Bioinformatics, 33(22), 2017, 3619–3626
doi: 10.1093/bioinformatics/btx487
Advance Access Publication Date: 29 July 2017
Original Paper

OXFORD

Systems biology

Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers

Jonghwan Choi[1], Sanghyun Park[2], Youngmi Yoon[3] and Jaegyoon Ahn[1,*]

[1]Department of Computer Science and Engineering, Incheon National University, Incheon, The Republic of Korea, [2]Department of Computer Science, Yonsei University, Seoul, The Republic of Korea and [3]Department of Computer Engineering, Gachon University, Seongnam-si, Gyeonggi-do, The Republic of Korea

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on April 21, 2017; revised on June 27, 2017; editorial decision on July 25, 2017; accepted on July 27, 2017

Abstract

Motivation: Identification of genes that can be used to predict prognosis in patients with cancer is important in that it can lead to improved therapy, and can also promote our understanding of tumor progression on the molecular level. One of the common but fundamental problems that render identification of prognostic genes and prediction of cancer outcomes difficult is the heterogeneity of patient samples.

Results: To reduce the effect of sample heterogeneity, we clustered data samples using K-means algorithm and applied modified PageRank to functional interaction (FI) networks weighted using gene expression values of samples in each cluster. Hub genes among resulting prioritized genes were selected as biomarkers to predict the prognosis of samples. This process outperformed traditional feature selection methods as well as several network-based prognostic gene selection methods when applied to Random Forest. We were able to find many cluster-specific prognostic genes for each dataset. Functional study showed that distinct biological processes were enriched in each cluster, which seems to reflect different aspect of tumor progression or oncogenesis among distinct patient groups. Taken together, these results provide support for the hypothesis that our approach can effectively identify heterogeneous prognostic genes, and these are complementary to each other, improving prediction accuracy.

Availability and implementation: https://github.com/mathcom/CPR

Contact: jgahn@inu.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Identification of genes that can be used to predict prognosis in cancer patients is an important goal in bioinformatics, because these genes can be used as biomarkers to provide patients with appropriate therapies, as well as to help us understand molecular mechanisms of tumor progression (Bullinger *et al.*, 2004; Sotiriou *et al.*, 2006). Numerous studies have utilized expression levels of several prognostic genes to predict prognosis of various cancer types (Abba *et al.*, 2010; Buyse *et al.*, 2006), using feature selection methods such as Support Vector Machine (Sun *et al.*, 2011) or Lasso (Sohn *et al.*, 2009) regression to obtain prognostic genes.

Recently biological networks have been extensively studied as another potential means to understand disease (Barabasi *et al.*, 2011; Furlong, 2013). Some approaches predict cancer outcome and identify key genes that result in poor prognosis using gene expression and various types of omics data, such as protein–protein interaction (PPI) networks or gene regulatory networks (Chuang *et al.*, 2007; Dao *et al.*, 2011), or use gene co-expression networks to

Table 1. Summary of gene expression datasets

Name	#Total samples	#Poor samples	#Good samples	#Total genes	Characteristic for label	Reference
BRCA	189	90	99	20 501	days_to_death	GDAC firehose
GSE3494	157	36	121	13 181	Disease-Specific Survival Time	Miller *et al.* (2005)
GSE4922	175	69	106	13 181	Disease Free Survival Time	Ivshina *et al.* (2006)
GSE7390	140	24	116	13 181	Time of Overall Survival	Desmedt *et al.* (2007)
GSE24450	121	25	96	25 106	10years followup time	Heikkinen *et al.* (2011)
NKI	120	47	73	10 703	TIMEsurvival	van de Vijver *et al.* (2002)

identify prognostic genes in network form (Ren *et al.*, 2016; Wu and Stein, 2012). We previously proposed a network-based method to identify gene networks for the accurate diagnosis of prostate cancer (Ahn *et al.*, 2011); this approach exploits the differences of correlations of gene pairs in the biological networks between two patient groups, but does not show good performance in application to prognosis data. Other approaches have been proposed for improving prediction of cancer outcomes (Roy *et al.*, 2014; Winter *et al.*, 2012). These methods identify prognostic genes using gene expression measurements and several gene networks including PPI networks. Genes are ranked according to their prognostic relevance using both expression and network information using NetRank, which iteratively calculates gene scores in a similar way to PageRank, but additionally uses correlation between gene expression values and patient survival time (Winter *et al.*, 2012), or t statistics or fold-change of gene expressions between patients with good and poor prognosis (Roy *et al.*, 2014).

Those methods commonly suffer from the heterogeneity of samples from cancer patients (Polyak, 2011), which makes prediction difficult. Since the main known source of the heterogeneity is genomic instability (Burrell *et al.*, 2013), poor prediction may result from difficulties in identifying prognostic genes or biomarkers that are specific to certain cancer patients. Breast cancer has several genetic subtypes that have disparate clinical response and tumor progression (Russnes *et al.*, 2011), so there have been many studies to deal with heterogeneity of breast cancer. Even one type of breast cancer, TNBC (triple-negative breast cancer), has six subtypes that shows different biological characteristics (Lehmann *et al.*, 2011).

A common approach to overcome the cancer heterogeneity problems is an integrative analysis of various omics data. Szeto *et al.* analyzed genomic, transcriptomic and proteomic data of TNBC patients at various clinical time-points to see temporal heterogeneity of TNBC (Szeto *et al.*, 2017). Wang *et al.* proposed the network approach that integrates heterogeneous omics data and demonstrated that resulting glioblastoma and breast cancer patient groups show different survival profiles (Wang *et al.*, 2016).

In the present study, we first clustered samples to deal with heterogeneity, using principal components of whole genes. For each cluster, we gave weights to edges in the functional interaction (FI) networks using t statistics of gene expressions between samples with good and poor outcomes in the cluster, and applied a modified PageRank algorithm to prioritize prognostic genes. Hub genes among resulting prioritized genes were selected as biomarkers to guarantee stable prediction accuracy among independent datasets. Then, we used Random Forest (Breiman, 2001) to predict prognosis of patient samples from gene expressions of the biomarkers.

This proposed approach outperformed the co-expression network approaches and NetRank, as well as traditional feature selection methods such as Lasso for predicting the prognosis of patients with breast cancer. We confirmed that clustering of heterogeneous samples actually contributes to prediction accuracy. In fact, we were able to find many cluster-specific prognostic genes for each dataset and functional study showed that distinct biological processes or pathways were enriched in each cluster; that is, our approach can actually enable detection of functional modules reflecting characteristics of patient groups. These results together provide support for the hypothesis that our approach can effectively identify heterogeneous prognostic genes and that these are complementary to each other, improving prediction accuracy.

2 Materials and methods

2.1 Data preparation

We downloaded one high-throughput sequencing (HTS) data from Broad Institute GDAC Firehose (Bitgda, 2016) and five microarray (MA) datasets from Gene Expression Omnibus (GEO) (Barrett *et al.*, 2013) and van de Vijver *et al.* We used RSEM information for 20 501 genes from the HTS data of 189 patients. Four breast cancer prognostic datasets, GSE3494 (Miller *et al.*, 2005), GSE4922 (Ivshina *et al.*, 2006), GSE7390 (Desmedt *et al.*, 2007), GSE24450 (Heikkinen *et al.*, 2011) and their clinical information were collected from GEO. We also used 120 samples of van de Vijver dataset (van de Vijver *et al.*, 2002). For all data, each sample was labeled as good prognosis if the patient survived more than 10 years (5 years for HTS dataset and GSE24450), and labeled as poor prognosis if the patient did not survive more than 5 years. Table 1 shows a summary of the described datasets. For all data, gene expression values were normalized for each sample by z-scoring.

FIs network is derived from curated pathways, protein-protein interactions, gene coexpression, Gene Ontology (GO) annotations and text-mined protein interactions (Wu *et al.*, 2010). We downloaded the network data from Reactome (Croft *et al.*, 2014; Fabregat *et al.*, 2016), and the number of interactions is 150 322.

2.2 Cancer gene identification

We first computed principal components of gene expressions as attributes, and applied K-means clustering to samples with k with maximal silhouette score. For each cluster of samples, we computed weights of edges in FI networks with t statistics of genes. Then we applied the modified PageRank. It ranks genes by ratios of two PageRank scores which are calculated using weighted and unweighted networks. The modified PageRank uses the following formula:

$$s_i^n = (1-d)s_i^0 \; + \; d\sum_{j\in NE_i} w_{ij}s_j^{n-1}$$

where s_i^n denotes a score of gene i after n iterations, d is a damping

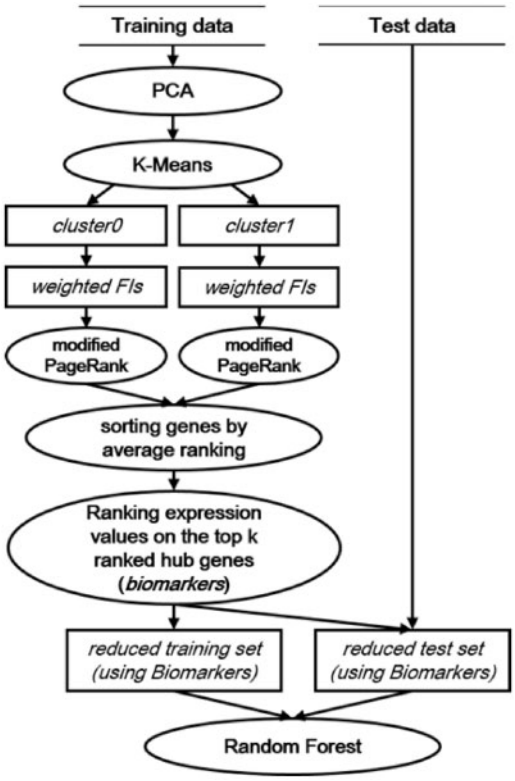

Fig. 1. Gene selection procedure. K-means clustering partitions training data using two principal components of whole genes. For each cluster, we give weights to edges in FI networks using t statistics, and apply modified PageRank to identify prognostic genes. The hub genes among them (degrees with the top 2%) were selected as biomarkers that are used to predict prognosis

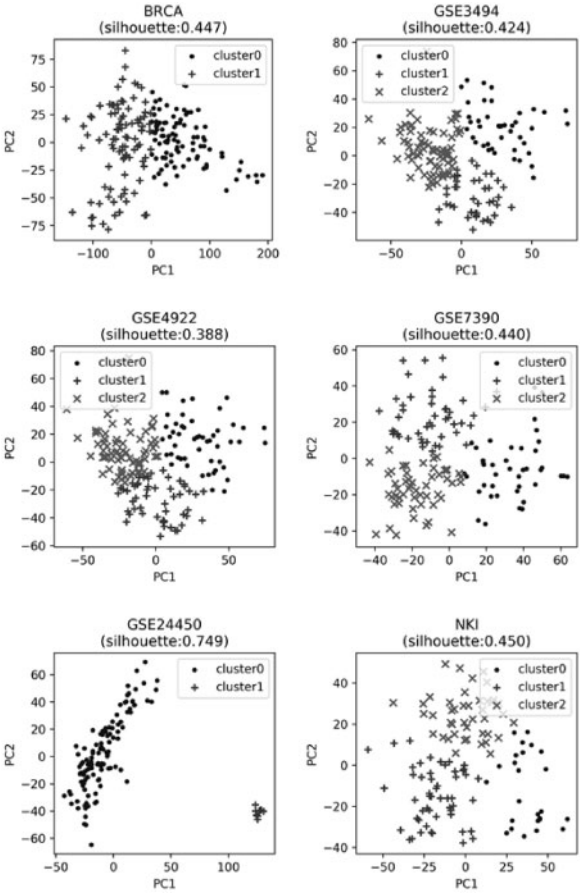

Fig. 2. PCA plots. PCA plots using two highest principal components (PC) for each dataset

factor (user parameter, $0 < d < 1$), NE_i is a set of genes which are neighbors of gene i, and w_{ij} is calculated using:

$$w_{ij} = \begin{cases} \dfrac{|t_i|}{\sum_{k \in NE_i} |t_k|} & \text{on weighted networks} \\ \dfrac{1}{|NE_j|} & \text{on un}-\text{weigthed networks,} \end{cases}$$

where $|NE_j|$ is the size of NE_i. A damping factor d adjusts an influence of the FI network information. As a damping factor increases, an influence of the network also increases. The default value of d is 0.7, which is found to be optimal through iterative experiments. t_i is t statistics to test whether the means of expressions of gene i are different between samples with poor and good outcomes. The initial score s_i^0 has $1/|G|$, where $|G|$ is the number of genes, and thus the sum of scores is always 1. We decided that a score is converged if its change is less than 0.005, and observed the convergence after five iterations. So we computed average genes ranks to integrate gene sets from clusters, after five iterations. We can consider high-ranked genes as driver genes that are responsible for different prognoses.

The high-ranked genes do not all necessarily contribute to prediction of prognosis, because they may not show differences in expressions between poor and good outcomes, as effector genes do. Thus, we selected *cut* (user parameter, $50 \leq cut \leq 150$) hub genes whose degrees were in the top 2% (in our FIs, the degree of the top 2% is 185) as biomarkers. We found that 2% is optimal percentage rate through iterative experiments. We converted expression values of biomarkers to rank values for each sample to avoid test set bias (Patil *et al.*, 2015), and applied Random Forest to the converted training and test data. We selected Random Forest since it showed the best result among widely used classification algorithms (Supplementary Fig. S1). The entire pipeline is provided in Figure 1 and implemented in Python with scikit-learn module (Pedregosa *et al.*, 2011).

3 Results

3.1 Effect of sample clustering

The first step of our approach was clustering of samples, as they show different gene expression profiles. We were able to confirm that samples are roughly divided into two or three groups in PCA plots (Fig. 2).

Note that those clusters did not show strong correlation with estrogen receptor (ER) status. Supplementary Table S1(a) shows that cluster1 and cluster2 of GSE4922, and cluster1 and cluster2 of GSE7390 have high ER+ sample ratios. However, clusters with many ER- samples also had many ER+ samples, which means that clustering does not necessarily reflect ER status. In fact, we found many ER+ samples also had poor outcome, although generally ER- samples shows poor outcome. As well as ER, PR status was not related with outcome (Supplementary Table S1b). We will show that the sample clustering actually contributes prediction accuracy and identification of driver genes, in subsequent chapters (Figs 3–5).

3.2 Performance of prognosis prediction

We performed 10-fold cross-validation and independent tests for different sets of two user parameters *cut* and *d*. Complete results are provided in Supplementary Table S2.

While performing independent tests, we found that datasets showed very different distributions of gene expression, as shown in Figure 3. That is, training and test data were not compatible when we applied Random Forest and other classification algorithms. We focused on the identification of reproducible biomarkers that can be used for different independent datasets. Therefore, we designed independent tests as 10-fold cross-validation of test data with the gene set identified using training data. We used one dataset as training data and the other two as test datasets for independent tests. k in K-means clustering was set to 2 or 3 which maximized silhouetted scores. Optimal parameters are those with the best area under curve (AUC) in cross-validations (Supplementary Table S3).

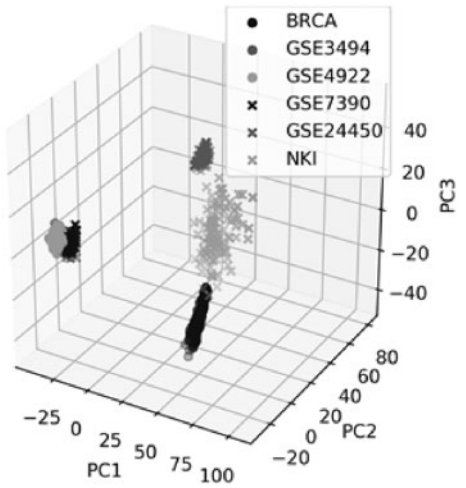

Fig. 3. PCA 3D scatter plot. PCA plots using three highest PCs for all datasets

We named our approach CPR (Clustering and PageRank), and compared it with CPR without clustering (hereafter, NCPR), Lasso (Sohn *et al.*, 2009), NetRank (Roy *et al.*, 2014), Wu and Stein (2012) and WGCNA (Langfelder and Horvath, 2008). Figure 4 shows ROC curve with BRCA as training data. We can see that our classification accuracy in terms of AUC is higher than others (0.001–0.094), except BRCA (training)-GSE4922 (test) and BRCA (training)-GSE24450 (test) case. CPR is generally superior to NCPR, which shows the effect of sample clustering. ROC curves from other datasets as training data are provided in Supplementary Figure S2, and it also shows CPR outperforms other approaches in general. The numbers of genes used for all methods and datasets, and the detailed information of genes using CPR is provided in Supplementary Tables S4 and S5, respectively.

We also investigated the minimal sample (good + poor) size that guarantees stable performance. Supplementary Figure S3 shows AUCs varying training sample sizes for all datasets, and we can check that the performance is stable if the sample size is greater than 80.

3.3 Analysis of oncogene inclusion ratio and reproducibility

Next, we counted known breast cancer oncogenes and tumor suppressor genes (TSG) collected from COSMIC (Forbes *et al.*, 2017) and calculated the p-value using hypergeometric test for each dataset. Figure 5 shows that our genes had significantly low p-values overall, which means that the genes we identified have higher

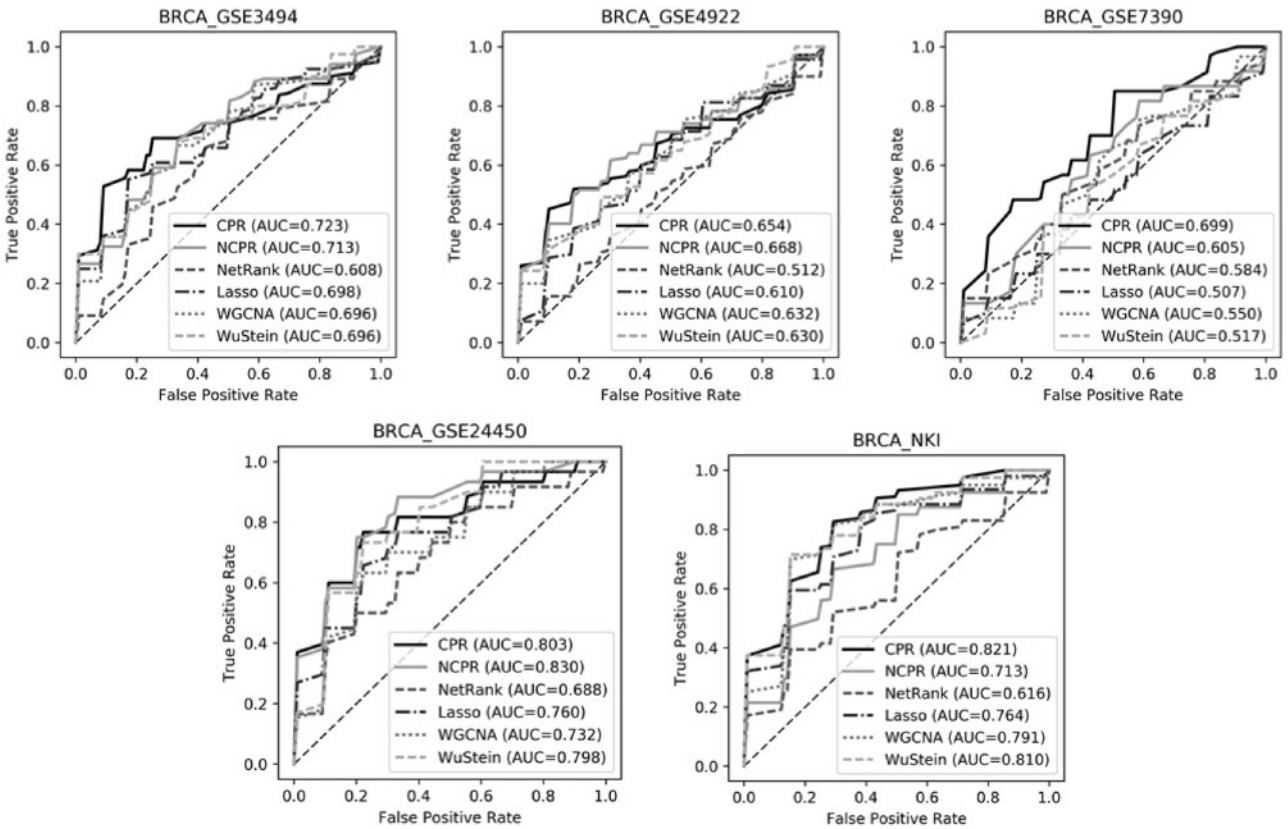

Fig. 4. ROC curves. CPR is compared with NCPR (CPR without clustering), NetRank, Lasso, WGCNA and Wu & Stein, when using BRCA as training data. Complete ROC curves are provided in Supplementary Figure S2. Parameters and number of genes used are shown in Supplementary Tables S4 and S5, respectively

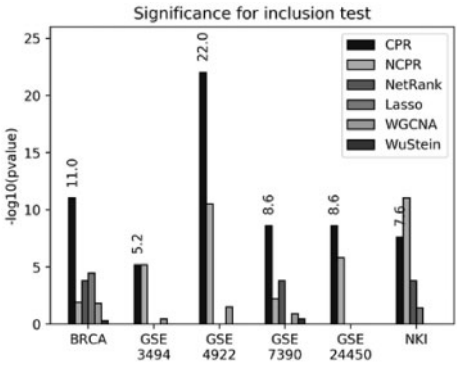

Fig. 5. Significances of breast cancer oncogene and TSG inclusion ratio. For each method and dataset, a significance of the ratio of identified oncogenes and TSGs collected from COSMIC was calculated using hypergeometric test. Complete contingency tables are provided in Supplementary Table S6

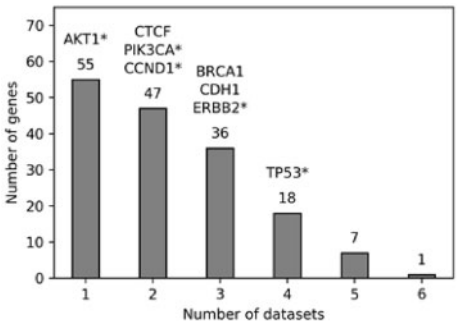

Fig. 6. Reproducibility of biomarker prediction. Each bar indicates the number of genes that are selected as biomarkers for n datasets ($1 \leq n \leq 6$). For example, 55 genes were selected as biomarker for only one dataset. Genes with and without asterisk on the bars are oncogenes and TSGs, respectively

probability of including breast cancer oncogenes or TSGs. Note that CPR shows better results than NCPR, which shows another effect of sample clustering. Complete contingency tables are provided in Supplementary Table S6.

We also calculated the ratio r of overlapping prognostic genes among different datasets by following formula:

$$r = \frac{(\text{number of genes included in at least three data})}{(\text{number of a union of the genes})}.$$

We observed that 37.8% of identified prognostic genes were overlapped at least three among different datasets in Figure 6, while 28.6%, 12.6%, 0.6% and 58.8% of genes were overlapped for Wu & Stein, WGCNA, Lasso and NetRank, respectively. NetRank has higher reproducibility, but identified genes are small number of extreme hub genes, and did not show good prediction accuracy.

The 37.8% genes that we identified include two famous oncogenes ERBB2 and TP53, and two tumor suppressor genes BRCA1 and CDH1. AKT1 was found only in GSE4922, CTCF was found in GSE4922 and NKI, and PIK3CA and CCND1 were found in both BRCA and GSE4922 (Fig. 7a–f). The PI3K-AKT1 signaling pathway has been studied as a clinical target for the breast cancer. Although AKT1 and PIK3CA1 are closely located on the pathway, PIK3CA1 may promote cancer through AKT-independent pathway (Vasudevan *et al.*, 2009) and it has been also reported that their mutations have distinct effects on sensitivity to targeted pathway inhibitors in breast cancer model (Beaver *et al.*, 2013).

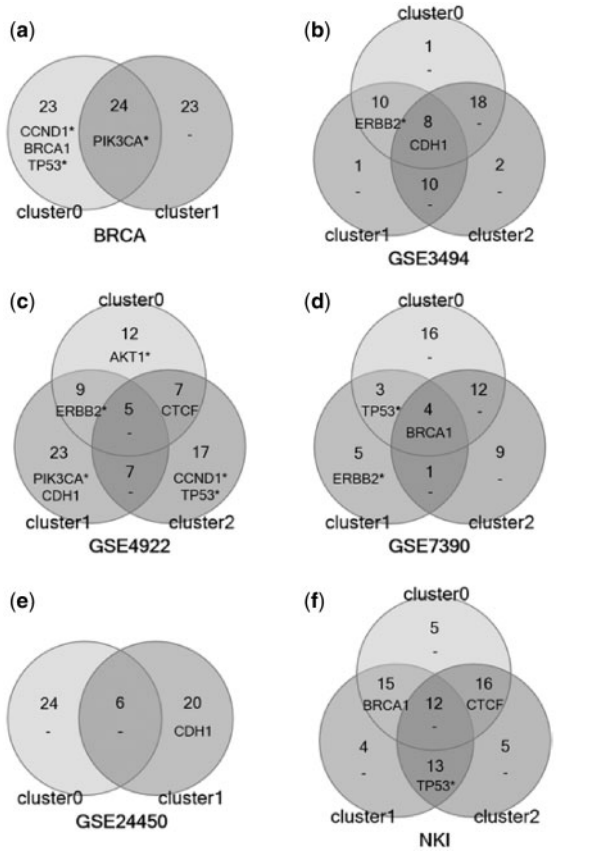

Fig. 7. Oncogene overlapping results among clusters. Overlap of identified biomarkers including known oncogenes and TSGs among clusters for each dataset. Genes with and without asterisk are oncogenes and TSGs, respectively; A number represents the number of genes

3.4 Functional analysis of the gene module

We were able to identify many common biomarkers on both clusters, but there were also many cluster-specific biomarkers, as shown in Figure 7(a–f). Famous oncogenes and TSGs including BRCA1, TP53, ERBB2, CDH1, CTCF and PIK3CA were in common set, but also many oncogenes and TSGs were able to be found in cluster-specific sets. These results tell us that patient samples in different cluster can have different tumor growth patterns or tumorigenesis factors.

For more details, we performed functional analysis on biomarkers of clusters 0 and 1 of BRCA data. A biomarker belongs to cluster x if its rank computed by our PageRank in cluster x is (i) less than 1000 or (ii) less than ranks in all other clusters. The biomarkers of cluster 0 and 1, respectively, were test using DAVID (Huang da *et al.*, 2009a,b) with default settings. Complete list of enriched GO terms (biological process) is provided in Supplementary Table S7. Among them, we selected some interesting GO terms (P-value < 0.05) enriched in each cluster exclusively and visualized them using Cytoscape (Shannon *et al.*, 2003) in Figure 8.

We can observe many cluster-specific GO terms enriched, but we will first focus on the 'response to leptin' of cluster 0 and 'integrin-mediated signaling pathway' of cluster 1. Leptin is known to enhance breast cancer growth and progression (Andò *et al.*, 2014), and it contributes metastasis in ER+ breast cancer cell (Strong *et al.*, 2015). Integrins are receptors which are known to promote survival of breast cancer cells and to control metastasis in breast cancer (Felding-Habermann *et al.*, 2001; Parvani *et al.*, 2015). Hedrick et al. also suggested that integrin-mediated signaling pathway as a

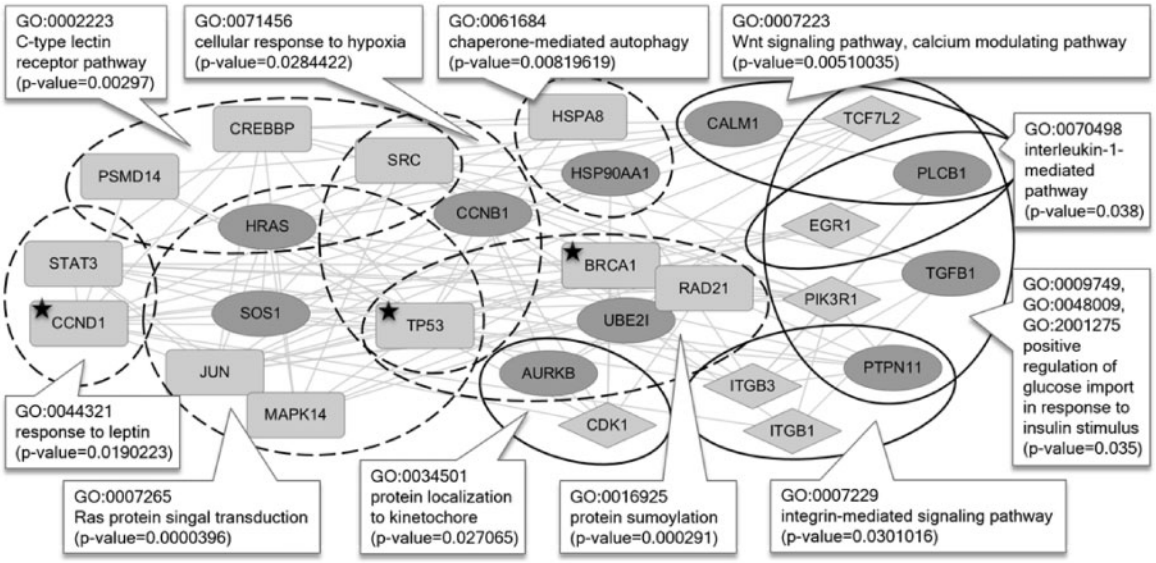

Fig. 8. GO terms enriched in each cluster. Rectangle node: gene in cluster0; diamond node: gene in cluster1; ellipse node: gene in both clusters; Dotted line and solid line circle block: GO term enriched in cluster0 and cluster1 genes, respectively; genes with star mark indicate known oncogenes or TSGs

novel target for treating high-risk breast cancer patients (Hedrick *et al.*, 2016).

We investigated the possibility that leptin, integrins, or their related genes can be used as prognostic biomarkers. Figure 9(a) and (b) shows that the Kaplan–Meier survival curves of STAT3 and CCND1 which are closely related to leptin, show prognostic significance (P-value < 0.05) in cluster0 but not in cluster1. On the other hand, the curves of ITGB1, ITGB3 and PTPN11 show prognostic significance in only cluster1 in Figure 9(c) and (d). These results show that STAT3, CCND1, ITGB1, ITGB3 and PTPN11 can be cluster-specific prognostic biomarkers, which means STAT3/CCND1 and ITGB1/ITGB3/PTPN11 may be mutually exclusive.

Beside leptin and integrin related terms, the relationships between identified GO terms and breast cancer outcome have been investigated. C-type lectin receptor pathway was enriched for biomarkers of cluster 0. C-type lectin receptor is known to regulate adaptive immunity and immunopathogenesis, and has been reported as a new target for cancer immunotherapy (Yan *et al.*, 2015). Ras signaling pathway is known to hyper-activate breast cancer, and it has been studied extensively as a metastasis suppressor (Niemitz, 2013). Elevated Ras signaling was enriched mainly in basal-like and Human Epidermal growth factor Receptor 2-positive (HER2+) subtype tumors, but Ras combined with PIK3CA may be associated with luminal B-like tumor that is ER+ (Wright *et al.*, 2015). PIK3CA was identified in both clusters in Figure 7(a). Hypoxia enriched in cluster0 is associated with risk of metastasis and patient mortality. Since the hypoxic tumor microenvironment affects cancer progression, recent preclinical studies have suggested a therapy with drugs that inhibit hypoxia-inducible factors for good outcomes of patients (Semenza, 2016).

GO terms enriched from biomarkers of cluster1 are somewhat different from those of cluster0. Wnt signaling pathway enriched from genes of cluster1 is known to overexpress in TNBC (Dey *et al.*, 2013). Many researchers studied the pathway as a biomarker for patient's survival and therapeutic target in TNBC (Holland *et al.*, 2013; Jang *et al.*, 2015). The kinetochore is known to play a role in correcting segregation of chromosomes in mitosis, and its aberration causes chromosomal instability. Recent study has reported that a centromere and kinetochore gene can be used as a marker for

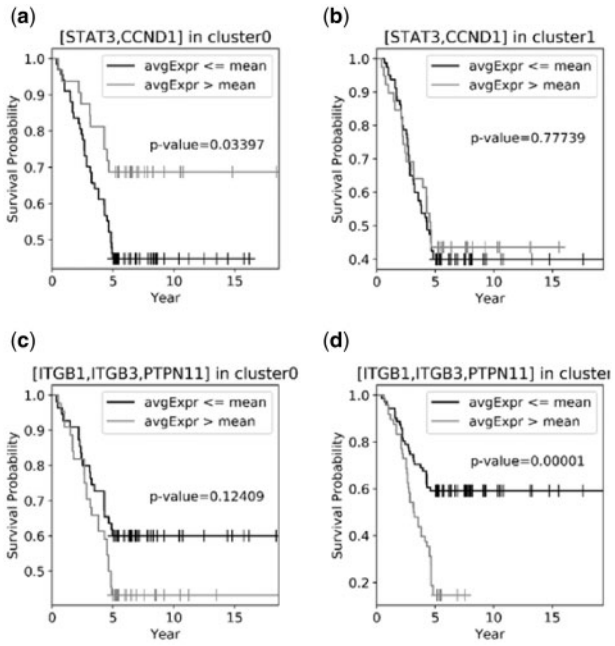

Fig. 9. Kaplan–Meier survival graphs. Kaplan–Meier survival graphs with mean expression of STAT3 and CCND1 for (**a**) cluster 0 and (**b**) cluster 1. Same graphs with mean expression of ITGB1, ITGB3 and PTPN11 for (**c**) cluster 0 and (**d**) cluster1

prognosis and prediction of response to radiotherapy and chemotherapy (Zhang *et al.*, 2016). Cluster 1 has also Interleukin-1 (IL1) mediated pathway. IL1 is a cytokine that involved in cell growth and death. It is reported to promote cancer cell invasion and metastasis in TNBC (Bouchard *et al.*, 2017).

4 Discussion

The proposed approach is capable of the identification of prognostic genes in the form of networks that provide rich information about

molecular functions related to cancer development. More importantly, our approach identifies different prognostic gene networks for heterogeneous patient sample sets, and allowing us to infer different biological processes related to specific groups of samples by analyzing those networks separately. In addition, hub genes of the union of those gene networks were good prognostic biomarkers that showed better prediction performance in terms of AUC than those of existing algorithms.

Existing studies to overcome the cancer heterogeneity problems generally distinguish samples by known cancer subtype like ER+/- or aim to cluster patient groups with good and poor outcome using various omics data. They study subtype specific biological processes and investigate their relationship with cancer outcome. The proposed approach effectively partitions the samples using principal components of gene expression data. Those groups are not necessarily subtype dependent, but we indirectly proved that those reflect breast cancer heterogeneity by showing that resulting cluster-specific gene networks enrich distinguished biological processes or functions, in Figure 8. Moreover, the clustering step helps better outcome prediction, which additionally supports our claim.

In this study, we clustered samples without prior information, such as ER subtypes, because subtype was not strongly correlated with prognosis (Supplementary Table S1a). To cluster with comprehensive information of whole gene expression, we used PCA on whole genes and selected the two best principal components. However, clustering strategies can vary for different datasets. For example, more principal components can be used, or specific groups of genes can be useful to divide samples that show different expression patterns for those genes. In the future, we will apply our approach to various kinds of cancer data, varying clustering strategies.

Funding

This work was supported by Incheon National University (International Cooperative) Research Grant in 2016 [2016-2295]. This research was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [NRF-2016R1D1A1B03934135].

Conflict of Interest: none declared.

References

Abba,M.C. *et al*. (2010) Breast cancer biomarker discovery in the functional genomic age: a systematic review of 42 gene expression signatures. *Biomark. Insights*, **5**, 103–118.

Ahn,J. *et al*. (2011) Integrative gene network construction for predicting a set of complementary prostate cancer genes. *Bioinformatics*, **27**, 1846–1853.

Andò,S. *et al*. (2014) The multifaceted mechanism of leptin signaling within tumor microenvironment in driving breast cancer growth and progression. *Front. Oncol*., **4**, 340.

Barabasi,A.L. *et al*. (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet*., **12**, 56–68.

Barrett,T. *et al*. (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*., **41**, D991–D995.

Beaver,J.A. *et al*. (2013) PIK3CA and AKT1 mutations have distinct effects on sensitivity to targeted pathway inhibitors in an isogenic luminal breast cancer model system. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res*., **19**, 5413–5422.

Broad Institute TCGA Genome Data Analysis Center (2016) Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. Broad Institute of MIT and Harvard.

Bouchard,G. *et al*. (2017) Induction of interleukin-1β by mouse mammary tumor irradiation promotes triple negative breast cancer cells invasion and metastasis development. *Int. J. Radiat. Biol*., **93**, 507–516.

Breiman,L. (2001) Random forests. *Mach. Learn*., **45**, 5–32.

Bullinger,L. *et al*. (2004) Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N. Engl. J. Med*., **350**, 1605–1616.

Burrell,R.A. *et al*. (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**, 338–345.

Buyse,M. *et al*. (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst*., **98**, 1183–1192.

Chuang,H.Y. *et al*. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol*., **3**, 140.

Croft,D. *et al*. (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res*., **42**, D472–D477.

Dao,P. *et al*. (2011) Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, **27**, i205–i213.

Desmedt,C. *et al*. (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res*., **13**, 3207–3214.

Dey,N. *et al*. (2013) Wnt signaling in triple negative breast cancer is associated with metastasis. *BMC Cancer*, **13**, 537.

Fabregat,A. *et al*. (2016) The reactome pathway knowledgebase. *Nucleic Acids Res*., **44**, D481–D487.

Felding-Habermann,B. *et al*. (2001) Integrin activation controls metastasis in human breast cancer. *Proc. Natl. Acad. Sci. USA*, **98**, 1853–1858.

Forbes,S.A. *et al*. (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*., **45**, D777–D783.

Furlong,L.I. (2013) Human diseases through the lens of network biology. *Trends Genet. TIG*, **29**, 150–159.

Hedrick,E. *et al*. (2016) NR4A1 antagonists inhibit beta1-integrin-dependent breast cancer cell migration. *Mol. Cell. Biol*., **36**, 1383–1394.

Heikkinen,T. *et al*. (2011) Variants on the promoter region of PTEN affect breast cancer progression and patient survival. *Breast Cancer Res. BCR*, **13**, R130.

Holland,J.D. *et al*. (2013) Combined Wnt/beta-catenin, Met, and CXCL12/CXCR4 signals characterize basal breast cancer and predict disease outcome. *Cell Rep*., **5**, 1214–1227.

Huang da,W. *et al*. (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*., **37**, 1–13.

Huang da,W. *et al*. (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc*., **4**, 44–57.

Ivshina,A.V. *et al*. (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res*., **66**, 10292–10301.

Jang,G.B. *et al*. (2015) Blockade of Wnt/beta-catenin signaling suppresses breast cancer metastasis by inhibiting CSC-like phenotype. *Sci. Rep*., **5**, 12465.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Lehmann,B.D. *et al*. (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest*., **121**, 2750–2767.

Miller,L.D. *et al*. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. USA*, **102**, 13550–13555.

Niemitz,E. (2013) Ras pathway activation in breast cancer. *Nat. Genet*., **45**, 1273–1273.

Parvani,J.G. *et al*. (2015) Silencing β3 integrin by targeted ECO/siRNA nanoparticles inhibits EMT and metastasis of triple-negative breast cancer. *Cancer Res*., **75**, 2316–2325.

Patil,P. *et al*. (2015) Test set bias affects reproducibility of gene signatures. *Bioinformatics*, **31**, 2318–2323.

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res*., **12**, 2825–2830.

Polyak,K. (2011) Heterogeneity in breast cancer. *J. Clin. Investig*., **121**, 3786–3788.

Ren,W. *et al*. (2016) Protein-protein interaction (PPI) network and significant gene analysis of breast cancer. *Int. J. Clin. Exp. Med*., **9**, 9033–9043.

Roy,J. *et al*. (2014) Network information improves cancer outcome prediction. *Brief. Bioinf*., **15**, 612–625.

Russnes,H.G. *et al*. (2011) Insight into the heterogeneity of breast cancer through next-generation sequencing. *J. Clin. Invest*., **121**, 3810–3818.

Semenza,G.L. (2016) The hypoxic tumor microenvironment: A driving force for breast cancer progression. *Biochimica Et Biophysica Acta*, **1863**, 382–391.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Sohn,I. *et al*. (2009) Gradient lasso for Cox proportional hazards model. *Bioinformatics*, **25**, 1775–1781.

Sotiriou,C. *et al*. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst*., **98**, 262–272.

Strong,A.L. *et al*. (2015) Leptin produced by obese adipose stromal/stem cells enhances proliferation and metastasis of estrogen receptor positive breast cancers. *Breast Cancer Res*., **17**, 112.

Sun,B.Y. *et al*. (2011) Combined feature selection and cancer prognosis using support vector machine regression. *IEEE/ACM Trans. Comput. Biol. Bioinf*., **8**, 1671–1677.

Szeto,C. *et al*. (2017) Investigating tumoral and temporal heterogeneity through comprehensive -omics profiling in patients with metastatic triple negative breast cancer. *J. Clin. Oncol*., **35**, 1093–1093.

van de Vijver,M.J. *et al*. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med*., **347**, 1999–2009.

Vasudevan,K.M. *et al*. (2009) AKT-independent signaling downstream of oncogenic PIK3CA mutations in human cancer. *Cancer Cell*, **16**, 21–32.

Wang,H. *et al*. (2016) Integrating Omic Data with a Multiplex Network-based Approach for the Identification of Cancer Subtypes. *IEEE Trans. Nanobiosci*., **15**, 335–342.

Winter,C. *et al*. (2012) Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput. Biol*., **8**, e1002511.

Wright,K.L. *et al*. (2015) Ras signaling is a key determinant for metastatic dissemination and poor survival of luminal breast cancer patients. *Cancer Res*., **75**, 4960–4972.

Wu,G. *et al*. (2010) A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*., **11**, R53.

Wu,G. and Stein,L. (2012) A network module-based method for identifying cancer prognostic signatures. *Genome Biol*., **13**, R112.

Yan,H. *et al*. (2015) Targeting C-Type Lectin Receptors for Cancer Immunity. *Front. Immunol*., **6**, 408.

Zhang,W. *et al*. (2016) Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nat. Commun*., **7**, 12619.