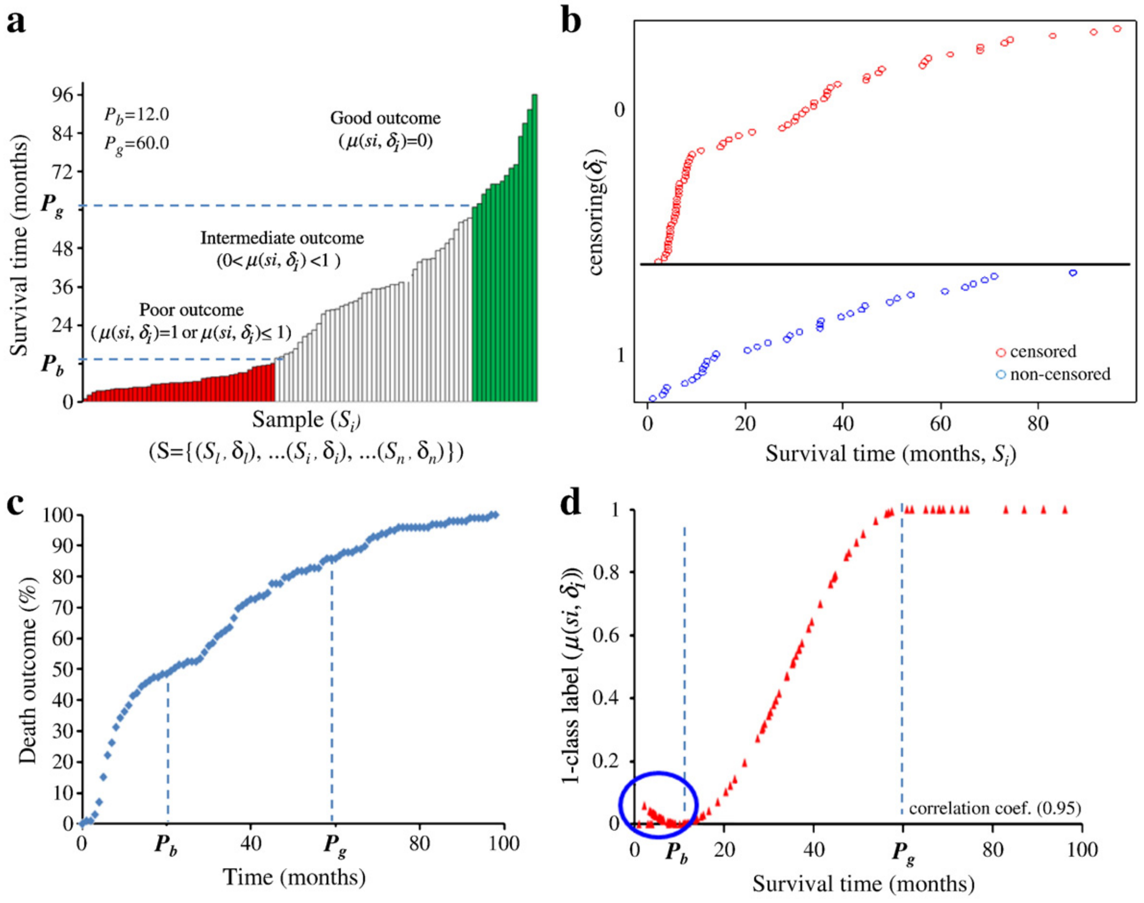

Fig. 2. Class labeling scheme for disease outcome groups. a) Distribution of survival times. P_b and P_g denote the thresholds for the classification of poor and good outcome groups, respectively. b) Presentation of observed survival times in the prepared data set ($S=\{(s_i,\delta_i),\ldots(s_n,\delta_n)\}$, $1\leq i\leq n$). c) Distribution of survival times. d) Results of class labeling for survival outcomes with the membership function $\mu(s_i,\delta_i)$.

and expression values that were complete, with no values missing for interaction training. Similarly, CORE-F was superior to the ensemble tree when assaying the overall set of samples with partial missing values for interaction training (Fig. 3a; mean AUC, 0.77 vs 0.72), indicating that CORE-F tolerated missing values. Typical missing values were missing genotypes arising from the preprocessing of arrays. Because the ensemble tree also utilized the best-performing set of SNPs and expression signatures to predict survival class labels, the better performance of CORE-F highlights the advantages of genetic architecture-based classification methods.

Due to the successful performance of CORE-F, the degree of survival differences between classified labels was presented graphically as KM survival curves. The predicted true class (poor survival outcome) from the CORE-F algorithm resulted in clearly distinct survival trends (p-value = 5.06E-08), in contrast with the labeling results of the ensemble tree (p-value = 0.0001) (Fig. 3c). The better performance of CORE-F was due to the application of a simple model for genetic architecture with parsimonious features. The results of these evaluations support the hypothesis that inspired the development of CORE-F, specifically that the genetic architecture underlying the transcriptional signature and survival traits could produce precisely classify survival outcomes.

Table 1
Selected SNPs and expression signatures.

Step 1) Significance of expected value for survival class label (p-value<0.05)[a]

Pathway	Genetic variations (SNPs)	Expression signatures
Platinum-related[b]	138	8
Taxane-related[b]	119	1
Membrane transport ABC[c]	77	3
Cytochrome P450 (CYP450)[c]	81	2

Step 2) Difference in class values according to feature variations (p-value<0.1)[d]

Pathway	Genetic variations (SNPs)	Expression signatures
Platinum-related[b]	**30**	**6**
Taxane-related[b]	17	1
Membrane transport ABC[c]	**10**	**2**
Cytochrome P450 (CYP450)[c]	15	1

[a] Background distribution describing the expected values in an analysis of the p-value of each expected value is presented based on random permutations.

[b] Resource of PharmGKB.

[c] Resource of KEGG.

[d] Using the KS (Kolmogorov–Smirnov) test. The number of SNPs and expression signatures in bold were utilized finally for the training of interactions and performance validations.

Table 2
Optimized classifiers for CORE-F and ensemble tree methods.

	Pathway	Genetic variants (SNPs)	Expression signatures	Parameters
CORE-F	Membrane transport ABC	10	2	$C=0.14$[a], $\theta=0.27$[b]
Ensemble tree	Platinum-related	19[c]	4[c]	$C'=0.63$[d]

[a] In Eq. (7), the true class of the survival class was defined by C as $\mu(s_i,\delta_i)>C\rightarrow 1$.

[b] In Eq. (7), the predicted class of survival was determined as $\theta{:}\sigma(s_i)>\theta\rightarrow 1$.

[c] After pruning of the ensemble tree.

[d] Using the ensemble tree, the predicted class of survival was determined by C': *predicted class by the ensemble tree*$>C'\rightarrow 1$.

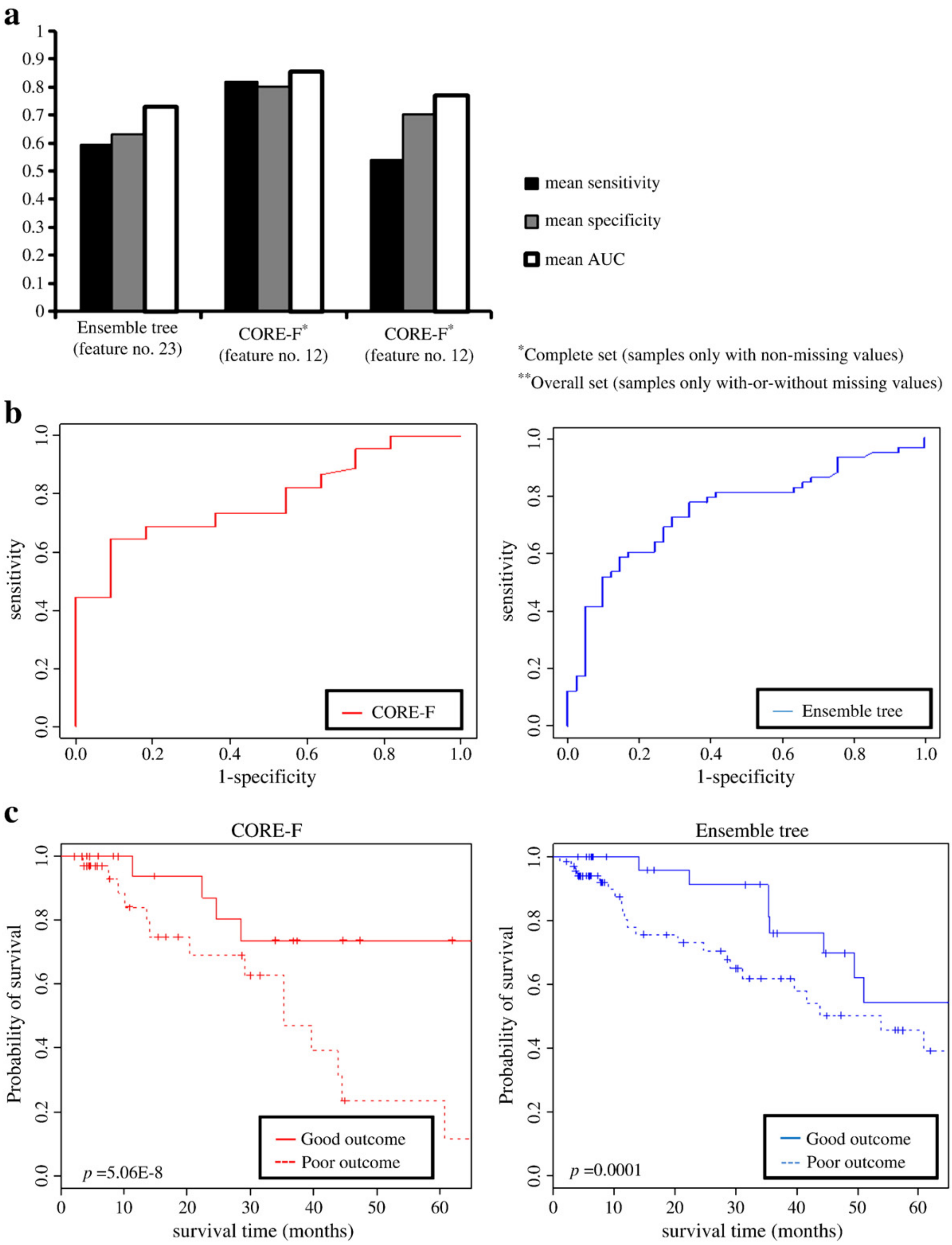

Fig. 3. Performance validation. a) Ten-fold cross validation of the CORE-F and ensemble tree methods. b) ROCs of CORE-F (left panel) and the ensemble tree (right panel). c) Kaplan–Meier survival curves for the predicted classes of CORE-F (left panel) and the ensemble tree (right panel). The *p*-values for survival differences are presented at the bottom of the plots.

3. Discussion

Although there is great interest in biological interactions (e.g., interactions in eQTL network and gene-environment interactions), typical survival prediction algorithms using tree-based methods have generally ignored feature interaction effects *per se* [28]. Using a model of genetic architecture, we developed a novel and effective method, CORE-F, that predicts survival outcomes from the interactions of genetic and expression signatures. The major contribution of CORE-F is its establishment of a classification method that learns the genetic architecture for quantitative traits, such as cancer prognosis. Since genetic architecture presents a model of genetic influence for trait variation and transcriptional signatures, the utilization of CORE-F may ameliorate predictions for other disease phenotypes associated with genetic variations and gene expression profiles. Cross-validation and KM survival curve analysis demonstrated that CORE-F more precisely classified samples from cancer patients into classes with poor or good outcome than did the latest tree-based method, the ensemble tree (13% enhanced AUC, distinct survival trend in the KM curve, p-value = 5.06E-08). Thus, the proposed method successfully unraveled the genetic issues surrounding survival prediction, showing that tree-based methods ineffectively covered the interaction effects described by genetic architectures, such as genetic regulation of transcriptional variations and prognostic differences.

The kernel function of CORE-F may be a promising avenue for identifying the functional roles of genetic variations relative to the underlying genetic architecture. For example, polymorphisms in ABCB1 are promising genetic markers [29,30] for ovarian cancer prognosis, although there is no experimental support, such as transcriptional profiles, for the functional roles of these SNPs. Through the training of interactions between SNPs and expression signatures, the kernel function of CORE-F presents the hypothesis that the genetic architecture of ABCB1 is a *trans*-effect model; SNPs in ABCB1 (rs1202171 and rs1202172) contribute to the expression of ABC transporters (ABCD4 and ABCG5), leading to differences in survival outcomes (Supplemental Figure 1). A previous study of cooperating genetic variations and transcriptional profiles [31] suggested that target phenotype-associated SNPs are enriched in expression quantitative loci and that the interactions between SNPs and gene expression are important for susceptibility to anti-cancer agents. In support of this, the trained interactions in kernel function also showed that genetic variations in ABCB1 (rs1202171 and rs1202172) result in survival differences via transcriptional variations in the membrane transport ABC pathway (p-value for interactions < 0.05, Supplemental Figure 1). Therefore, exploring SNP-gene expression interactions in other prognostic pathways, such as the platinum-related pathway, may increase our understanding of the effects of individual variations on survival outcomes in certain diseases including cancers.

Since early detection of cancer is associated with improved patient prognosis [32], tumor stage may contribute to predictions of survival outcomes. The CORE-F algorithm we utilized was designed based on genetic and transcriptomic profiles rather than by integrating information from cancer stages. To test whether adding stage information improved CORE-F, we compared the results of 10-fold-cross-validation with different data sets, including all 99 selected samples in Table 3 and the subset of 73 samples from patients with stage 3 cancers. We found, however, that information on cancer stage had little effect on both CORE-F and the ensemble tree (AUC differences of 0.02 and 0.03, respectively; Supplemental Figure 2). Using an alternative method, the Wilcoxon-rank sum test of survival outcome, we also found that higher cancer stage (≥III) was an underpowered predictor of patient survival (p-value 0.9). Despite the minimal effect of higher cancer stage in predicting survival, it may be of interest to estimate the prognostic power of early stage cancer samples by expanded analysis using larger data sets including samples at various stages.

Although we confirmed that CORE-F tolerated missing values in the interaction training between SNPs and gene expressions, such as the absence of several genotypes due to quality control of SNP arrays, it was unclear for the degree of tolerances for missing values and patient prognosis with missing values. This may be addressed by the development of feature imputation methods. Although CORE-F outperformed the ensemble tree in predicting binary class of survivorships, such as good and poor prognosis, their relative ability to predict intermediate survival is unclear and may require the development of a multiclass classifier. To our knowledge, however, CORE-F is the first algorithm to predict survival based on biological principles: genes were selectively expressed according to given genotypes under given environmental (or disease) conditions. Including its biomedical benefits in successfully predicting patient prognosis, trained interactions (Supplemental Figure 1) represent the biological value of CORE-F in detecting the underlying genetic mechanisms and the functional roles of genetic variations at the transcript-network level.

Table 3
Overview of the selected samples.

Features of the TCGA[a]	Total[b]	Selected[c]
Number of ovarian cancer samples	213	99
Age		
Mean, years (SD[d])	60.2 (11.1)	59.2 (10.8)
Range	35 - 83.5	35 – 83.5
Tumor stage		
II	2 (0.9%)	1 (1%)
III	171 (80.2%)	73 (73.7%)
IV	40 (18.7%)	25 (25.2%)
Survival outcome[e]		
Mean overall survival, months (SD[d])	32.3 (25.7)	28 (24.2)
> 1 year survival (%)	150 (70.4%)	58 (58.5%)
> 5 year survival (%)	29 (13.6%)	14 (14.1%)
Race		
Caucasian	186 (87.3%)	99 (100%)
African American or Black	9 (4.2%)	
Asian	5 (2.3%)	
Others	13 (6.1%)	
Vital status		
Alive (censored)	96 (45.0%)	64 (64.6%)
Deceased (non-censored)	117 (54.9%)	35 (35.3%)

[a] Date of latest update, 9 September 2009.
[b] Age and/or survival outcome records; missing cases were discarded.
[c] Caucasian, platinum–taxane treated pathway; total number of agents ≤ 3.
[d] SD: Standard deviation.
[e] Wilcoxon rank sum test (total vs. selected set) p-value > 0.05.

4. Conclusion

In conclusion, we developed a novel algorithm, CORE-F, based on underlying genetic architecture to predict the prognosis of cancer patients. By increasing knowledge of genetic architectures and eQTL networks, CORE-F may be used to better predict survival outcomes, as well as enhancing biological insight.

5. Methods

5.1. Labeling of class survival outcome

Using the thresholds for poor/good survival outcomes, we generated a membership function based on fuzzy set theory [21] to transform the unbounded survival outcomes into bounded class labels [0,1]. S was defined as a set of survivorship vectors that included survival time (s_i) and a censoring indicator (δ_i) of the i-th sample ($S = \{(s_1, \delta_1), \ldots(s_i, \delta_i), \ldots(s_n, \delta_n)\}\ 1 \le i \le n$). The censoring indicator denotes the status of censoring ($\delta_i = 1$ for non-censored, 0 for censored status). The survivorship vectors were graded by using a developed membership function to transform the i-th vector of survival into a class label for survival outcomes, yielding labeling results between 0 and 1 ($0 \le \mu(s_i, \delta_i) \le 1$). Since the poor outcome class was defined as a true class (survival time $\le P_b \rightarrow \mu(s_i, \delta_i) = 1$), the survival time and censoring condition of the i-th sample in a non-fuzzy case was transformed into a class label of survival time using the equations:

$$P_b = \text{threshold of poor prognosis;}$$
$$P_g = \text{threshold of good prognosis;}$$

$$\text{if } s_i \le P_b \text{ and } \delta_i = 1 \rightarrow \mu(s_i, \delta_i) = 1; \tag{1}$$

$$\text{if} s_i > P_g \rightarrow \mu(s_i, \delta_i) = 0. \tag{2}$$

For intermediate survivorship with continuous values ($0 \leq \mu(s_i, \delta_i) \leq 1$), the following equation of membership function determined the class labels:

$$\text{if } P_\text{b} < s_i \leq P_\text{g} \text{ and } (\delta_i = 1 \text{ or } 0) \rightarrow \mu(s_i, \delta_i) = \frac{1}{2} + \frac{1}{2}\, cos\left(\frac{s_i - P_\text{b}}{P_\text{g} - P_\text{b}}\pi\right). \tag{3}$$

The uncertainty in survival time for the remaining cases, in which a short period of survivorship was determined according to the early censoring event, was also modeled by Eq. (3):

$$\text{if } s_i \leq P_\text{b} \text{ and } \delta_i = 0 \rightarrow \mu(s_i, \delta_i) = \frac{1}{2} + \frac{1}{2}\, cos\left(\frac{s_i - P_\text{b}}{P_\text{g} - P_\text{b}}\pi\right). \tag{3}$$

Transformed class labels of survival times were utilized for feature selection and learning of classifiers for both CORE-F and the ensemble tree.

5.2. Survival associated feature selections using statistical analysis

Before the training of interactions between SNPs and gene expression levels, prognosis associated features (SNP and expression signatures) were determined in two steps: 1) identification of features that yielded the expected, significantly high or low value of survival class, and 2) capture of features that showed significant variations in expected value of survival class within the observed variations of each feature. Since the training steps followed feature selection produced information about these interactions, we independently selected survival class-associated features without considering the interactions among SNPs and expression signatures (Fig. 1b and c).

In the statistical analysis for genotype selection, where $n^{j=a}$ was the total number of samples in the membership function $\mu({s_i}^{j=a}, {\delta_i}^{j=a})$ that presented a class label for survival outcomes in the i-th sample, with the alleles of the j-th locus being of the a-th genotype; for example, AA, AB, and BB (Fig. 1b). Therefore, the expected value of the survival class determined for the a-th genotype of the j-th locus was:

$$\text{Exp}\left(g^{j=a}\right) = \sum \mu\left(s_i^{j=a}, \delta_i^{j=a}\right) \Big/ n^{j=a}. \tag{4}$$

Similarly, the expected value of the survivorship class for the h-th expression label, for example, high and low values in the k-th gene, was calculated as (Fig. 1c):

$$\text{Exp}\left(e^{k=h}\right) = \sum \mu\left(s_i^{k=h}, \delta_i^{k=h}\right) \Big/ m^{k=h}, \tag{5}$$

where $m^{k=h}$ is the total number of samples for which e^k is the h-th expression label.

Application of the permutation approach [33] with these expected values permitted identification of the features related to the significant class label for survival time (p-value<0.05). The Kolmogorov–Smirnov (KS) test was used to determine the significance of class label alterations by the symbolic labeling of features, such as genotype differences at the j-th locus (p-value<0.1).

5.3. Training and building of CORE-F

A model of genetic architecture was used to suggest our classifier via CORE-F learning processes (Fig. 1d):

(1) The true class of survival outcomes was defined using the cut-off parameter, C; $\mu(s_i, \delta_i) \geq C \rightarrow 1$.

(2) The degree of transcriptional associations for the true class of survival outcome was determined: $\Pr(\mu(s_i, \delta_i) \geq C \mid e^{k=h})$.

(3) The genetic associations for the expression signatures were determined: $\Pr(e^{k=h} \mid g^{j=a})$.

Previously, the membership function graded the degree of survivorship class in a continuous manner ($0 \leq \mu(s_i, \delta_i) \leq 1$). Therefore, the cut-off parameter, C, dichotomized the class label into true (poor-survival outcome) and false (good-survival outcome) classes for the supervised learning of CORE-F (step 1). The trained associations (steps 2 and 3) were introduced into the kernel function of CORE-F as interaction terms.

The interactions between transcriptional activity and a given genetic condition suggest the mechanism underlying the phenotype variation [34]. To highlight the genetic architecture underlying gene expression and survival outcome, a kernel function was designed to present a high score with strong associations among the expression signatures and the genetic variations. Therefore, the kernel function of the CORE-F, $\sigma(s_i)$, utilized the inverse of the associations among genetic and transcriptional features as the denominator of the transcriptional associations for the true class: $Pr(\mu(s_i, \delta_i) \geq C \mid e^{k=h})$. The equation for the kernel function was calculated as (Fig. 1e):

$$if\ \Pr\left(e_i^k \mid g_i^j\right) > 0, \quad \sigma(s_i) = \frac{1}{p*q} * \sum_{k=1}^{q} \sum_{j=1}^{p} \left(\frac{\Pr\left(\mu(s_i, \delta_i) >= C \mid e_i^k\right)}{\Pr\left(e_i^k \mid g_i^j\right)^{-1}}\right) \tag{6}$$

where $1 \leq k \leq q$ for e_i^k, $1 \leq j \leq p$ for g_i^j and $0 \leq C \leq 1$ indicates the cut-off parameter for the true class of survival class label $\mu(s_i, \delta_i)$.

The kernel function $\sigma(s_i)$ predicted the score of the true class (poor-survival outcomes) using the given genotypes (g_i^j) and expression signatures (e_i^k) in the i-th sample (s_i). In the kernel function, the collected score was divided by the total number of feature combinations ($p*q$) to prevent a false-positive classification due to a large number of features.

With kernel function derived values, the predicted class label for the i-th sample ($\hat{s_i}$) was determined by the decision parameter θ ($0 \leq \theta \leq 1$) as:

$$\hat{s}_i = \begin{cases} \text{if } \sigma(S_i) \leq \theta, 0 \\ \text{else}, 1 \end{cases}. \tag{7}$$

The values of the parameters C and θ were tested and recursively introduced over a set of values bounded by 0 and 1 and separated by intervals of 0.01.

5.4. Data set

We selected 99 tissue samples from ovarian cancer patients that had been enrolled in the TCGA (The Cancer Genome Atlas, http://www.cancergenome.nih.gov/). All samples were from Caucasian patients who underwent homogeneous treatment (Table 3). These samples adequately represented the range of survival outcomes without stratification by therapeutic trial or ethnic background (p-value>0.05). Overall patient survival was estimated from the date of diagnosis to the date of death or latest follow-up.

Tumor materials in the TCGA project were excised prior to administration of anticancer treatment, and genomic DNA and RNA were extracted as described by the Biospecimen Core Resource (BCR), a component of the TCGA. WG-SNP6.0 was used to detect genetic variations (SNPs) and HG-U133 (Affymetrix, Inc., USA) was used to measure gene expression levels. Expression signatures and SNPs were profiles according to the guidelines of the platform manufacture (Affymetrix, Inc., USA).

5.5. Construction of classifiers for the ensemble tree method to validate CORE-F

To validate the CORE-F algorithm, we compared its performance with that of the recently developed ensemble tree method [14]. Because the key feature of CORE-consists of the "feature interactions in the model of genetic architecture," the ensemble tree method was trained without assuming a genetic architecture involving genetic associations that influence gene expression. Both the ensemble tree and the CORE-F routines predicted the class labels for survival ($\mu(s_i, \delta_i)$) to determine the power of the CORE-F method in the absence of class labeling effects. The sensitivity and specificity of the ensemble tree were measured using the cut-off parameter ($0 \leq C' \leq 1$) of the true class (poor outcome), with the predicted class by the ensemble tree being $> C' \rightarrow 1$.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.03.005.

Acknowledgments

This work was supported by the World Class University program (R32-2008-000-10218-0), the Korean Systems Biology Research Project (20100002164), and the Basic Research Laboratory grant (2009-0086964) of the Ministry of Education, Science and Technology through the National Research Foundation of Korea. DL and EL were also supported by the Samsung Biomedical Research Institute (grant C-A7-101-2). HP was supported by grants from an Omics-Based Integration Database for Cancer Interpretation (KGM1161012 and KGM0661113) funded by the Korea Research Institute of Bioscience and Biotechnology (KRIBB). The results published here are based in whole or in part upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at http://www.cancergenome.nih.gov.

References

[1] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, J.L. Wrana, Dynamic modularity in protein interaction networks predicts breast cancer outcome, Nat. Biotechnol. 27 (2009) 199–204.

[2] X. Yang, Y. Huang, J.L. Chen, J. Xie, X. Sun, Y.A. Lussier, Mechanism-anchored profiling derived from epigenetic networks predicts outcome in acute lymphoblastic leukemia, BMC Bioinform. 10 (Suppl 9) (2009) S6.

[3] H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis, Mol. Syst. Biol. 3 (2007) 140.

[4] L. Tian, Y. Wang, D. Xu, J. Gui, X. Jia, H. Tong, X. Wen, Z. Dong, and Y. Tian, Serological AFP/Golgi protein 73 could be a new diagnostic parameter of hepatic diseases. Int J Cancer.

[5] B. Zhang, A. zur Hausen, M. Orlowska-Volk, M. Jager, H. Bettendorf, S. Stamm, M. Hirschfeld, O. Yiqin, X. Tong, G. Gitsch, and E. Stickeler, Alternative splicing-related factor YT521: an independent prognostic factor in endometrial cancer. Int J Gynecol Cancer 20 492–9.

[6] R. McPherson, A. Pertsemlidis, N. Kavaslar, A. Stewart, R. Roberts, D.R. Cox, D.A. Hinds, L.A. Pennacchio, A. Tybjaerg-Hansen, A.R. Folsom, E. Boerwinkle, H.H. Hobbs, J.C. Cohen, A common allele on chromosome 9 associated with coronary heart disease, Science 316 (2007) 1488–1491.

[7] S.F. Grant, G. Thorleifsson, I. Reynisdottir, R. Benediktsson, A. Manolescu, J. Sainz, A. Helgason, H. Stefansson, V. Emilsson, A. Helgadottir, U. Styrkarsdottir, K.P. Magnusson, G.B. Walters, E. Palsdottir, T. Jonsdottir, T. Gudmundsdottir, A. Gylfason, J. Saemundsdottir, R.L. Wilensky, M.P. Reilly, D.J. Rader, Y. Bagger, C. Christiansen, V. Gudnason, G. Sigurdsson, U. Thorsteinsdottir, J.R. Gulcher, A. Kong, K. Stefansson, Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes, Nat. Genet. 38 (2006) 320–323.

[8] E.E. Schadt, C. Molony, E. Chudin, K. Hao, X. Yang, P.Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, J. Zhu, J. Millstein, S. Sieberts, J. Lamb, D. GuhaThakurta, J. Derry, J.D. Storey, I. Avila-Campillo, M.J. Kruger, J.M. Johnson, C.A. Rohl, A. van Nas, M. Mehrabian, T.A. Drake, A.J. Lusis, R.C. Smith, F.P. Guengerich, S.C. Strom, E. Schuetz, T.H. Rushmore, R. Ulrich, Mapping the genetic architecture of gene expression in human liver, PLoS Biol. 6 (2008) e107.

[9] Y. Li, C.C. Sheu, Y. Ye, M. de Andrade, L. Wang, S.C. Chang, M.C. Aubry, J.A. Aakre, M.S. Allen, F. Chen, J.M. Cunningham, C. Deschamps, R. Jiang, J. Lin, R.S. Marks, V.S. Pankratz, L. Su, Z. Sun, H. Tang, G. Vasmatzis, C.C. Harris, M.R. Spitz, J. Jen, R. Wang, Z.F. Zhang, D.C. Christiani, X. Wu, and P. Yang, Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. Lancet Oncol 11 321–30.

[10] N.P. Crawford, R.C. Walker, L. Lukes, J.S. Officewala, R.W. Williams, K.W. Hunter, The Diasporin Pathway: a tumor progression-related transcriptional network that predicts breast cancer survival, Clin. Exp. Metastasis 25 (2008) 357–369.

[11] D. Etemadmoghadam, A. deFazio, R. Beroukhim, C. Mermel, J. George, G. Getz, R. Tothill, A. Okamoto, M.B. Raeder, P. Harnett, S. Lade, L.A. Akslen, A.V. Tinker, B. Locandro, K. Alsop, Y.E. Chiew, N. Traficante, S. Fereday, D. Johnson, S. Fox, W. Sellers, M. Urashima, H.B. Salvesen, M. Meyerson, D. Bowtell, Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas, Clin. Cancer Res. 15 (2009) 1417–1427.

[12] J.D. Piette, O. Intrator, S. Zierler, V. Mor, M.D. Stein, An exploratory analysis of survival with AIDS using a nonparametric tree-structured approach, Epidemiology 3 (1992) 310–318.

[13] I. Bray, K. Bryan, S. Prenter, P.G. Buckley, N.H. Foley, D.M. Murphy, L. Alcock, P. Mestdagh, J. Vandesompele, F. Speleman, W.B. London, P.W. McGrady, D.G. Higgins, A. O'Meara, M. O'Sullivan, R.L. Stallings, Widespread dysregulation of MiRNAs by MYCN amplification and chromosomal imbalances in neuroblastoma: association of miRNA expression with survival, PLoS ONE 4 (2009) e7850.

[14] J.A. Koziol, A.C. Feng, Z. Jia, Y. Wang, S. Goodison, M. McClelland, D. Mercola, The wisdom of the commons: ensemble tree classifiers for prostate cancer prognosis, Bioinformatics 25 (2009) 54–60.

[15] A.P. Bremner, R.H. Taplin, Modified classification and regression tree aplitting criteria for data with interations, Aust. NZ J. Stat. 44 (2002) 169–176.

[16] M. Garcia-Magarinos, I. Lopez-de-Ullibarri, R. Cao, A. Salas, Evaluating the ability of tree-based methods and logistic regression for the detection of SNP–SNP interaction, Ann. Hum. Genet. 73 (2009) 360–369.

[17] R.J. Guerreiro, D.R. Gustafson, J. Hardy, The genetic architecture of Alzheimer's disease: beyond APP, PSENs and APOE, Neurobiol. Aging (2010).

[18] D. Altshuler, M.J. Daly, E.S. Lander, Genetic mapping in human disease, Science 322 (2008) 881–888.

[19] A.M. Glazier, J.H. Nadeau, T.J. Aitman, Finding genes that underlie complex traits, Science 298 (2002) 2345–2349.

[20] T.C.G.A.R. Network, Comprehensive genomic characterization defines human glioblastoma genes and core pathways, Nature 455 (2008) 1061–1068.

[21] B.M. Mohan, A. Sinha, Mathematical models of the simplest fuzzy PI/PD controllers with skewed input and output fuzzy sets, ISA Trans. 47 (2008) 300–310.

[22] J. Engel, R. Eckel, G. Schubert-Fritschle, J. Kerr, W. Kuhn, J. Diebold, R. Kimmig, J. Rehbock, D. Holzel, Moderate progress for ovarian cancer in the last 20 years: prolongation of survival, but no improvement in the cure rate, Eur. J. Cancer 38 (2002) 2435–2445.

[23] T.J. Duncan, A. Al-Attar, P. Rolland, S. Harper, I. Spendlove, and L.G. Durrant, Cytoplasmic p27 expression is an independent prognostic factor in ovarian cancer. Int J Gynecol Pathol 29 8–18.

[24] K.D. Steffensen, M. Waldstrom, A. Jakobsen, The relationship of platinum resistance and ERCC1 protein expression in epithelial ovarian cancer, Int. J. Gynecol. Cancer 19 (2009) 820–825.

[25] R.S. Huang, S. Duan, S.J. Shukla, E.O. Kistner, T.A. Clark, T.X. Chen, A.C. Schweitzer, J. E. Blume, M.E. Dolan, Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach, Am. J. Hum. Genet. 81 (2007) 427–437.

[26] M. Hewett, D.E. Oliver, D.L. Rubin, K.L. Easton, J.M. Stuart, R.B. Altman, T.E. Klein, PharmGKB: the pharmacogenetics knowledge base, Nucleic Acids Res. 30 (2002) 163–165.

[27] K.F. Aoki-Kinoshita, M. Kanehisa, Gene annotation and pathway mapping in KEGG, Methods Mol. Biol. 396 (2007) 71–91.

[28] H.J. Cordell, Detecting gene–gene interactions that underlie human diseases, Nat. Rev. Genet. 10 (2009) 392–404.

[29] A. Hamidovic, K. Hahn, J. Kolesar, Clinical significance of ABCB1 genotyping in oncology, J. Oncol. Pharm. Pract. 16 (2010) 39–44.

[30] E. Balcerczak, M. Panczyk, S. Piaskowski, G. Pasz-Walczak, A. Salagacka, M. Mirowski, ABCB1/MDR1 gene polymorphisms as a prognostic factor in colorectal cancer, Int. J. Colorectal Dis. (2010).

[31] E.R. Gamazon, R.S. Huang, N.J. Cox, and M.E. Dolan, Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proc Natl Acad Sci U S A 107 9287–92.

[32] D. Badgwell, R.C. Bast Jr., Early detection of ovarian cancer, Dis. Markers 23 (2007) 397–410.

[33] E. Czwan, B. Brors, D. Kipling, Modelling *p*-value distributions to improve theme-driven survival analysis of cancer transcriptome datasets, BMC Bioinform. 11 (2010) 19.

[34] S.I. Lee, A.M. Dudley, D. Drubin, P.A. Silver, N.J. Krogan, D. Pe'er, D. Koller, Learning a prior on regulatory potential from eQTL data, PLoS Genet. 5 (2009) e1000358.