Published in final edited form as:
Proc Int Joint Conf Bioinforma Syst Biol Intell Comput. 2009 August 3; : 428–434. doi:10.1109/IJCBS.2009.29.

Using Frequent Co-expression Network to Identify Gene Clusters for Breast Cancer Prognosis

Jie Zhang[1,2], **Kun Huang**[1,2], **Yang Xiang**[3], and **Ruoming Jin**[3]

Jie Zhang: jie.zhang@osumc.edu; Ruoming Jin: jin@cs.kent.edu

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA

[2]OSUCCC Biomedical Informatics Shared Resource, The Ohio State University, Columbus, OH, USA

Department of Computer Science, Kent State University, Kent, OH, USA

Abstract

In this paper, we investigated the use of gene co-expression network analyses to identify potential biomarkers for breast carcinoma prognosis. The network mining algorithm CODENSE is used to identify highly connected genome-wide gene co-expression networks among a variety of cancer types, and the resulted gene clusters are applied to a series of breast cancer microarray sets to categorize the patients into different groups. As a result, we have identified a set of genes that are potential biomarkers for breast cancer prognosis which can categorize the patients into two groups with distinct prognosis. We also compared the gene clusters we discovered with gene subsets identified from similar studies using other clustering algorithms.

Keywords

CODENSE; breast cancer prognosis; co-expression network; gene cluster

I. INTRODUCTION

Breast carcinoma is known for its remarkable diversity in terms of the pathology, histology, and prognosis [1]. Since the beginning of breast cancer research, many attempts have been made to categorize this disease and understand the molecular basis for the significant phenotype difference among subgroups. But before the emergence of “molecule profiling” based on the genechip technology, only crude biomarkers have been identified using methods such as immunohistochemical staining, histology, pathological phenotyping, and individual protein characterization. Breast carcinoma was first categorized into ER+ and ER − subgroups, based on the histochemical staining of oestrogen receptor (ER) on the surface of mammary epithelial tumor cells, which was later called basal and luminal breast carcinomas respectively. In fact, it is the first biomarker to be used to characterize tumor cells, and has been a useful biomarker for prognosis for three decades. ER+ and ER− type of breast cancers exhibit fundamental difference in almost every stage of the disease progression, and there have been studies suggesting distinct disease entities and cell lineage among subtypes of breast carcinomas [1–4]. Roughly two-third of the breast carcinomas are ER+, which means a lower rate of cell proliferation, tumor differentiation, recurrence and a longer survival time. ER+ patients are also likely to benefit from tamoxifen and herceptin treatment, as well as hormone therapy [5]. Despite such knowledge, this classification method is still crude, and heterogeneity exists within each subgroup [2, 6, 7]. Clinicians lack a reliable biomarker to predict the prognosis of the patients and their response to chemotherapy [8].

With the rapid development of genome-wide gene expression profiling study, molecule profiling has become a powerful tool to characterize primary tumor subtypes even metastasis, and several studies have discovered subsets of signature genes using molecule profiling, which link to breast carcinoma features such as histological and pathological features, developing stages, and ER status subtypes [2, 3, 6, 9–19]. Among them, the most reliable signature gene subsets are for prediction of ER+ and ER− subtypes[6], while gene patterns related to other features are still imperfect and not stable to became new stand-alone prognosis method for cancer sub-classification[1, 7]. To identify signature gene subsets that can be used to predict breast carcinoma subtypes or its metastasis/survival, a variety of clustering and feature selection algorithms, including PAM, ANN, SAM, CGI, PCA, and hierarchical clustering, has been applied to gene expression profiles, generating signature gene lists from hundreds to thousands [2, 6, 9–19]. However, most of the above algorithms are data-driven and do not directly incorporate the functional relationships among genes in the gene selection process.

Recently, another class of methods focuses on clustering genes with similar expression patterns in multiple samples and identifying the so-called co-expression networks. These co-expression networks can then be used as functional modules in the investigation of disease mechanisms [20–22]. In this paper, we explore the use of the co-expression networks in discovering biomarkers for breast cancer prognosis. Instead of using the co-expression network for each individual study as in [21, 22], we first establish frequent co-expression networks from multiple sets of cancer gene expression data. This approach provides us with a list of “background” co-expression networks that are common in different disease states and reflect the essential functional units in the gene network. In order to achieve this goal, we resort to an established network mining algorithm CODENSE, which efficiently identifies frequent gene co-expression networks [23].

Specifically, we applied the CODENSE algorithm on gene expression profiles from different types of carcinoma cell cDNA microarray studies, identified 44 clusters of genes which are highly connected in the co-expression network, then applied these clusters to a breast cancer microarray study, which consists of 286 breast cancer samples, from 209 ER+ and 77 ER− patients [24]. We isolated differentially co-expression genes clusters correspondent to different ER status co-expression pattern, and identified a group of signature genes that can be indicative of a conceptual network for breast tumor cell. Then we applied one of our gene clusters containing 41 genes to another breast carcinoma microarray studies including 159 tumor samples, and predicted the ER status as well as the disease outcome in term of relapse. We also compared our finding with signature gene lists discovered from other breast carcinoma research using different clustering algorithms, and attempt to understand it in terms of pathway and ontology analysis.

Our work suggests a new approach to select biomarkers for disease. Instead of selecting individual genes by comparing patients with different disease states or subtypes, we treat the gene clusters as a functional units and their responses in different disease states not only help to choose biomarkers but also directly contribute to the gaining of new insights on the molecular mechanisms such as key pathways or gene functions of disease prognosis and subtyping.

II. METHODS

A. Data Selection and PCC Computation

GEO was queried using terms "metastatic cancer”. We further select the datasets (GDS data only) containing both normal and tumor tissues obtained from primary tissue biopsy (cell

lines and secondary cultures were excluded). 23 datasets were selected based on this criterion, 19 were from human, 3 from mouse and 1 from rat tissue (GDS389, GDS505, GDS619, GDS1070, GDS1110, GDS1209, GDS1210, GDS1220, GDS1222, GDS1272, GDS1250, GDS1321, GDS1479, GDS1363, GDS1375, GDS1649, GDS1650, GDS1665, GDS1732, GDS2250, GDS2609, GDS2617, GDS2635). Pearson correlation coefficient (PCC) was calculated for every pair of genes in the expression profile for each of the 23 datasets, using a MATLAB script. In order to identify only the highly connected gene clusters, we retained the pair of highly correlated genes with the absolute values of PCC (|PCC|) of 0.75 or higher. All resulted gene pairs constructed the initial network for CODENSE to identify potential gene clusters.

B. Frequent co-expression network (gene cluster) discovery using CODENSE

The CODENSE algorithm was first developed for discovery networks of genes in multiple microarray datasets and is therefore extremely suitable for our study [23]. We applied the CODENSE algorithm to the gene pairs of 23 datasets of selected cancer microarray studies (see Supplement table for dataset list and reference), and set the parameters so that networks were constructed where each edge linking a pair of genes appeared in at least 4 datasets. The network motifs that had a connectivity ratio $r > 0.4$ (i.e., given a co-expression network with K nodes and L edges, $r = L/(n(n-1)/2)$) were selected for further analyses. The r-value of 0.4 is also the default cutoff value for CODENSE algorithm, which defines a highly connected network motif.

C. Estrogen receptor (ER) status-linked gene cluster identification

The gene clusters found in the above step were applied to a lymph-node negative breast cancer microarray study (GEO dataset GSE2034, using Affymetrix GeneChip Human Genome U133 Array Set HG-U133A), which consists of 286 samples from 209 ER+ patients and 77 ER− patients from age 18–54 [24]. Student's t-tests comparing the two groups of patients were performed and genes with p-values < 0.05 and mean-fold change > 1.5 were selected. For each gene cluster, the percentage of the selected genes in each cluster was computed and Fisher's exact test was carried out to determine the significance of the enrichment of the selected genes in the cluster Bonferroni test is used to compensate for the multiple Fisher's exact tests.

D. Gene ontology (GO) and pathway Analysis

Preliminary GO term enrichment analysis was carried out using the DAVID Bioinformatics Resources (http://david.abcc.ncifcrf.gov/). Further pathway analysis was done using Ingenuity Pathway Analysis (IPA, http://www.ingenuity.com) [25]. Each gene list was annotated using IPA's annotate function and placed in a new pathway and connected under standard settings using the "Build" and "Connect" functions. The threshold for function list and canonical pathway list is determined by IPA using p-value of 0.05, which is shown as the straight orange lines on Fig. 2.

E. Test selected gene cluster on breast cancer datasets (GSE2990 and NKI)

For the genes in the selected cluster, we test them against other public breast carcinoma datasets for their predictive capacity for ER status and survival. Specifically, supervised learning algorithm K-nearest neighbor (KNN) with 20% holdout cross validation (repeat 500 times) was used to test the predictive capacity on the ER status. To test the potential of the selected gene list as prognostic gene markers, we use the select gene list as features and carry out unsupervised clustering on the selected dataset. Then we carry out survival analysis to compare the different groups generated by the clustering. Specifically, we use K-

mean algorithm (K=2, 100 random initialization) for clustering. The survival analysis includes Kaplan-Meijer analysis and log-rank test.

Two sets of large scale study on breast cancer have been chosen to test the gene cluster selected. One is GEO dataset GSE2990 [15] and the other is the Netherlands Kanker Instituut (NKI) NKI-295 dataset with 295 patients (226 ER+ and 69 ER−) [26]. To be consistent with GSE2034 which contains only lymph node negative patients, we selected only the lymph-node negative patients (34 ER− and 113 ER+) patients in GSE2990.

F. Comparison with other gene signature subsets

Eight gene signature subsets from genome-wide gene expression profiling studies on breast carcinoma were compared with genes in cluster 2. Official gene symbols from HUGO were used to query the eight gene lists to find shared genes.

III. RESULTS

A. Identification of frequent co-expression gene clusters

Using the CODENSE algorithm, we identified 44 gene clusters composed of frequently co-expressed genes from the 23 datasets. The sizes of the clusters range from 21 to 74 with an average of 44. The connectivity ratio r ranges from 0.41 to 0.78.

B. Selecting ER status-linked gene cluster

Using Bonferroni test, we set the threshold on the p-value of the Fisher's exact test to be 0.05/44 = 0.00111. As shown in Table 1, eleven genes clusters satisfy this criterion and contain significant numbers of genes that are differentially expressed between patient groups with different ER-status in GSE2034.

Among the eleven clusters, cluster 2 contains the highest ratio (56%) of selected genes and the most significant p-value for the Fisher's exact test. In addition, GO term enrichment analysis shows that it is the only cluster that is not primarily enriched with immune response related genes (Table-1). Therefore, we focus on the cluster 2 in this paper (shown in Fig. 1 Left).

C. Functional and pathway analysis of the signature gene cluster using IPA and literature search

We further conducted IPA analysis on the 41 genes in cluster 2. As shown in Fig. 1 (Right), a highly connected network is formed by a subset of the 41 genes. In addition, the canonical pathways identified from gene lists in cluster 2 and the enriched functional groups/diseases are shown in Fig. 2.

Clearly, the cell cycle related genes form the most enriched functional group. Specifically, many of the genes are related to spindle control and centrosome formation. Among them, HMMR has recently been shown to co-express with the well-known breast cancer related gene BRCA1[27]. In addition, knockout of HMMR or BRCA1 in breast cancer cell lines results in supernumeracy of centrosomes in the cells, which is an abnormality found in breast cancer.

D. Prediction of ER status in other datasets

In order to confirm that our findings are not just by chance, we test the correlation between the expression levels of genes in cluster 2 and the ER status using other datasets. Specifically, we test the predictive capacity of the cluster 2 genes in predicting the ER status

using supervised learning method K-nearest neighbor. For cross validation, we use 20% holdout with 500 repeats. The average accuracy is computed for each gene. In the NKI datasets, 27 genes are present with 30 probesets. The predication accuracies for these 30 the 60 probesets that are present, 47 probesets show prediction accuracies more than 65%. In total, 22 genes show prediction accuracy more than 65% in both datasets. Table-2 shows ten of such genes sorted by their prediction accuracy in the NKI dataset.

E. Selection of gene signatures for survival prediction

We further test if the genes in cluster 2 have predictive capacity for survival. Since the original GSE2034 dataset contains only lymph node negative patients, we test the genes using the lymph node negative patient data in the NKI dataset. Since the NKI dataset only contains 30 probesets for 27 genes out of the 41 genes, we use the 30 probesets as features to cluster the 150 lymph node negative patients into two groups using the K-mean algorithm (cityblock distance metric and 100 times random repeat). This process segments the patients into a good prognosis group (67 patients) and a poor prognosis group (83 patients) (log rank test p value = 2.32×10^{-6}). The Kaplan-Meijer curves of the two groups are shown in Fig. 3.

F. Comparing with other known breast cancer signatures

There are many studies on selecting gene signatures for breast cancer prognosis. By searching the literatures, we found many genes in cluster 2 has been previously included in the various such gene signatures. Some of them were listed in Table-3.

IV. DISCUSSION

Cancer subclassification using gene co-expression profiling has been a rapid growing field in cancer research since the turn of this century. Based on the strategy used, these studies can be divided into three categories, i.e. supervised top-down approach, unsupervised bottom-up approach, and hypothesis-driven approach [1]. The top-down approach uses gene expression data from cohorts of patients with known pathological subtypes or clinical outcomes to identify gene lists associated with those subtypes for prognosis, therefore it is a supervised clustering method. The bottom-up approach is an unsupervised clustering method, using available microarray datasets to search for gene co-expression patterns that can be linked to specific subtypes or clinical outcomes. The hypothesis-driven approach isolates candidate genes from already known biological processes or pathways involved in cancer, then tests them in different cohorts of patients to see their prediction capacity. The hierarchical clustering methods with either top-down or bottom-up approach has been proved not stable, adding more samples may dramatically changes the dendogram [7], while hypothesis-driven gene candidates can be limited by the hypothesis itself and our lack of knowledge of tumor cell physiology and development. Our network-based approach differs from the three above in that we fill the gap among different approaches, and by isolating and applying highly connected gene co-expression network as a functional unit, we achieve a better understanding of tumor cell physiology as well as obtaining signature gene classifiers to predict disease outcome and do subtyping. We pre-identify 44 highly connected gene network clusters from a variety of tumor sample microarray data with a unsupervised network mining algorithm CODENSE, then apply the clusters to cancer studies with known outcome to narrow down to 11 clusters, which are tightly associated with certain cancer subtypes. We further test the predicting power of our candidate cluster (cluster 2) as an intact functional network group as well as each individual gene in ER status subtyping using two different breast carcinoma studies, and test its predicting accuracy on the survival data of different cohorts of patients. The wide range of cancer types of data ensures our gene cluster selection to represent the common traits in the gene co-expression profile among

different types of carcinomas, and this is the major difference between our and other breast carcinoma signature selection methods, with the latter focused entirely on breast carcinoma cases. The resulted clusters clearly group genes into cell proliferation and immune response categories. Not surprisingly, the genes from the best candidate cluster are mostly involved in the cell proliferation, including cell cycle, cellular assembly and organization, DNA replication etc. (Fig. 2), which confirms previous founding that genes from proliferation group forms the universal signature gene subset in the molecule profiling study of breast carcinoma samples [7, 28]. The significant overlapping of cluster 2 genes with signature gene lists from other studies also shows its importance in cancer prognosis (Table-3). Studies have also shown proliferation related genes drives metastasis and relapse in ER+ tumors [7].

BRCA-1 is an important biomarker in breast cancer research. Mutant BRCA-1 carriers usually have high risk of breast cancer and have the worst disease outcome. BRCA-1 protein is a tumor suppressor preventing uncontrollable cell proliferation, and is directly involved in DNA repair [29]. Although it is not found in our gene cluster, several members of cluster 2, including HMMR, AURKA, CENPE, CENPF, UBE2C and UBE2S, are either highly related to BRCA1 or involved in the functions of BRCA1 as a ubiquitin ligase and regulator of centrosome development [27] [29] [30]. Therefore, the discovery of our proliferation cluster also helps in understanding the critical role of BRCA1 as one of the key genetic factors in breast cancer. In addition, our finding is also consistent with one study, which associated cell proliferation signature genes with extremely poor outcome patients in breast cancer [19]. The proliferation cluster is also a good classifier for survival analysis, which clearly distinguishes the good vs. the poor outcome patients. Each gene in this functional network clusters is also proved to have strong prediction power even by itself. When tested to classify ER status, each single gene can predict with the accuracy in a range of 65% to 83.9%. Because the NKI dataset is mixed with lymph-node negative and positive patients, it may influence the prediction accuracy slightly, as studies have shown that lymph-node negative datasets generally can be classified more accurately with gene signatures [1].

Compared with signature genes that can predict ER status very well, genes subsets which can classify chemotherapy sensitivity, metastasis and recurrence are not very reliable. We plan to use the clusters identified in this study to these fields, further testing their prognostic performance.

Studies have demonstrated that simple proliferation-base small gene subsets performs as good as complex large-number gene list [31], and CODENSE cluster mining algorithm provides an excellent way to identify such small gene subsets, each in a functional unit, with a reliable prognostic performance. The genes isolated as a functional network also help us understand the physiology and development of tumor cells. Our successfully identified gene cluster in breast cancer prognosis suggests that the co-expression networks identified in this method can serve as a set of generic building blocks for biomarker selection and gene interaction studies in other cancer or disease scenarios.

REFERENCES

1. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med. 2009 Feb 19.vol. 360:790–800. [PubMed: 19228622]
2. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001 Sep 11.vol. 98:10869–10874. [PubMed: 11553815]

3. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000 Aug 17.vol. 406:747–752. [PubMed: 10963602]
4. Jones C, Mackay A, Grigoriadis A, Cossu A, Reis-Filho JS, Fulford L, Dexter T, Davies S, Bulmer K, Ford E, Parry S, Budroni M, Palmieri G, Neville AM, O'Hare MJ, Lakhani SR. Expression profiling of purified normal human luminal and myoepithelial breast cells: identification of novel prognostic markers for breast cancer. Cancer Res. 2004 May 1.vol. 64:3037–3045. [PubMed: 15126339]
5. Osborne CK. Steroid hormone receptors in breast cancer management. Breast Cancer Res Treat. 1998; vol. 51:227–238.
6. Kristensen VN, Sorlie T, Geisler J, Langerod A, Yoshimura N, Karesen R, Harada N, Lonning PE, Borresen-Dale AL. Gene expression profiling of breast cancer in relation to estrogen receptor status and estrogen-metabolizing enzymes: clinical implications. Clin Cancer Res. 2005 Jan 15.vol. 11:878s–883s. [PubMed: 15701881]
7. Loi S, Piccart M, Sotiriou C. The use of gene-expression profiling to better understand the clinical heterogeneity of estrogen receptor positive breast cancers and tamoxifen response. Crit Rev Oncol Hematol. 2007 Mar.vol. 61:187–194. [PubMed: 17088071]
8. Cleator S, Ashworth A. Molecular profiling of breast cancer: clinical implications. Br J Cancer. 2004 Mar 22.vol. 90:1120–1124. [PubMed: 15026788]
9. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A. 2003 Jul 8.vol. 100:8418–8423. [PubMed: 12829800]
10. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan 31.vol. 415:530–536. [PubMed: 11823860]
11. Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res. 2001 Aug 15.vol. 61:5979–5984. [PubMed: 11507038]
12. Lu X, Lu X, Wang ZC, Iglehart JD, Zhang X, Richardson AL. Predicting features of breast cancer with gene expression patterns. Breast Cancer Res Treat. 2008 Mar.vol. 108:191–201. [PubMed: 18297396]
13. Zhao H, Langerod A, Ji Y, Nowels KW, Nesland JM, Tibshirani R, Bukholm IK, Karesen R, Botstein D, Borresen-Dale AL, Jeffrey SS. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. Mol Biol Cell. 2004 Jun.vol. 15:2523–2536. [PubMed: 15034139]
14. Martin KJ, Patrick DR, Bissell MJ, Fournier MV. Prognostic breast cancer signature identified from 3D culture model accurately predicts clinical outcome across independent datasets. PLoS ONE. 2008; vol. 3:e2994. [PubMed: 18714348]
15. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh J, Piccart M, Delorenzi M. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst. 2006 Feb 15.vol. 98:262–272. [PubMed: 16478745]
16. Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, Lindahl T, Pawitan Y, Hall P, Nordgren H, Wong JE, Liu ET, Bergh J, Kuznetsov VA, Miller LD. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. 2006 Nov 1.vol. 66:10292–10301. [PubMed: 17079448]
17. Farmer P, Bonnefoi H, Anderle P, Cameron D, Wirapati P, Becette V, Andre S, Piccart M, Campone M, Brain E, Macgrogan G, Petit T, Jassem J, Bibeau F, Blot E, Bogaerts J, Aguet M, Bergh J, Iggo R, Delorenzi M. A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer. Nat Med. 2009 Jan.vol. 15:68–74. [PubMed: 19122658]

18. Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, Macgrogan G, Bergh J, Cameron D, Goldstein D, Duss S, Nicoulaz AL, Brisken C, Fiche M, Delorenzi M, Iggo R. Identification of molecular apocrine breast tumours by microarray analysis. Oncogene. 2005 Jul 7.vol. 24:4660–4671. [PubMed: 15897907]
19. Dai H, van't Veer L, Lamb J, He YD, Mao M, Fine BM, Bernards R, van de Vijver M, Deutsch P, Sachs A, Stoughton R, Friend S. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. Cancer Res. 2005 May 15.vol. 65:4059–4066. [PubMed: 15899795]
20. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005; vol. 4 p. Article17.
21. Presson AP, Sobel EM, Papp JC, Suarez CJ, Whistler T, Rajeevan MS, Vernon SD, Horvath S. Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. BMC Syst Biol. 2008; vol. 2:95. [PubMed: 18986552]
22. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet. 2006 Aug 18.vol. 2:e130. [PubMed: 16934000]
23. Hu H, Yan X, Huang Y, Han J, Zhou XJ. Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics. 2005 Jun; vol. 21(Suppl 1):i213–i221. [PubMed: 15961460]
24. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005 Feb 19–25.vol. 365:671–679. [PubMed: 15721472]
25. Ingenuity Pathway Analysis Software. 2008; vol. 2008
26. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002 Dec 19.vol. 347:1999–2009. [PubMed: 12490681]
27. Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, Elshamy WM, Rual JF, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Sole X, Hernandez P, Lazaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, Livingston DM, Gruber SB, Parvin JD, Vidal M. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet. 2007 Nov.vol. 39:1338–1349. [PubMed: 17922014]
28. Desmedt C, Sotiriou C. Proliferation: the most prominent predictor of clinical outcome in breast cancer. Cell Cycle. 2006 Oct.vol. 5:2198–2202. [PubMed: 16969100]
29. Sankaran S, Crone DE, Palazzo RE, Parvin JD. BRCA1 regulates gamma-tubulin binding to centrosomes. Cancer Biol Ther. 2007 Dec.vol. 6:1853–1857. [PubMed: 18087219]
30. Sankaran S, Crone DE, Palazzo RE, Parvin JD. Aurora-A kinase regulates breast cancer associated gene 1 inhibition of centrosome-dependent microtubule nucleation. Cancer Res. 2007 Dec 1.vol. 67:11186–11194. [PubMed: 18056443]
31. Haibe-Kains B, Desmedt C, Piette F, Buyse M, Cardoso F, Van't Veer L, Piccart M, Bontempi G, Sotiriou C. Comparison of prognostic gene expression signatures for breast cancer. BMC Genomics. 2008; vol. 9:394. [PubMed: 18717985]

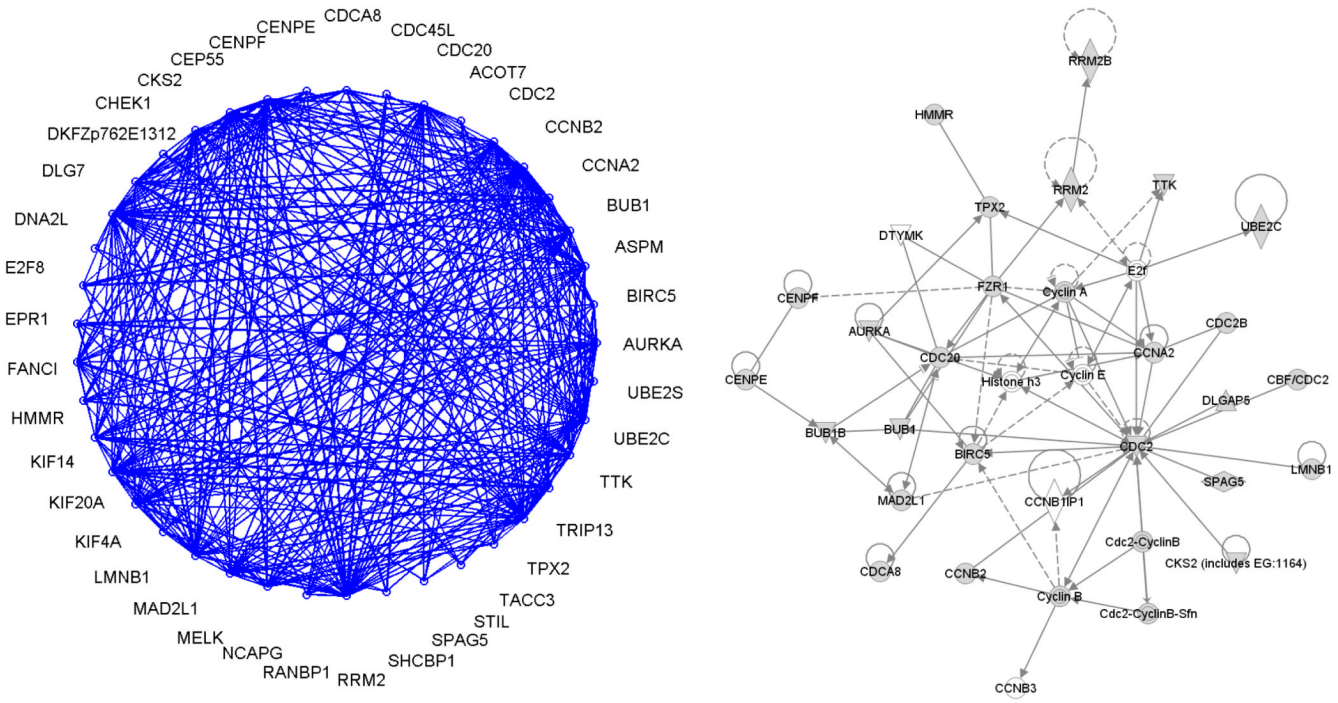

Figure 1.

Left: Cluster 2 with 41 genes showing connections between each other. *Right*: A network identified by IPA from the 41 genes.

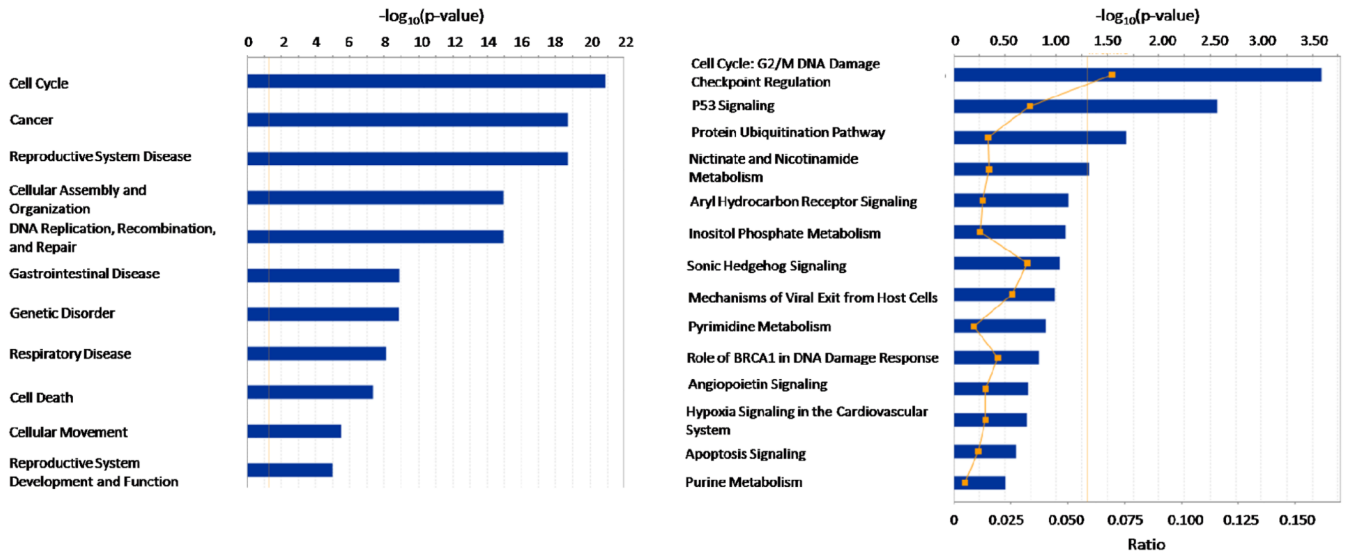

Figure 2. Functional and canonical pathway analyses of genes in cluster 2 using Ingenuity Pathway Analysis

Left: Function list of the top 11 functions carried out by genes in cluster 2, ranked by p-value. *Right:* Canonical pathways with cluster 2 genes involved, ranked by p-value. The straight orange line represents a p-value of 0.05.

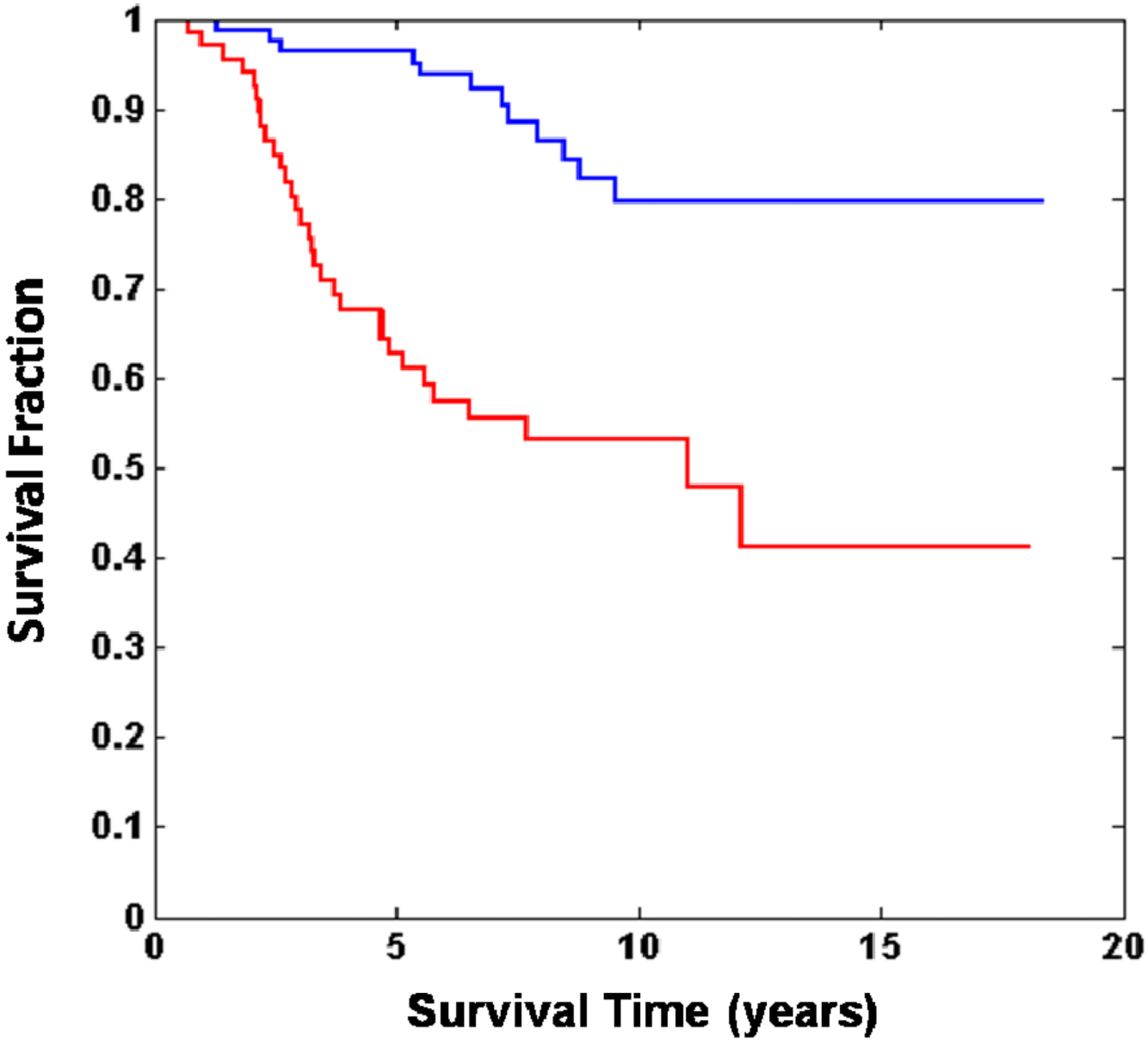

Figure 3. The Kaplan-Meijer curves of the good prognosis group (blue) and the bad prognosis group (red) of lymph-node negative patients on NKI dataset

Table-1

Clusters related to ER status identified from breast cancer microarray study GSE2034

Cluster ID	Total gene number	No. of Genes involved	P-value	Top GO biological process
2	41	23	2.16E-12	M phase of mitotic cell cycle
11	64	19	3.67E-05	Immune response
12	30	9	3.85E-03	Immune response
30	65	25	7.31E-09	Immune response
32	23	8	2.26E-03	Immune response
34	27	14	1.79E-07	Immune response
35	72	23	1.41E-06	Immune response
38	67	23	3.32E-07	Immune response
41	37	13	9.19E-05	Immune response
42	50	15	2.09E-04	Immune response
44	52	22	7.21E-09	Immune response

Table-2

Individual gene prediction accuracy in ER status of different microarray datasets

Gene	Mean accuracy in NKI dataset	Mean accuracy in GSE2990
AURKA	83.9%	69.1%
TTK	77.5%	74.7%
CENPE	76.9%	69.0%
CCNB2	75.0%	73.4%
CCNA2	73.4%	75.1%
BIRC5	72.9%	68.6%
TACC3	72.8%	70.5%
CENPF	72.7%	72.9%
KIF4A	72.5%	70.4%
HMMR	72.0%	67.0%

Table-3

Common genes between cluster 2 and other breast carcinoma gene signature sets:

Reference	Genes also found in cluster 2
Martin, et al. [14]	AURKA, ASPM, CEP55, TRIP13, CKS2, RRM2
Sotiriou, et al. [15]	CDC2, CDC20, CCNB2, CCNA2, BUB1, MELK, BIRC5, ASPM, CDC45L, CDCA8, CENPE, CENPF, CHEK1, AURKA, SCKS2, DKFZp762E1312,DLG7, KIF14, KIF20A, TPX2, HMMR, LMNB1, MAD2L1, RRM2, TACC3, SPAG5, TRIP13, TTK, UBE2C
Dai, et al. [19]	BIRC5, DKFZp762E1312, CCNB2, CDC45L, BUB1, MAD2L1, AURKA
Ivshina, et al. [16]	MELK, TPX2, TTK, CDCA8, CENPE, AURKA
Van ‘t veer, et al. [10]	MELK