Comparisons of cancer classifiers based on RNA_seq and miRNA_seq

Shinuk Kim*

Department of Civil Engineering,
Sangmyung University,
Cheonan 31066, South Korea
Email: kshinuk@smu.ac.kr
*Corresponding author

Hyowon Lee

Department of Biomedical Technology,
Sangmyung University,
Cheonan 31066, South Korea
Email: hw11290711@gmail.com

Mark Kon

Department of Mathematics and Statistics,
Boston University,
Boston, MA 02215, USA
Email: mkon@bu.edu

Abstract: Studies in computational cancer genomics have been faced with the challenge of increasing prediction accuracy of molecular datasets. Here we outline how a feature selection method combined with machine learning may help overcome this challenge for BRCA microRNA-Seq datasets, BRCA RNA-Seq and mRNA microarray datasets, and BLCA microRNA_seq and RNA_seq datasets. We used three different computational approaches: (a) support vector machine, (b) decision tree and (c) *k* nearest neighbours, and two different feature selection methods: (a) Fisher feature criterion and (b) infinite feature selection. Our computational approaches performed consistently better with RNA_seq datasets rather than with miRNA_seq or RNA_array datasets.

Keywords: feature selection; machine learning methods; classification methods; RNA_sequence; miRNA_sequence; breast invasive carcinoma; bladder urothelial carcinoma.

Reference to this paper should be made as follows: Kim, S., Lee, H. and Kon, M. (2017) 'Comparisons of cancer classifiers based on RNA_seq and miRNA_seq', *Int. J. Data Mining and Bioinformatics*, Vol. 17, No. 4, pp.359–368.

Biographical notes: Shinuk Kim received her PhD in Engineering Applied Mathematics from The University of Akron, USA in 2004. She worked at Dr. DeLisi Lap, Bioinformatics, Boston University as a research associate, and now is an Assistant Professor in the Department of Civil Engineering, Sangmyung University, South Korea. Her research interests include bioinformatics, development of machine learning algorithm and systems biology.

Hyowon Lee received her BS in Biomedical Technology from Sangmyung University, South Korea, in 2017. Her research interest includes bioinformatics.

Mark Kon received a PhD in Mathematics from MIT. He has served as departmental director of graduate studies at Boston University, and he is currently affiliated with the Bioinformatics graduate program. His research interests include quantum probability and information, bioinformatics and machine and statistical learning.

1 Introduction

Rapid advancement in biotechnology has raised an urgent need for developing new computational methods with prediction accuracy for the analysis of cancer genomics datasets. Many solutions including subtype classification methods have been developed and implemented for the purpose, yet achieving prediction accuracy for cancer classifications still remains a challenge in many instances.

RNA-Seq data, which unveil transcripts that correspond to existing but to be determined genomic sequences, has emerged as a powerful tool in new biomarker discovery. Compared to microarray, RNA-Seq had technological and practical advantages such as unbiased detection of novel transcripts due to low background noise and unique identification of genomic regions, resulting in a broader dynamic range of genomic profiling with no upper limit (Carter et al., 1990). Thus, RNA-Seq-based gene expression analysis has become routine, which makes RNA-based clinical assays for cancers possible and common.

On the other hand, downstream computational analysis of RNA-seq data remains complex because the technology generates a staggering amount of data. However, machine learning (ML) algorithms have in many cases been developed for the less definitive microarray datasets. Recently researchers have begun to focus on applying such existing algorithms to new datasets (Kim et al., 2015). An example involves the works of Kim (Kim et al., 2014) and others who comparatively studied accuracies of miRNA versus mRNA datasets.

Here, we outline how RNA-seq datasets as new biomarkers are utilised to discriminate tumour tissues from normal ones using existing machine learning algorithms. These algorithms combine the support vector machine (SVM) (Nouretdinov et al., 2011), k-nearest neighbours (kNN) (Weinberger et al, 2006), decision trees (DT) with two feature selection methods; the Fisher score (Fisher), and so-called infinite feature selection (IFS) (Roffo et al., 2015).

2 Materials and methods

2.1 Materials

Breast invasive carcinoma (BRCA) and Bladder urothelial carcinoma (BLCA) datasets were downloaded in June of 2016 from The Cancer Genome Atlas (TCGA), a comprehensive source of cancer datasets, available at http://cancergenome.nih.gov/. We downloaded a total of 832 BRCA miRNA_seq data sets with 1046 genes, which were

generated from Illuminahiseq_miRNAseq platform, and separated them into 87 normal and 745 tumour samples. We also downloaded a total of 1212 BRCA RNA_seq datasets with 20531 genes, which were generated from Illuminahiseq_RNAseq platform, and separated them into 98 normal and 1114 tumour samples. We matched RNA_seq samples with miRNA_seq ones, ending up with 745 tumour and 87 normal samples.

In addition, we also downloaded 533 BRCA microarray tumour and 57 normal samples with 17814 genes, which were generated from Agilentg4502A platform. For BLCA, we downloaded 410 tumour and 19 normal samples with 1046 genes of miRNA_seq and 408 tumour and 19 normal samples with 20531 genes of mRNA_seq datasets. By matching miRNA_seq and RNA_seq, we obtained 405 tumour and 19 normal samples. All datasets were transformed to log2, and genes with zero raw values were eliminated. Consequently, we obtained 12951 RNA_seq and 193 miRNA_seq genes with 745 tumour and 87 normal for BRCA_seq and 12544 RNA_seq and 207 miRNA_seq genes with 405 tumour and 19 normal samples for BLCA. For biomarker discriminators and feature selector, we used Matlab packages, and R-packages (https://cran.r-project.org).

2.2 Methods

We employed two feature selection methods. One is Fisher score, which selects features computed by $(\mu_A - \mu_B)^2/(\sigma_A^2 + \sigma_B^2)$ where μ_A indicates mean score of class A and σ_A indicates standard deviation. The other method is infinite feature selection (IFS), an unsupervised method that computes a central measure on a graph consisting of all possible feature subsets (Roffo et al., 2015). The measurement is $a_{ij} = \alpha\sigma_{ij} + (1-\alpha)c_{ij}$ where $\sigma_{ij} = \max(\sigma(i), \sigma(j))$, $c_{ij} = 1 - |Spearman(f^{(i)}, f^{(j)})|$ f's are features and c_{ij}'s are Spearman ranking correlation. α represents loading coefficients. Then, the common genes obtained from top 20 genes of both Fisher and IFS methods were used to select meaningful biomarkers.

First, with respect to BRCA sequence data sets, we randomly selected 87 samples from 745 tumour samples, matching to normal datasets of 87 samples, and separated these into two groups of size 44 and 43. We randomly separated two groups of size 44 and 43 size out of the normal samples. Consequently, we used 88 samples (44 tumour and 44 normal) as training data and 86 samples (43 tumour and 43 normal) as test data. Second, we selected top 20 genes using either Fisher or IFS methods and verified the test classes using the features. All experiments for discrimination were performed 30 times. In the case of BLCA, the procedures were as the same as BRCA performance, except that training and test sample size are 20 and 18, respectively.

3 Results and discussion

As mentioned above, we selected diagnostic signatures based on two feature selection methods, Fisher score (Fisher) and infinite feature selection (IFS). We tested three machine learning classifiers consisted of support vector machine (SVM), decision tree (DT), and k- nearest neighbours (kNN).

3.1 Comparison of feature selection methods based-on miRNA sequence datasets.

We tested two feature selection methods to obtain meaningful markers. The feature-selected gene sets from each method are presented in Table 1. The two different methods produced three common genes such as let-7c, miR-145, and miR-200c.

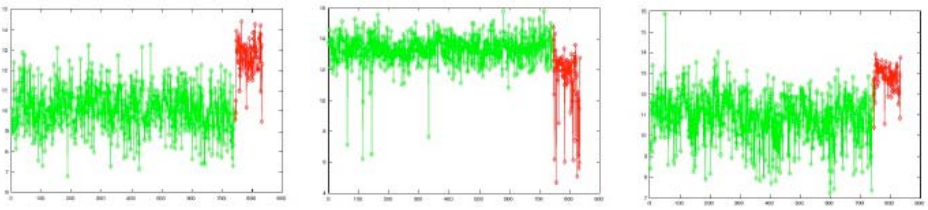
Table 1 A list of genes selected from fisher score (Fisher FS) or infinite feature selection (infinite FS) in miRNA_seq datasets

Fisher FS		*Infinite FS*	
'hsa-mir-10b'	'hsa-mir-96'	'hsa-mir-1307'	'hsa-mir-199a-1'
'hsa-mir-21'	'hsa-mir-183'	'hsa-mir-379'	'**hsa-mir-200c**'
'hsa-mir-139'	'hsa-mir-141'	'hsa-mir-151'	'hsa-mir-203'
'**hsa-mir-145**'	'hsa-mir-182'	'hsa-mir-146b'	'hsa-mir-93'
'hsa-mir-99a'	'hsa-mir-3199-2'	'hsa-mir-181a-1'	'hsa-mir-25'
'hsa-mir-100'	'hsa-mir-28'	'**hsa-let-7c**'	'hsa-mir-451'
'**hsa-let-7c**'	'hsa-mir-497'	'hsa-mir-28'	'hsa-mir-103-1'
'hsa-mir-125b-1'	'hsa-mir-429'	'hsa-mir-150'	'hsa-mir-30d'
'hsa-mir-200a'	'**hsa-mir-200c**'	'**hsa-mir-145**'	'hsa-mir-29a'
'hsa-mir-195'	'hsa-mir-125b-2'	'hsa-mir-142'	'hsa-mir-455'

Notes: Genes in **bold** represents overlapping genes that are commonly selected from both Fisher and Infinite feature selection methods in miRNA datasets.

The common genes were found to be cancer-related genes. Wang (Wang et al., 2009) proposed the role of miR-145 as an inhibitor of breast cancer cell growth. In support of Wang's study, we found that mir-145 expression was greater in normal samples than in tumour ones. Unlike previous studies reporting miR-200c as an inhibitor of breast cancer growth, however, we found that miR-200c expression based on miRNA-seq datasets appeared to be upregulated in activated breast cancer (Figure 1b). Sun (Sun et al., 2016) suggested that let- 7c inhibits the oestrogen receptor of Wnt signalling by binding in the 3 UTR in breast cancer. In an agreement with Sun's argument, we also found that RNA-Seq of let-7c was substantially lower in tumour sample than in normal samples (Figure 1c).

Figure 1 Comparison of miRNA-Seq datasets between tumour samples (green) and normal (red); (a) miRNA-145, (b) miRNA-200c, and (c) let-7c. The x-axis represents samples, and y-axis represents base 2 logarithm of miRNA-seq levels



(a) miRNA-145 (b) miRNA-200c (c) let-7c

In addition, we were interested in exploring common biomarkers from two different datasets. For common biomarkers, we selected top 20 ranked genes generated from Fisher feature selection method based on both RNA_seq and microarray datasets, and marked ten common genes in bold that were found in both datasets, as shown in Table 2. Among them, COL10A1 (Kim et al., 2010), MMP11 (Ma et al., 2009; Paik et al., 2006), NEK2 (Hayward et al., 2004), FIGF (Ma et al., 2009), SDPR (Ma et al., 2009), and SPRY2 (Lo et al., 2004; Qian et al., 2009) were previously identified as breast cancer related genes. LRRC3B (Kim et al., 2008), ADAMS5 (Mochizuki and Okada, 2007), CAV1 (Williams and Lisanti, 2005), CA4 and RP17 (Carter et al., 1990) were identified as oncogenic genes.

Table 2 Top 20 genes selected from Fisher feature selection in RNA_seq or RNA microarray datasets

RNA_seq		*RNA microarray*	
'**MMP11**'	'**SPRY2**'	'**MMP11**'	'WDR51A'
'**COL10A1**'	'PAMR1'	'**COL10A1**'	'**FIGF**'
'**FIGF**'	'**NEK2**'	'**CA4**' **(RP17)**	'IGSF10'
'**CA4**' **(RP17)**	'TMEM220'	'TSLP'	'**SDPR**'
'**ADAMTS5**'	'CPA1'	'**LRRC3B**'	'**NEK2**'
'**SDPR**'	'**CAV1**'	'MME'	'FXYD1'
'CEP68'	'CD300LG'	'**SPRY2**'	'**CAV1**'
'PPP1R12B'	'HIF3A'	'COL11A1'	'KLHL29'
'LRRC3B'	'MESTIT1'	'HOXA4'	'**ADAMTS5'**
'DMD'	'LOC572558'	'CXCL2'	'MAMDC2'

Notes: Genes in **bold** represents overlapping genes that are commonly selected from both RNA_seq and RNA microarray datasets in Fisher feature selection.

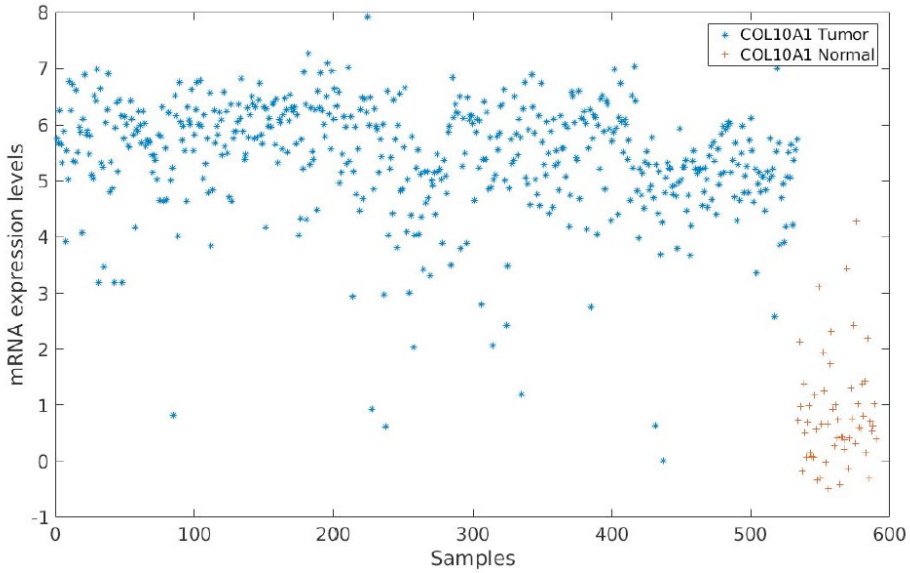
3.2 Computational analysis for BRCA datasets

From 832 RNA_seq datasets, we randomly selected 87 tumour datasets and balanced to 87 normal data sets. Then, each 44 samples out of the tumour and normal datasets, respectively, were selected for training datasets. In test datasets, we randomly selected 43 tumour and 43 normal samples from the remaining data. A base 2-logarithm transformation was applied to normalise the variances of RNA-seq datasets.

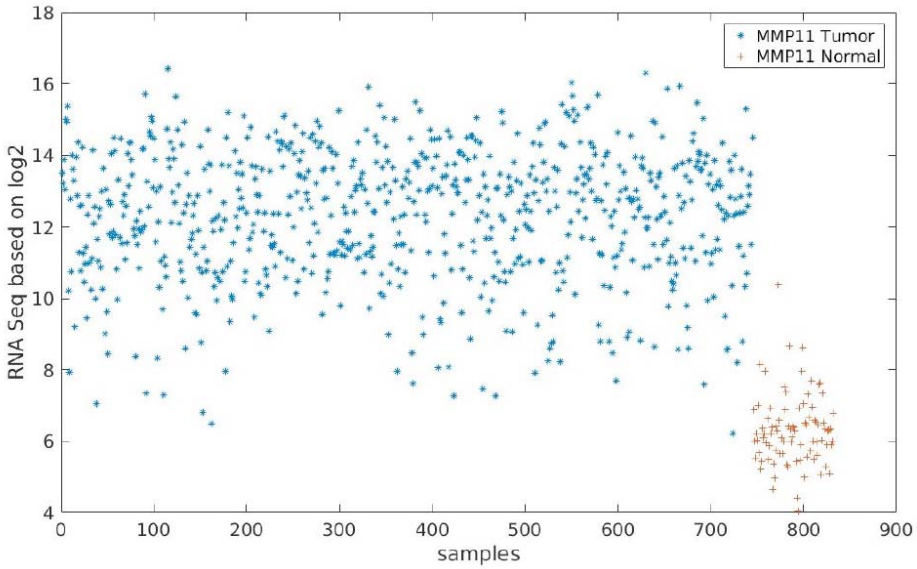
We performed each machine learning method 30 times, with 20 features selection from each time. IFS-based genes selected repeatedly more than 20 times included A1BG, A2LD1, A2M, and MMP11. Fisher method-based genes selected repeatedly more than 20 times included ADAMTS5, CEP68, MMP11, PPP1R12B, and SDPR.

While performing mRNA_arrary datasets 30 times, we selected 20 genes each time. IFS- based common genes included COL10A1, CREB3L1, ELMO2, PNMA1, and RPS11, whereas Fisher-based common genes included CA4, COL10A1, MMP11, and TSLP. COL10A1 was a common gene selected by both feature selection methods based on mRNA_array, whereas MMP11 was a common gene selected by both feature selection methods based on RNA_seq (refer to Figure 2).

Figure 2 (a) A overlapping gene, COL10A1, that are commonly selected from both IFS and Fisher selection methods in mRNA_array datasets. (b) A overlapping gene, MMP11, that are commonly selected from both IFS and Fisher selection in RNA_seq



(a) COL10A1

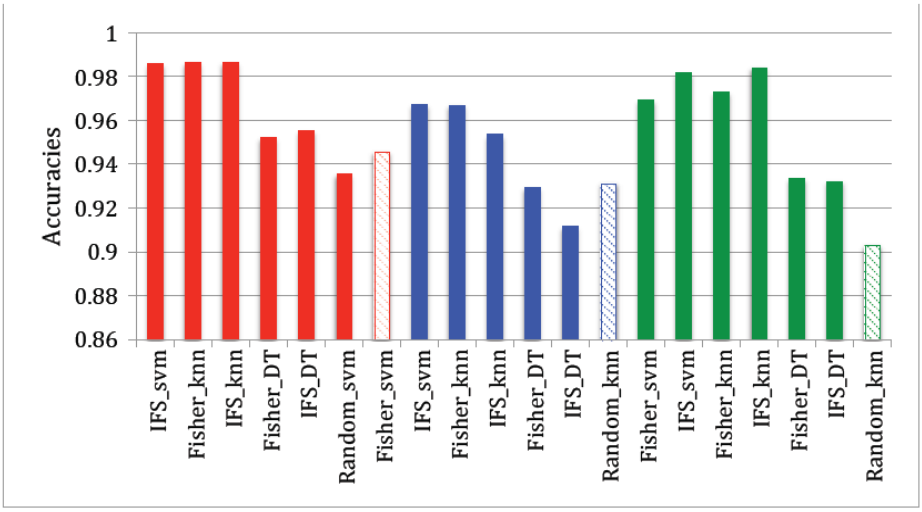


(b) MMP11

Figure 3 represents the results of the feature selection methods and classification methods. Red, blue and green bars indicate RNA_seq, miRNA_seq, and mRNA_array, respectively. The results are expressed as the averages in 30-time performance. To assess

selection accuracy of each performance, leave-one-out-cross validation was performed. Note that each bar indicates feature selection and classifier methods. For example, IFS_SVM represents SVM classification based on infinite feature selection, while Fisher_kNN represents k-nearest neighbour based on Fisher score selection. As implemented, kNN method used a fixed choice of $k = 3$.

Figure 3 Comparison of the accuracies performed by feature selection and machine learning classifiers based on BRCA datasets. Red, blue, and green bars indicate the results of RNA_seq, miRNA_seq and mRNA_array respectively. Oblique pattern bars represent best performance using random feature selection



Limited to the three machine learning methods, DT performance was worse than any of the other methods, as illustrated in Figure 3. Comparing the three different datasets used in both feature selection methiods, RNA_seq datasets performed better than any of the other datasets. RNA_seq datasets were well performed with both SVM and kNN with both feature selection methods. Accuracy performance was better with RNA_seq datasets than with microarray datasets.

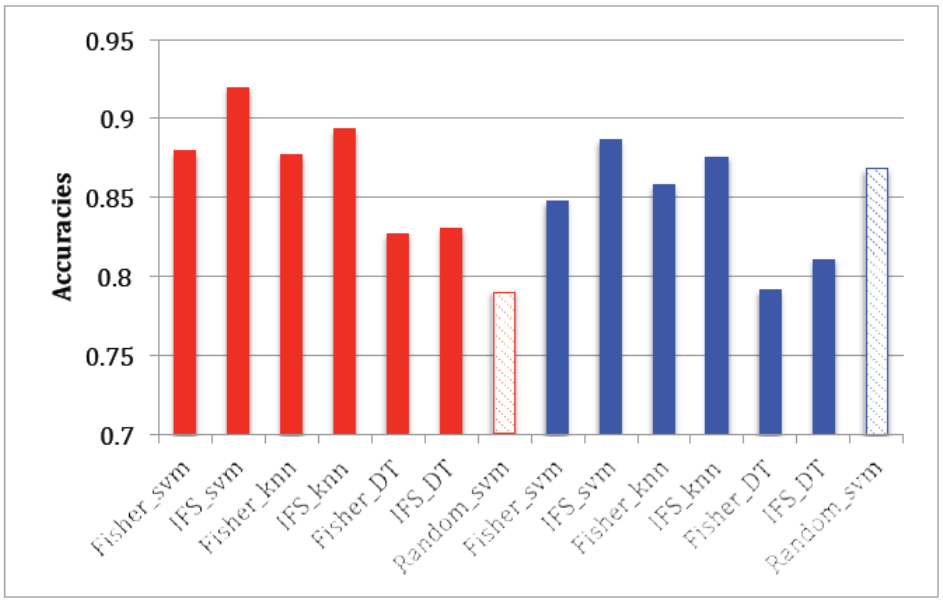
3.3 Computational analysis for BLCA datasets

We tested three machine learning methods with two feature selection methods for BLCA datasets. The results are presented in Figure 4. Among 405 tumour samples and 19 normal samples, we randomly selected 19 tumour samples and assigned to ten training and nine test subjects, balancing to ten training and nine test subjects randomly selected from 19 normal samples.

As results, all machine learning methods performed better with RNA_seq datasets than with miRNA_seq datasets. The genes selected repeatedly more than 20 times while performing Fisher feature selection 30 times included mir-106b, mir-130b, mir-133a-1, mir- 148b, mir-21, mir210, mir-331, mir345, and mir-671. The genes selected repeatedly more than 20 times while performing IFS methods 30 times included let-7a-1(2,3),

let-7b(c,d,e), let-7f-1(2). There were no common genes from random selection. The results of random feature selection were 0.7698 (kNN) and 0.7620 (DT) based on miRNA_seq and 0.7725 (kNN), and 0.6833(DT) based on RNA_seq.

Figure 4 Comparison of RNA_seq and miRNA_seq accuracies performed by feature selection and machine learning algorithms based on BLCA. Red and blue bars indicate the result of RNA_seq and miRNA_seq respectively. Oblique pattern bars represent best performance using random feature selection



4 Conclusions

We studied three classification algorithms of SVM, kNN, and DT combined with two feature selection methods for classifying BRCA and BLCA patients from normal subjects. In both BRCA and BLCA datasets, classification accuracies based on RNA_seq datasets are better and stable than those based on miRNA_seq. DT-based classification performances are worse than those of kNN or SVM for both BRCA and BLCA datasets. In addition, we found ten common genes between microarray and RNA_seq BRCA datasets, and they included six breast cancer-related genes and four cancer-related genes. We also found three common genes selected repeatedly by both Fisher and IFS selection methods, and they included miRNA-145, miRNA-200c, and let-7c. With respect to miRNA-145 and let-7c, we found that their expression levels were downregulated in cancerous tissues as compared to non-cancerous ones, which were in agreement with those of previous findings (Sun et al., 2016; Wang et al., 2009). Interestingly enough, however, we found that miR-200c was upregulated in cancerous tissues as compared to non-cancerous ones, which was contradictory to the findings of Song's study (Song et al., 2015). For this reason, further study will be necessary to resolve the controversy regarding the association direction between this gene and cancer.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2015R1D1A1A01060287) and the Ministry of Science, ICT & Future Planning (NRF-2017R1A2B4010684).

References

Carter, N.D., Fryer, A., Grant, A.G., Hume, R., Strange, R.G. and Wistrand, P.J. (1990) 'Membrane specific carbonic anhydrase (CAIV) expression in human tissues', *Biochim Biophys Acta*, Vol. 1026, No. 1, pp.113–116.

Hayward, D.G., Clarke, R.B., Faragher, A.J., Pillai, M.R., Hagan, I.M. and Fry, A.M. (2004) 'The centrosomal kinase Nek2 displays elevated levels of protein expression in human breast cancer', *Cancer Research*, Vol. 64, No. 20, pp.7370–7376.

Kim, H., Watkinson, J., Varadan, V. and Anastassiou, D. (2010) 'Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1', *BMC Medical Genomics*, Vol. 3, No. 1, pp\.51.

Kim, M., Kim, J.H., Jang, H.R., Kim, H.M., Lee, C.W., Noh, S.M., Song, K.S., Cho, J.S., Jeong, H.Y., Hahn, Y., Yeom, Y.I., Yoo, H.S. and Kim, Y.S. (2008) 'LRRC3B, encoding a leucine-rich repeat-containing protein, is a putative tumor suppressor gene in gastric cancer', *Cancer Res*, Vol. 68, No. 17, pp.7147–7155.

Kim, S., Kon, M. and Kang, H. (2015) A method for generating new datasets based on copy number for cancer analysis', *BioMed Research International*, Vol. 2015, Article ID 467514.

Kim, S., Park, T. and Kon, M. (2014) 'Cancer survival classification using integrated data sets and intermediate information', *Artif IntellMed*, Vol. 62, No. 1, pp.23–31.

Lo, T.L., Yusoff, P., Fong, C.W., Guo, K., McCaw, B.J., Phillips, W.A., Yang, H., Wong, E.S.M., Leong, H.F. and Zeng, Q. (2004) 'The ras/mitogen-activated protein kinase pathway inhibitor and likely tumor suppressor proteins, sprouty 1 and sprouty 2 are deregulated in breast cancer', *Cancer Research*, Vol. 64, No. 17, pp.6127–6136.

Ma, X.J., Dahiya, S., Richardson, E., Erlander, M. and Sgroi, D.C. (2009) 'Gene expression profiling of the tumor microenvironment during breast cancer progression', *Breast Cancer Res*, Vol. 11, No. 1, pp.R7.

Mochizuki, S. and Okada, Y. (2007) 'ADAMs in cancer cell proliferation and progression', *Cancer Sci*, Vol. 98, No. 5, pp.621–628.

Nouretdinov, I., Costafreda, S. G., Gammerman, A., Chervonenkis, A., Vovk, V., Vapnik, V. and Fu, C. H. (2011) 'Machine learning classification with confidence: application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression', *Neuroimage*, Vol. 56, No. 2, pp.809–813.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F.L., Watson, D., Bryant, J., Costantino, J.P., Geyer, C. E., Jr., Wickerham, D. L. and Wolmark, N. (2006) 'Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer', *J Clin Oncol*, Vol. 24, No. 23, pp.3726–3734.

Qian, B., Katsaros, D., Lu, L., Preti, M., Durando, A., Arisio, R., Mu, L. and Yu, H. (2009) 'High miR-21 expression in breast cancer associated with poor disease-free survival in early stage disease and high TGF-beta1', *Breast Cancer Res Treat*, Vol. 117, No. 1, pp.131–140.

Roffo, G., Melzi, S. and Cristani, M. (2015) 'Infinite feature selection', *Proceedings of the IEEE International Conference on Computer Vision.*

Song, C., Liu, L. Z., Pei, X. Q., Liu, X., Yang, L., Ye, F., Xie, X., Chen, J., Tang, H. and Xie, X. (2015) 'miR-200c inhibits breast cancer proliferation by targeting KRAS', *Oncotarget*, Vol. 6, No. 33, pp.34968–34978.

Sun, X., Xu, C., Tang, S. C., Wang, J., Wang, H., Wang, P., Du, N., Qin, S., Li, G., Xu, S., Tao, Z., Liu, D. and Ren, H. (2016) 'Let-7c blocks estrogen-activated Wnt signaling in induction of self-renewal of breast cancer stem cells', *Cancer Gene Ther*, Vol. 23, No. 4, pp.83–89.

Wang, S., Bian, C., Yang, Z., Bo, Y., Li, J., Zeng, L., Zhou, H. and Zhao, R. C. (2009) 'miR- 145 inhibits breast cancer cell growth through RTKN', *Int J Oncol*, Vol. 34, No. 5, pp.1461-1466.

Weinberger, K. Q., Blitzer, J. and Saul, L. (2006) 'Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems*, Vol. 18, pp.1473.

Williams, T.M. and Lisanti, M.P. (2005) 'Caveolin-1 in oncogenic transformation, cancer, and metastasis', *Am J Physiol Cell Physiol*, Vol. 288, No. 3, pp.C494–C506.