Identifying the Relationship between Non-Stem Cancer Cell and Cancer Stem Cell Genes for Breast Cancer: An In-Silico based Approach

Monalisa Mandal
Xavier University, Bhubaneswar
Odisha-752050, India
monalisa@xub.edu.in

Sanjeeb Kumar Sahoo
Institute of Life Sciences,
Bhubaneswar Odisha-751023, India
sanjeeb@ils.res.in

***Abstract*—It has not been established yet that the cancer stem cell genes have any relation with the non-stem cell cancer genes. It could be interesting to find which non-stem cancer genes are more prone to be affected by cancer stem cell genes. In this paper, an in-silico model has been developed to find out the breast cancer non-stem cell genes which have shown a strong relationship with the breast cancer stem cell genes. First, a set of genes (appearing at antecedent) having a high association with the class labels (as consequent) have been computed using the Apriori algorithm. Then, the cancer non-stem cell genes are identified from the association where one of the genes is cancer stem cell gene. Finally, these cancer non-stem cell genes are optimized according to their differential ability. The experiment has been done on a publicly available Breast Cancer data set. The resultant genes are cross-validated, and accuracy has been reported.**

I. Introduction

Cancers, the deadliest disease are most likely curable if it is diagnosed at the earlier stage. The process involves conventional treatments such as surgery, chemotherapy and radiotherapy. But the cancers that are also diagnosed at a later stage have become progressive and metastasize to other organs. In the former case, even after treating cancer at the early stage, some residual cells still persist and sometime later it causes tumour recurrence. Then it becomes more aggressive which leads to metastasis. From the years of the experiment, the researchers concluded that these residual cells which could be found during any stage of cancer progression are the cancer stem cells (CSCs) [1]. These CSCs are regarded as the origin of the diseases which have stem-like properties and are responsible for causing the therapeutic resistance [2], [3]. However, this should not infer that CSCs are originated from normal stem cells. The primary function of a stem cell is self-renewal irrespective of the fact that a cell being normal or malignant. In the process of self-renewal, a stem cell produces one non-CSC (asymmetric division) or two CSCs (symmetric division) that keep holding the self-renewal property again ensuring long-term sustainability [4], [5]. The spontaneous interconversion between non-CSCs and a CSC state seems to be a very rare and slow event, which can be induced by several factors such as the infiltration of inflammatory cells, cytokines, chemokines and hypoxia. Signals from the tumour microenvironment [6], [7] and interactions between cells within the tumour could induce and regulate the level of tumour stemness [8], [9]. Experimental evidence suggests that this reversible transition is, in some cases regulated by the process of epithelial-to-mesenchymal transition (EMT) during tumour progression [10]. Importantly, CSC-rich tumours are also associated with aggressive disease and poor prognosis, indicating that an understanding of CSCs biology is pertinent to developing effective therapies [11]. In CSC paradigm, CSCs are thought to be self-renewing and to reside at the top of the cellular hierarchy [12]. Through asymmetric division and differentiation, these stem cells generate more differentiated progeny that lack self-renewal capacity. However, recent studies also indicated that CSCs can be spontaneously generated from non-stem cancer cells (NSCCs) [10], [13]. The relationships between CSCs and NSCCs [10], [14], [15] have received enormous attention but remain controversial. However, the basis of this phenomenon is not well understood. Therefore, it is believed that if CSCs theory will revolutionize the development of cellular and molecular events during the cancer progression contributing to therapy resistance, recurrence and metastasis [16], [17].

The goal of this study was a bit different than most. Generally, researchers are more interested to find drugs that make tumors grow or shrink. However, what feature genes are more affected by these cells that initiate tumor growth is interest to us [8]–[10], [12]. These feature genes might be more prone to transform into CSCs. In the current study, Apriori association rule mining algorithm has been applied on the breast cancer gene expression data where the class label has been set as RHS of a rule. A set of rules having one of the genes is CSC gene are identified and the associated non-stem cancer genes are collected for the next step. Lastly, by applying t-test, the top ten genes with the least p-values are distinguished and validated using linear Support Vector Machine (SVM).

II. Materials and Methods

A. Association Rules Mining

Mining frequent patterns is an important aspect in Data Mining. Association rule mining [18] is used to find the frequent patterns among the features in a data set [2]. These

rules work on the basis of $if/then$ statements. In association analysis, the antecedent or LHS (if) and consequent or RHS $(then)$ are sets of items (called itemsets) that are disjoint. These statements help to reveal associations between independent data in a database, relational database or other information repositories. These rules are used to identify the association relationships between the objects which are usually used together. An example of a standard association rule have the form $X \rightarrow Y$ which implies if X is true of an instance in a database, so is Y true of the same instance, with a certain level of significance as measured by two indicators, support and confidence. These two measures express the degree of uncertainty about the rule. The goal of standard association rule mining is to output all rules whose support and confidence are respectively above some given support and coverage thresholds [3].

1) Minimum Support Threshold: The support simply determines how often a rule is applicable to a given data set. The support [4] of an association pattern is the percentage of task-relevant data transaction for which the pattern is true. It is a very important measure of the quality of the rule as low support may occur simply by chance. An item set satisfies minimum support if the occurrence frequency of the item set (A set of items) is greater than or equal to minimum support. If an item set satisfies minimum support, then it is a frequent item set.

$$Supp(A \rightarrow B) = Supp\,of\,the\,set(A \cup B)/total\,tuples \quad (1)$$

2) Minimum Confidence Threshold: Confidence, on the other hand, measures the reliability or trustworthiness associated with each discovered pattern. For a given rule $X \rightarrow Y$, the higher the confidence, the more likely it is for Y to be present in transactions that contain X. Confidence also provides an estimate of the conditional probability of Y given X. Confidence [4] is defined as the measure of certainty

$$Conf(A \rightarrow B) = Supp\,of\,the\,set(A \cup B)/Supp\,of\,A \quad (2)$$

3) Lift of a rule: The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. In other words, it can be expressed by the ratio between the rules confidence and the support of the item set in the rule consequence.

$$Lift(A \rightarrow B) = Conf\,of\,the\,set(A \cup B)/Supp\,of\,B \quad (3)$$

B. Apriori Algorithm:

Generating association rules are meaningful in different field of research but how do we generate them? One of the common methods is to bruteforce all possible rules. This is the most inefficient way to generate rules as it is computationally infeasible to compute the support and confidence of all these rules if the data set is medium or large. However, the Apriori algorithm is introduced in [19] for mining frequent item sets and strong association rules in 1994. Apriori algorithm is, the most classical and important algorithm for mining frequent item sets. It reduces the number of candidates by having the property that *if an itemset is frequent, all of its subsets are frequent*. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where k-itemsets are used to explore (k+1) itemsets [5]. The algorithm terminates when no further successful extensions are found. But it has to generate a large amount of candidate itemsets and scans the data as many times as the length of the longest frequent itemsets. The advantage of the algorithm is that before reading the database at every level, it prunes many of the sets which are unlikely to be frequent sets by using the Apriori property, which states that all nonempty subsets of frequent sets must also be frequent. This property belongs to a special category of properties called anti-monotone in the sense that if a set cannot pass a test, all of its supersets will fail the same test as well [8].

Step1	First scan the database once to get frequent 1-itemset
Step2	Repeat
Step3	Generate length$(k+1)$ candidate itemsets from length k frequent itemsets
Step4	Test the candidates against database to find frequent $(k+1)$ itemsets
Step5	Set $k := k+1$
Step6	Until no freaquent itemsets can be generated

End users of ARM encounter problems as the algorithm do not return result in a reasonable time [4]. b. It only tells the presence and absence of an item in transactional database. c. It is not efficient in case of large dataset. d. ARM treats all items in database equally by considering only the presence and absence of an item within the transaction. It does not take into account the significance of item to user or business [4]. Apriori algorithm suffers from some weaknesses in spite of being clear and simple. The main limitation is costly wasting of time to hold vast number of candidate sets with much frequent itemsets, low minimum support or large itemsets. Further, Apriori algorithm also scans the database multiple times to calculate the frequency of the itemsets in k-itemset. So, Apriori algorithm turns out to be very slow and inefficient, especially when memory capacity is limited and the number of transactions is large.

C. Student's t-test:

Student's t-distribution [20], [21] is a continuous probability distribution that is considered when the sample size is less and standard deviation of the population is not known. It is also called t-distribution and estimates the mean of a normally distributed population where samples are drawn from a full population. Normally, for each sample, the t-distribution should be different from each other. But it is noticed that if the sample size increases, t-distribution follows a normal distribution. A t-test is a statistical hypothesis test in which the test statistic follows a Student's t-distribution if the null hypothesis is believed to be true. It is mostly applied to those hypothesis tests which would follow a normal distribution. All such two-sample location tests are usually called Student's t-tests, where the null hypothesis states that the means of two normally distributed populations are equal. Although, the two population variances are also presumed to be equal in t - test. However, when this assumption is dropped from t-test then that form of t-test is called Welch's t-test. Usually, this type of t-test is called as "unpaired" or "independent samples" t-tests because they are applied particularly for those tests where the non-overlapping samples are being compared. Two-

sample t-tests [20] for a data having difference in mean can be categorized into paired or unpaired t-test.

D. Problem Description

A key unresolved issue for cancer biology and therapy is whether the relentless growth of a tumour is driven by most of its cells or, as proposed by the CSC hypothesis, exclusively by a minor subpopulation capable of self-renewal, akin to the numerically rare normal stem cells that maintain tissues (Adams et al, 2007). For more effective cancer therapies, it is critical to determine which cancer cells have the potential to contribute to the disease progression. To this view, anti-tumour treatments are specifically designed to target CSCs, although theoretically unable to cause rapid shrinkage of tumour lesions, it might nonetheless achieve long-term disease eradication by exhausting self-renewal and growth potential of cancer tissues (Dalerba et al, 2007). Our results clarified the impact of CSCs on NSCCs. As the CSC database is still incomplete, the recent study can add some the probable CSC markers. In addition, these findings may correlate with the interconversion between CSCs and NSCCs.

E. Proposed Method

In the first phase, the association between two genes that infer the classes of the samples are investigated. Essentially the rule would look like $Gene_a, Gene_b \leftarrow ClassLabel$. A rule essentially indicates the strongly associated genes that infer the class labels of the samples. Therefore the genes at the left side not only differentiate the different classes of samples but they are also strongly correlated with each other. For each rule, there are rule judging qualifiers such as Support and Confidence. According to the minimum specified support and confidence, the rules are generated. Then the association rules having one CSC gene between the pair of genes presents in the rule are separated to get the set of cancerous genes having a high association with CSC genes. For any association rule to be meaningful, it is critical to have sufficiently high values of support and confidence. The thresholds for the support and the confidence are defined by the user as requires by the experiment. Therefore, different results may be received by assigning different sets of support and confidence. One has to try different settings in order to have a better result. Also, the minimum length of a rule can be the externally provided by the user as per the demand of the experiment. In this paper, the high dimensional breast cancer data is preprocessed by reducing the dimension according to the standard deviation (sd) which denotes the spread of a feature across the samples. For each of the feature gene, the sd is calculated and then these feature genes are sorted in descending order of their sd values. The "knee point" or the "elbow" denoting the maximum difference in between two consecutive sd values. So the reduced data consists of all samples and the feature genes that have larger sd values than the knee point. Next, the Apriori association rule mining algorithm has been applied on the reduced dataset with the set up of minimum rule length = 3, support= 0.3 and confidence = 1. This setup has been experimentally validated and even after setting minimum rule is 3, only rules having a length of 3 are generated by the algorithm. After executing the Apriori, 25765 rules are received. Then the rules that contain one CSC gene at the left side are identified and the list of other genes associated with the CSC genes are separated. Finally, 52 non-CSC cancer genes are the resultant set. Next, unpaired t-test has been applied on the reduced set to get the $p-values$ for each of the gene. So, the top 10 genes are taken from the set of genes which are sorted according to the increasing order of their $p-values$. The proposed method has been depicted in Fig. 1.

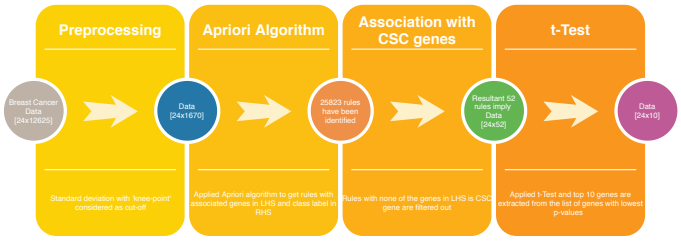

Fig. 1. Flow chart of the proposed method

F. DataSet

Data set name: GSE349_350 This data has been obtained from http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE349, http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE350. The data is being collected from an experiment where patients with breast cancer are given with neoadjuvant docetaxel treatment. In response to treatment, patients can be resistant or sensitive. The samples were obtained before treatment and the tumour response to neoadjuvant treatment was later assessed. The number of samples for resistant to docetaxel treatment (resistant) is 14 and sensitive to docetaxel treatment (sensitive) is 10 that leads 24 samples in the dataset. The total number of genes present in this data is 12625.

G. CSCdb Database

Cancer Stem Cell research is still in its beginning phase. Therefore, the databases of CSC genes are very limited. Moreover, the total number of CSC genes are not very large. Altogether, 1600 CSC genes are collected from the database [13]. One can find other useful information related to CSC related genes.

III. Results and Discussions

The high dimensionality of the caused the Apriori algorithm to hang. So the dimension of the dataset has been reduced by applying standard deviation on to it. The standard deviation of a feature represents how much a feature gene values are spread across the samples and higher the values mean better the feature. Then according to the 'Knee-Point' analysis, the feature gene with larger sd values are kept in the dataset. The sd values for the feature genes has been plotted in Figure 2 and the 'Knee-Point' is 1663 which has been shown in red. Therefore, 1670 feature genes have been considered for the next step. After applying the Apriori algorithm on the reduced dataset, a set of 25823 rules have been achieved. Some of the rules are listed on the Table I. Then from the rules, the non-stem cell genes where one of the genes is CSC gene in LHS of the rule are extracted. So, this subset of non-stem cell

genes does not only imply the classes of the samples but also they have a strong correlation with the CSC genes. Altogether, 52 non-stem cell genes are identified and listed in Table II.

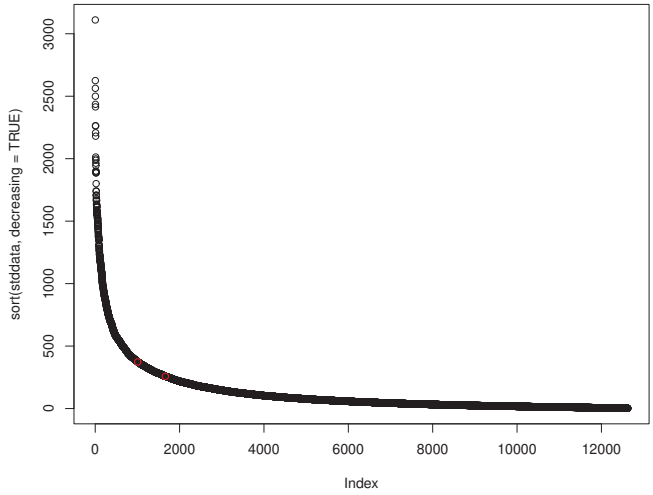

Fig. 2. Standard Deviation

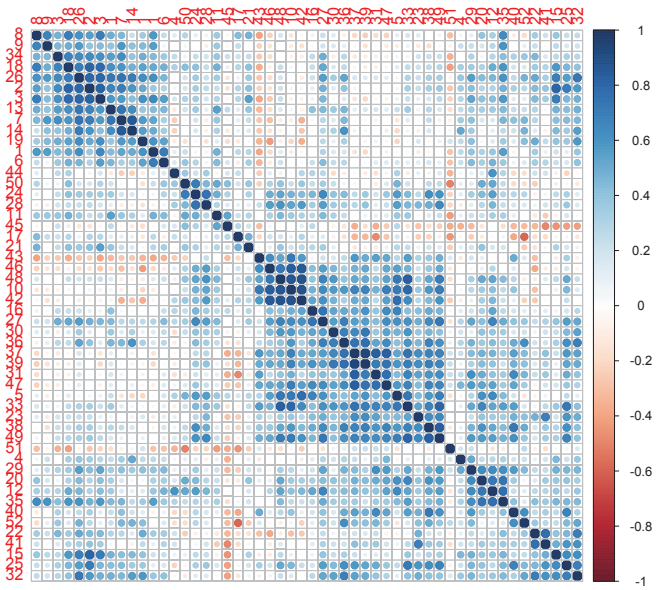

Fig. 3. Correlation Plot

TABLE I. TOP 10 RULES ACCORDING TO SUPPORT

lhs Imply	rhs	support	confidence	lift
{X33458_r_at,X37350_at} =>	{class}	0.4167	0.9951	1.568
{X36991_at,X41179_at} =>	{class}	0.375	1	1.7141
{X36991_at,X39329_at} =>	{class}	0.375	1	1.7141
{X36991_at,X34368_at} =>	{class}	0.375	1	1.7141
{X36991_at,X32803_at} =>	{class}	0.375	1	1.7141
{X33820_g_at,X40888_f_at} =>	{class}	0.375	0.9	1.5428
{X31863_at,X437_at} =>	{class}	0.375	0.9	1.5428
{X32530_at,X437_at} =>	{class}	0.375	0.8181	1.4026
{X36608_at,X437_at} =>	{class}	0.375	0.9	1.5428
{X41147_at,X437_at} =>	{class}	0.375	0.8181	1.4026

The correlation plot of these 52 genes has been shown in Figure 4 and it displayed the results of a correlation by cluster analysis. The blue and red colour represents positive and negative correlation respectively. Darker the blue/red colour means higher the positive/negative correlation respectively. The image shows two large subset of positively correlated genes and a few negatively correlated genes.

The heatmap of the 52 non-stem cell genes is plotted in Figure 3. The level of expression has been expressed in terms of colour and the range of the colour changes from yellow to red as the expression level changes. It is clear from the plot that the expression levels across the samples are very well defined, meaning all the genes have different expression levels in two classes of samples. As this data set contains 10 samples from one class and 14 samples from others, therefore the image is clearly able to differentiate the two classes of samples.

Next, on the subset of 52 genes, t-test has been applied. The p-values for each of the feature gene has been calculated by the unpaired t-test. After sorting the p-values in increasing of their order, the top 10 genes have been separated. From starting to till this point, four datasets has been identified with reduced di-

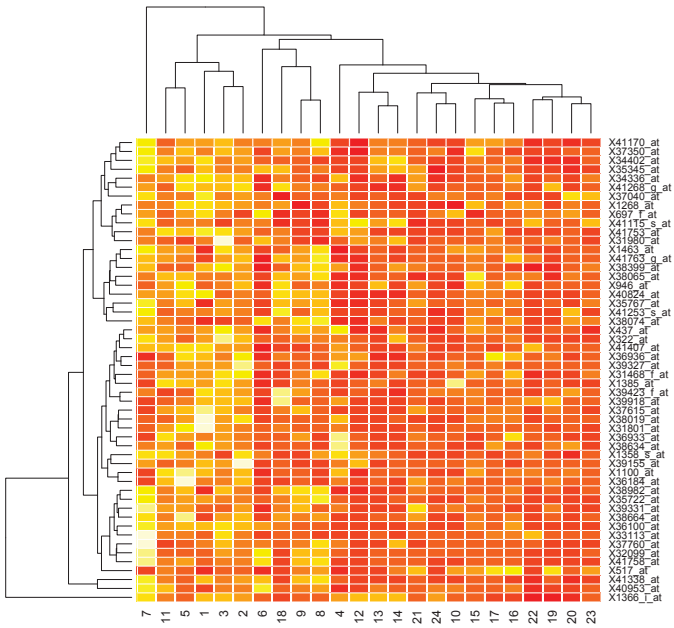

Fig. 4. Heatmap

mension in each stage. At the beginning of the experiment the dimensionality was $D_{24\times12625}$, then after applying standard deviation, the dimensionality becomes $D_{24\times1670}$, then Apriori algorithm has been applied on the reduced set and the resultant reduced dataset have the dimensionality of $D_{24\times52}$ and lastly, t-test has been applied to get the top 10 genes having least p-

values and the dimensionality becomes $D_{24\times10}$. In each stage, the dataset has been evaluated using linear SVM for calculating the classification accuracy. In this classification procedure, at first, the dataset has been divided into two subsets viz., training and testing. On the training dataset, 10-fold cross-validation has been performed with linear SVM and the resultant average accuracy has been reported in Table III. Then, the testing dataset is tested on the trained SVM and sensitivity, specificity, accuracy have been calculated and described in the same table. From Table III, it is evident that the average cross-validation accuracy for the dataset $D_{24\times12625}$ is 71.67 but for the dataset $D_{24\times1670}$ it becomes 68.33. However, the reduced dataset $D_{24\times52}$ and $D_{24\times10}$ gives better cross-validation accuracy than the dataset with a high dimension. In train-test model, the sensitivity and accuracy are gradually improving with the reduction of the dimension o the dataset. Such as the sensitivities are 61.54, 63.33, 67.68 and 1 for the datasets $D_{24\times12625}$, $D_{24\times1670}$, $D_{24\times52}$ and $D_{24\times10}$. Similarly, the accuracies in train-test model are 67.78, 85.54, 96.88 and 1 for the datasets $D_{24\times12625}$, $D_{24\times1670}$, $D_{24\times52}$ and $D_{24\times10}$. Although, the specificity for the dataset with dimension $D_{24\times12625}$ is 97.89, but for the rest of the cases it is 1. So, overall the dataset with top ten non-stem cancer cell genes performed better than the datasets with higher dimensions with respect classification ability. Also, it is should be noted that these top genes are also in the list of non-stem cell genes that are strongly connected with CSC genes.

IV. Conclusions

In this paper, we tried to find out the non-stem cancer cell genes which not only shows different expression level in two classes of samples but also, they have a strong association with the available CSC genes. For this study, two publicly available databases have been used namely, NCBI and CSCdb. Apriori association rule mining algorithm has been applied on the breast cancer gene expression data where the class label has been set as RHS of a rule. Also, other parameters such as minimum rule length, minimum support and confidence are selected experimentally. As a result, a set of rules have been identified with one of the genes in LHS of the rules is CSC gene. From these rules, non-stem cancer genes are collected for the next step. Lastly, by applying t-test, the top ten genes with the least p-values are distinguished and validated using linear SVM. Several performance metrics viz., sensitivity, specificity and accuracy have been reported along with the heatmap and correlation plot.

Acknowledgement

The work is supported by TARE scheme of DST-SERB, Govt. of India.

References

[1] G. Shukla, H. Khera, A. Srivastava, P. Khare, R. Patidar, and R. Saxena, "Therapeutic potential, challenges and future perspective of cancer stem cells in translational oncology: A critical review," *Current Stem Cell Research and Therapy*, vol. 12, no. 3, pp. 207–224, 2017.

[2] P. Dandawate, D. Subramaniam, R. Jensen, and S. Anant, "Targeting cancer stem cells and signalling pathways by phytochemicals: Novel approach for breast cancer therapy," *Seminars in Cancer Biology*, vol. 40-41, pp. 192–208, 2004.

TABLE II. 52 Gene symbols with their names

Gene Symbol	Gene Name
PLOD1	procollagen-lysine,2-oxoglutarate 5-dioxygenase 1(PLOD1)
IFI6	interferon alpha inducible protein 6(IFI6)
GRM1	glutamate metabotropic receptor 1(GRM1)
UPF2	UPF2 regulator of nonsense transcripts homolog (yeast)(UPF2)
PSMD10	proteasome 26S subunit, non-ATPase 10(PSMD10)
PXDN	peroxidasin(PXDN)
PSMD3	proteasome 26S subunit, non-ATPase 3(PSMD3)
MMP3	matrix metallopeptidase 3(MMP3)
BAIAP3	BAI1 associated protein 3(BAIAP3)
TGFBI	transforming growth factor beta induced(TGFBI)
CNN3	calponin 3(CNN3)
FBXW4P1	F-box and WD repeat domain containing 4 pseudogene 1(FBXW4P1)
TMF1	TATA element modulatory factor 1(TMF1)
PIK3R3	phosphoinositide-3-kinase regulatory subunit 3(PIK3R3)
PTPN12	protein tyrosine phosphatase, non-receptor type 12(PTPN12)
TCF25	transcription factor 25(TCF25)
TMEM184B	transmembrane protein 184B(TMEM184B)
MIR1236	microRNA 1236(MIR1236)
UBA1	ubiquitin like modifier activating enzyme 1(UBA1)
TUBB2A	tubulin beta 2A class IIa(TUBB2A)
CITED2	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2(CITED2)
STRAP	serine/threonine kinase receptor associated protein(STRAP)
AP3S1	adaptor related protein complex 3 sigma 1 subunit(AP3S1)
TUBGCP2	tubulin gamma complex associated protein 2(TUBGCP2)
SNRPB2	small nuclear ribonucleoprotein polypeptide B2(SNRPB2)
LOC400927-CSNK1E	LOC400927-CSNK1E readthrough(LOC400927-CSNK1E)
TIAL1	TIA1 cytotoxic granule associated RNA binding protein like 1(TIAL1)
HMGB2	high mobility group box 2(HMGB2)
TSTA3	tissue specific transplantation antigen P35B(TSTA3)
CGB1	chorionic gonadotropin beta subunit 1(CGB1)
GRB10	growth factor receptor bound protein 10(GRB10)
UBC	ubiquitin C(UBC)
CFDP1	craniofacial development protein 1(CFDP1)
RBP1	retinol binding protein 1(RBP1)
KARS	lysyl-tRNA synthetase(KARS)
spop	speckle type BTB/POZ protein(SPOP)
XPO7	exportin 7(XPO7)
DYNLT1	dynein light chain Tctex-type 1(DYNLT1)
IRAK1	interleukin 1 receptor associated kinase 1(IRAK1)
AES	amino-terminal enhancer of split(AES)
plekhb2	pleckstrin homology domain containing B2(PLEKHB2)
SAFB2	scaffold attachment factor B2(SAFB2)
LOC441155	zinc finger CCCH-type domain-containing-like(LOC441155)
VEGFA	vascular endothelial growth factor A(VEGFA)
TERF2IP	TERF2 interacting protein(TERF2IP)
HMGCS2	3-hydroxy-3-methylglutaryl-CoA synthase 2(HMGCS2)
FOXN1	forkhead box N1(FOXN1)
GABARAPL2	GABA type A receptor associated protein like 2(GABARAPL2)
BAIAP2	BAI1 associated protein 2(BAIAP2)
ACTN4	actinin alpha 4(ACTN4)
NDRG1	N-myc downstream regulated 1(NDRG1)
GANAB	glucosidase II alpha subunit(GANAB)

TABLE III. PERFORMANCES OF THE DATASETS WITH REDUCED DIMENSIONS IN DIFFERENT STAGES

Data set with Varying Dim.	Performance of Train-Test Model			Avg. Cross Val. Accuracy
	Sensitivity	*Specificity*	*Accuracy*	
$D_{24\times12625}$	61.54	97.89	67.78	71.67
$D_{24\times1670}$	63.33	1	85.54	68.33
$D_{24\times52}$	67.78	1	96.88	78.78
$D_{24\times10}$	1	1	1	78.78

[3] J. Moselhy, S. Srinivasan, M. ANKEM, and C. Damodatan, "Natural products that target cancer stem cells," *Anticancer Research*, vol. 35, pp. 5773–5788, 2015.

[4] S. K. Singh, I. D. Clarke, and M. T. et al., "Identification of a cancer stem cell in human brain tumors," *Cancer Research*, vol. 63, no. 18, 2003.

[5] A. Eramo, F. Lotti, and G. S. et al., "Identification and expansion of the tumorigenic lung cancer stem cell population," *Cell Death and Differentiation*, vol. 15, no. 3, 2008.

[6] Z. R. ZA and W. Matsui, "Csc cells that express stem cell marker genes, including oct4, sox2, nanog, c-kit, abcg2, and aldh()," *Gastroenterol Hepatol.*, vol. 15, no. 8, 2012.

[7] Z. F. Yang, D. W. Ho, and M. N. N. et al., "Significance of cd90+ cancer stem cells in human liver cancer," *Cancer Cell*, vol. 13, no. 2, 2008.

[8] M. W. C. e. a. S. Zhang, C. Balch, "Identification and characterization of ovarian cancer-initiating cells from primary human tumors," *Cancer Research*, vol. 68, no. 11, pp. 4311–4320, 2008.

[9] Y. Xiao, M. Lin, X. Jiang, J. Ye, T. Guo, Y. Shi, and X. Bian, "The recent advances on liver cancer stem cells: Biomarkers, separation, and therapy," *Analytical cellular pathology*, vol. 2017, p. doi:10.1155/2017/5108653, 2017.

[10] G. Yang, Y. Quan, and W. W. et al., "Dynamic equilibrium between cancer stem cells and non-stem cancer cells in human sw620 and mcf-7 cancer cell populations," *British Journal of Cancer*, vol. 106, no. 9, pp. 1512–1519, 2012.

[11] V. Mayank, "Molecular docking study of natural alkaloids as multi-targeted hedgehog pathway inhibitors in cancer stem cell therapy," *Computational Biology and Chemistry*, vol. 62, pp. 145–154, 2016.

[12] L. P. Chaitra, A. Prashant, C. Gowthami, B. Hajira, M. Suma, S. S. Mahesh, G. Manjunath, and C. Sheeladevi, "Detection of cancer stem cell related markers in different stages of colorectal carcinoma patients of indian origin by immuno histochemistry," *Genomics*, vol. 15, no. 1, pp. 75–81, 2019.

[13] S. Achuthan, T. Santhoshkumar, J. Prabhakar, S. Nair, and M. Pillai, "Drug-induced senescence generates chemoresistant stemlike cells with low reactive oxygen species," *Journal of Biological Chemistry*, vol. 286, pp. 37 813–37 829, 2011.

[14] S. P. and S. Chakraborty and RK. Maji and Z. Ghosh, "Elucidating the gene regulatory networks modulating cancer stem cells and non-stem cancer cells in high grade serous ovarian cancer," *Genomics*, vol. 111, no. 1, pp. 689–692, 2019.

[15] C. A. O'Brien, A. Pollett, S. Gallinger, and J. E. Dick, "A human colon cancer cell capable of initiating tumour growth in immunodeficient mice," *Nature*, vol. 445, no. 7123, 2007.

[16] A. Liskova, P. Kubatka, M. Samec, P. Zubor, M. Mlyncek, T. Bielik, S. M. Samuel, A. Zulli, T. K. Kwon, and D. Büsselberg, "Dietary phytochemicals targeting cancer stem cells," *Molecules*, vol. 24(5), p. 899, 2019.

[17] M. Chan, R. Chen, and D. Fong, "Targeting cancer stem cells with dietary phytochemical - repositioned drug combinations," *Cancer Letters*, vol. 433, pp. 53–64, 2018.

[18] A. K. C. Wong and W. Yang, "High-order pattern discovery from discrete-valued data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, no. 6, 1997.

[19] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *Proceedings of the 20th International Conference on Very Large Data Bases*, vol. VLDB, pp. 487–499, 1994.

[20] W. A. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments." *Bioinformatics*, vol. 18, no. 4, 2002.

[21] R. Mankiewicz, "The story of mathematics." *Princeton University Press*, 2004.