Detecting Long Non-Coding RNAs Responsible for Cancer Development

1st Mitra Datta Ganapaneni
Dept. of CSE.
SRM University,AP.
Andhra Pradesh, India
mitradatta_g@srmap.edu.in

2st Kundhana Harshitha Paruchuru
Dept. of CSE.
SRM University,AP.
Andhra Pradesh, India
kundhanaharshitha_paruchuru@srmap.edu.in

3stJaya Harshith Ambati
Dept. of CSE.
SRM University,AP.
Andhra Pradesh, India
harshith_ambati@srmap.edu.in

4st Mahesh Valavala
Dept. of CSE.
SRM University,AP.
Andhra Pradesh, India
mahesh_valavala@srmap.edu.in

5st Dr Sobin C.C
Dept. of CSE.
SRM University,AP.
Andhra Pradesh, India
sobin.c@srmap.edu.in

2022 OITS International Conference on Information Technology (OCIT) | 978-1-6654-9348-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/OCIT56763.2022.00040

***Abstract*—Long noncoding RNAs (lncRNA) have a vital role in tumor development. Variation in expressions of lncRNAs affect several target genes related to tumor initiation and development. Recent studies in Carcinogenesis have indicated the importance of lncRNA in cancer progression, diagnosis, and treatment. The purpose of our research is to identify the key cancer-related lncRNAs. It is considered a complex task to identify key lncRNAs in cancer with existing cancer data of tumor patients due to the high dimensionality nature of expression profiles. LncRNA expression profiles of 12309 lncRNAs and 2221 patients are gathered from TCGA. A Computational framework is proposed considering 5 cancer types (Bladder, Colon, Cervical, Liver, Head, and Neck) comprising four Machine learning classification models named K-Nearest Neighbor, Naive Bayes, Random Forest,and Support Vector Machine. An essential component in the framework is to use models along with the state-of-the-art Variance threshold, L1-based, and Tree-based feature selection algorithms for differential analysis. The study resulted in identifying 234 key lncRNAs capable of differentiating 5 cancer types. The capability of identified key lncRNAs is observed by the performance of classification models resulting in the highest 98.2% accuracy by SVM. Furthermore, the correlation analysis of 234 lncRNAs experimentally validated the results.**

***Index Terms*—lncRNA, Carcinogenesis, Expression Profiles, TCGA, UCSC-Xena, TANRIC, Machine learning, BLCA, COAD, CESC, HNSC, LIHC, SVM, Correlation Analysis**

I. Introduction

Cancer is a lethal disease with increasing fatality and morbidity rates [1]. Over the last few years, despite the significant advancements in cancer treatment, there are still some issues that need to be ameliorated, like inaccurate prognosis and slow diagnosis. Abnormal cell proliferation is the biological consequence of the disease. The spread of tumor cells to the neighboring tissues and organs is the main factor of morbidity and mortality in many cancer patients [2]. Cancers are classified by the type of cell from which they originate [3]. As a result, non-coding regions should receive greater attention than coding regions, as they contain more cancer mutations.The role of long non-coding RNAs (lncRNAs) in various biological pathways is becoming more and more important [22].They are anticipated to play a role in various diseases as well as developmental processes [3]. LncRNAs can operate as signal or scaffold lncRNAs to communicate with the chromatin remodeling machinery in a variety of ways. By interacting with proteins and controlling their activity, lncRNA functions as a scaffold [10]. In response to various stimuli, signal lncRNAs function as a molecular signal to control transcription [11]. These mechanisms are shown in Fig 1 from [25]. The only lncRNA discovered and verified till today is PCA3, a biomarker for diagnosis of prostate cancer from an early stage [7].For cancer diagnosis and treatment, lncRNA expression profiles provide several new opportunities and suggest eventual developments. Low throughput techniques that produce large sequence coverage per cell are used to detect lncRNAs at the single-cell level [4].

In this paper, we proposed a set of methods which uses machine learning to recognize cancer-related significant lncRNAs by using the expression profiles of several lncRNAs of various cancer patients. We have extracted lncRNA expression profiles of 2221 patients related to 5 different cancer types (BLCA, CESC, COAD, HNSC, LIHC). Applied feature selection algorithms resulting in a set of lncRNAs. Machine learning Algorithms are applied to perform Multi-Class Cancer Classification and compare the performance of results and to validate the identified "cancer-related" lncRNAs.

II. Background

In earlier days, protein-coding genes through the central dogma of DNA had more importance for gene regulation in biology. The number of distinct types of lncRNAs ,their mechanisms were less known. Long Non Coding RNAs have a considerable amount of ability to function as essential genes to identify cancer. Exhibiting some characteristics in the presence of diverse cancers by these genes had visible positive effects on several stages of cancer treatment. lncRNAs are

978-1-6654-9348-2/22/$31.00 ©2022 IEEE
DOI 10.1109/OCIT56763.2022.00040

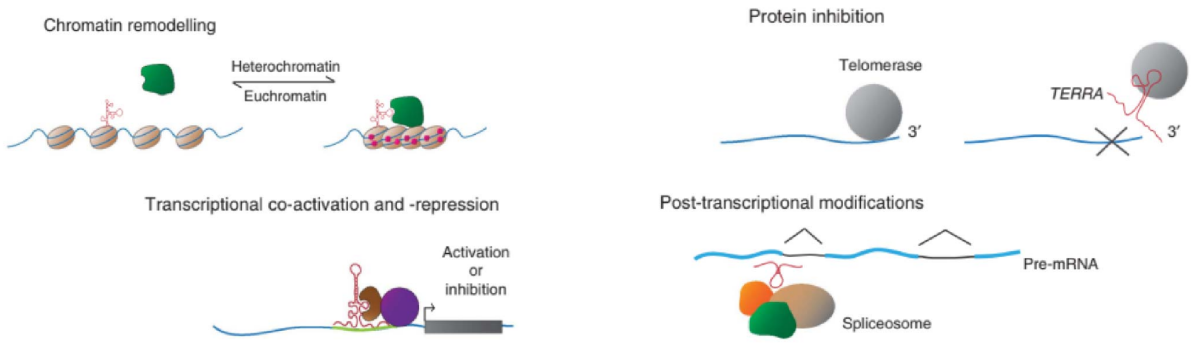


Fig. 1: Represents the lncRNA mechanisms

prospective cancer therapeutic targets due to their involvement in a range of cell activities and unique expressions [22]. Non-coding RNAs which are most likely to stay inside the nucleus post transcription and contain more than 200 nucleotides are LongNon-Coding RNAs [22]. Few curative diagnostics exist currently that target IncRNAs, which are important and associated with cancer [22]. Identification of association between IncRNA and cancer is crucial for effectively identifying signature molecules and creating cancer therapy regimens [22]. The role of LncRNAs in cancer is found in proliferation, motility, growth suppression, immortality, angiogenesis, and viability. An oncogenic lncRNA causes cancer. In tumor cells, expression values are highly expressed. Tumor suppressor lncRNAs are down-regulated in the cancer cells.

Using RNA-seq, transcriptomes are made from tiny amounts of initial material with higher transcript identification. [22]. Application and Usage of RNA-seq is found in various scenarios. This method is used for prediction of expression of different transcripts, genes and also helps in improving studies related to quantitative study of gene expressions [10]. RNA sequencing is an experimental technique used to get the lncRNA expression profiles.

III. Related Work

The aberrant expression of oncogenic lncRNA in tissues leads to cancer progression [15]. In the case of cancer diagnosis, many authors summarized that finding a lncRNA is either a tumor suppressor or oncogenic plays a significant role. Wetlab lncRNA experiments get the greatest results, but they are too costly and time-consuming. To further explore cancer-related lncRNAs, functional annotation of lncRNAs is required, which is more arduous. An uncomplicated solution for this is using in-silico techniques which are economical and time-saving [19]. Over-expression or under-expression of lncRNA can diagnose cancer [16].

The normal examination of lncRNAs differential expression in malignant and normal tissue does not support the necessary conditions for prediction [8].To distinguish cancer-related lncRNAs from tumor suppressor lncRNAs, characteristics such as co-expression relationship, exon mutation frequency, genomic location tissue specialty, and, somatic single nucleotide variations between lncRNAs and protein-coding genes were pooled together [8]. But grouping these features takes a lot of time. For this purpose, a model (CRlncRC) was developed with a combination of network, epigenetic, and gene expression features. It used machine learning to group Long noncoding RNA according to their correlation with cancer. Network-based characteristics and traits with the basis of epigenetics were crucial for categorization [8].

CRlncRC has remarkable dominance of prediction of sensitivity and accuracy. By using the CRlncRC method, the features used have strong correlation to oncogenic lncRNAs and their prediction can be used for further studies [8].

IV. Methodology and Implementation

Our Study integrates Machine learning methods to enhance the work . As lncRNA expression profiles are too complex in terms of dimensionality, we proposed a systematic computational framework involving ML as a tool. The proposed framework is described in Fig 2.

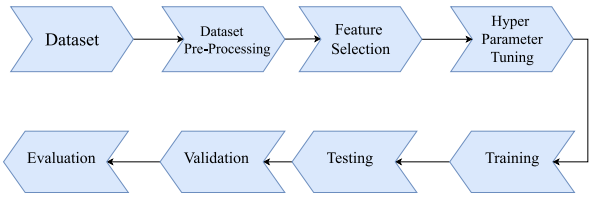


Fig. 2: Machine Learning Framework

A. Collection and Preprocessing of data

Using the UCSC Xena database, a hub for large-scale datasets related to cancers, we have downloaded the expression data of lncRNAs for five different cancers specifically, Liver, Head and Neck, Cervical, Colon and, Bladder Cancer. It consists of expression data about lncRNAs, miRNAs, and mRNAs together.The whole database contains data of 60,483 RNAs related to 8 different cancers comprising 3656 patients.

t

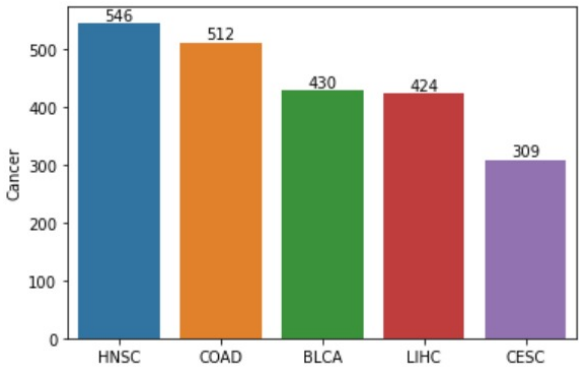


Fig. 3: Samples for each cancer

To get data related to lncRNAs from UCSC Xena we used the TANRIC database which contains only the lncRNA expressions. There are a total of 2221 patients in the filtered dataset with 12309 lncRNAs expression profile values. From the combined dataset, expression values of lncRNAs of 5 cancers are isolated using their unique identifiers, which are available in TANRIC(A web application to identify relation between lncRNA and development of cancer).

Data pre-processing is used to convert the raw data into a clean data set to build effecient Machine Learning models. Preprocessing of data is an essential step to perform on a dataset. The data is transformed and then provided as input to ML algorithms, which helps in improving overall performance. We used a mean lncRNA expression threshold of 0.3 to determine expressed lncRNAs from each cancer type.

The number of commonly expressed lncRNAs for 5 cancer types (using the union function) is 3435. Feature selection happens in data preprocessing. We used three feature selection algorithms L1 based, Univariate feature selection, and Tree based methods to identify important features that are most relevant to estimate the relationship, reduce the features that may be redundant, or not varied enough, etc.

B. Feature Selection

In our study,LncRNA expressions are used in classifying cancer types. Feature selection selects the related variables by reducing redundant and irrelevant lncRNAs. It improves the classification results, by reducing computational costs [1]. We have applied two embedded methods for feature selection to achieve this goal, one is L1 based and other is a tree based algorithm. A filter based feature selection method: Variance Threshold feature selection is used. In general, these feature selection algorithms perform very well for the task of classifying, with no limitations in terms of data type; they can deal with categorical and continuous variables, are tolerant of noise in data, and can handle missing values.

TABLE I: Summary of Expressed lncRNA's

Cancer Types	Expressed lncRNA's
BLCA	2502
COAD	2180
CESC	2328
LIHC	1773
HNSC	1831
Total(Unique)	3435

1) **L1 based feature selection:** This technique works by fitting a model to data along with regularization (L1 norm). Model used in applying this algorithm is SVM with a linear kernel. L1 norm penalizes the parameters which reduces the irrelevant features.It also handles large numbers of samples better and since it is based on the linear kernel technique.The features selected by applying this method are the ones whose coefficients are not zero, which produced a total of 1237 features.
2) **Variance Threshold:** It is an elementary method for selection of features. Features below a certain threshold are removed including those with 0 variance. Here, we presumptively consider features with substantial variance to be significant. This type of feature selection method only considers features and not target variables. The features produced by variance threshold feature selection are 1002.
3) **Tree Based Feature Selection:** This algorithm is used when there exists a non-linear relationship between the feature variables and target variables. A tree-based technique can be used to determine the importance of impurity-based features before removing any that are redundant. There are 562 features acquired through tree-based feature extraction.

For a given number of features, different algorithms for feature extraction such as L1, variance threshold, and Tree algorithm have resulted an optimal set of features. Common features produced by implementing these algorithms are a total of 234 from 2801 features.

C. Classification

Our study to identify cancer related lncRNA is modulated into a classification problem which helps in revealing key lncRNAs responsible for cancer development out of a vast number. A MultiClass Cancer Classification is carried out by training models on 3 different curated feature sets described in Table II. considering 5 cancer types(BLCA, COAD, CESC, LIHC, HNSC). The comparative analysis of the classification on trained models is made to know which features have contributed more towards classification.All our machine learning models are implemented with the help of scikit-learn library. We have trained and tested our models using a 70:30 split ratio.

Hyper-parameter tuning is performed on the models using grid search. After performing the tuning, Neighbour value "k"

TABLE II: Description of feature sets

Number of Features	Type of Feature Set
Block-A-12309	Initial feature set- Initial gathered lncRNAs from UCSC-Xena using TANRIC Database
Block-B-3435	Thresholded feature set-Expressed lncRNAs by using mean expression threshold 0.3
Block-C-234	Processed feature set-Features obtained from performing feature selection

in the K-nearest neighbors model is set to 7. 10 number of estimators are used in the random forest ,and the kernel used in Support Vector Machine is “linear”. A specific type of Naive Bayes Algorithm is used which is known as Gaussian Naive Bayes.

All machine learning models are executed on a system that incorporates 8GB RAM and a hard disk of 1 TB with an Intel Core i5 processor of the 8th generation. While performing classification of data samples are randomly selected by setting random seed to 0.

V. Performance Evaluation

A. Results Comparision

A specific set of metrics are used to assess and compare the models’ performance.The selection of metrics has an crucial role in identifying the performance of the model. We used F1score , accuracy , precision and recall as evaluation metrics for evaluating the classification models (SVM,KNN,RF,NB) computed using initial feature set ,thresholded feature set and processed feature set respectively. This enables the comparison of performance among the trained models and interpretation of the important features.

Among the considered feature sets , the best performance results are observed in BLock C. This indicates the importance of feature selection and the role of the selected features lie in classifying the cancer type.By comparison of the performance metrics, among all the four trained models SVM resulted with highest accuracy of 98.2% .

To evaluate the generalization of models,K-fold cross validation is used.Since Data considered has a large number of features and relatively less number of samples , this method is used to overcome this issue and evaluate the performance of the four considered models. Hyper parameter k was set to

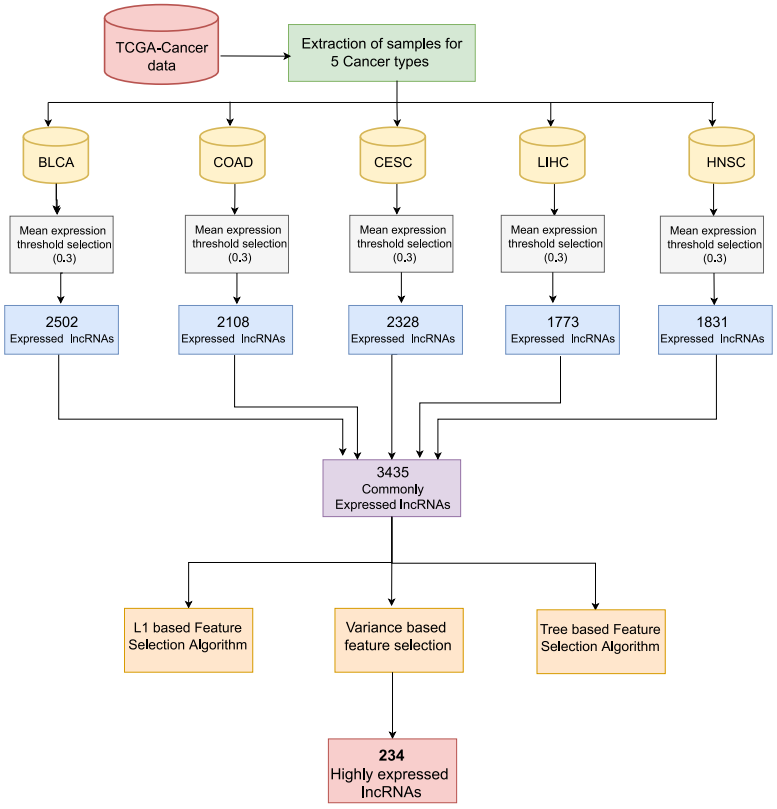


Fig. 4: Data Collection and Feature Selection

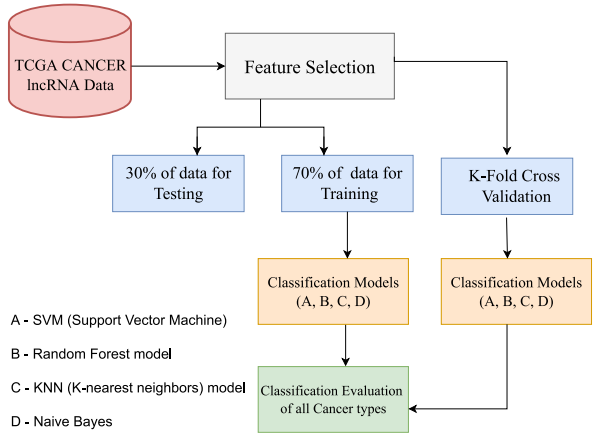


Fig. 6: Training, Cross Validation and Testing

10 and average of the all the values computed from each fold is taken as the performance metric for each model.

B. Result Analysis

By comparison of results, SVM contributed to the highest accuracy in classification of cancer types. Observation of results validates the importance of feature selection in dealing with high-dimensional gene expression data. Implementation of three feature selection methods have led to the separation of different subsets of features that served as a basis for comparative analysis in knowing the role in classifying cancer type.

The final result is the identification of important 234 lncRNAs that contributed highly to the detection of the cancer type. The expression of these lncRNAs is observed to vary when a patient has specific cancer and hence are "cancer-related". To get deep insights as to how much each identified lncRNA is

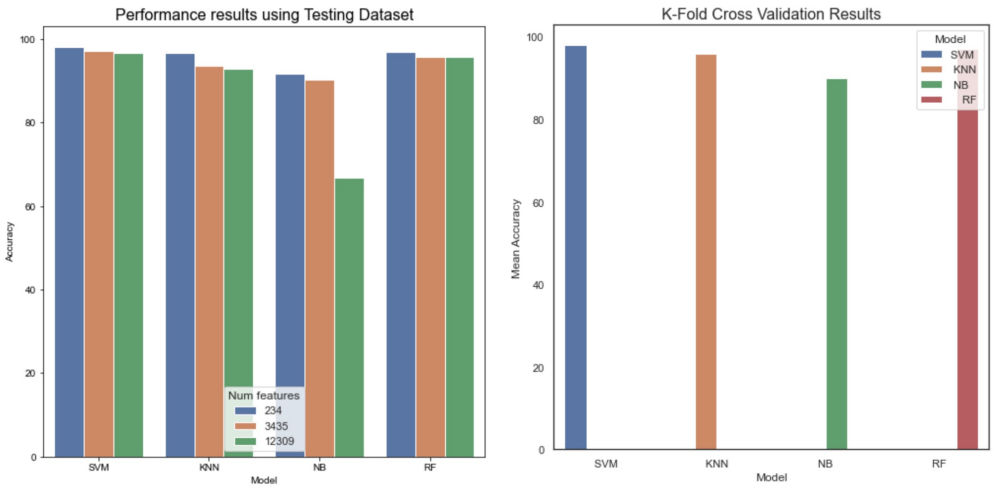
TABLE III: Result Comparison of different methods with our proposed method

Features	Model Name	Accuracy	Precision	Recall	F1-Score
Block-A	NB	66.9%	79%	62%	64%
	RF	95.68%	95%	95%	95%
	SVC	96.76%	97%	96%	97%
	KNN	92.98%	93%	92%	93%
Block-B	NB	90.25%	95%	93%	93%
	RF	95.65%	97%	97%	97%
	SVC	97.15%	97%	97%	97%
	KNN	93.70%	95%	93%	93%
Block-C	NB	91.75%	93%	90%	91%
	RF	96.85%	97%	97%	97%
	SVC	98.20%	98%	98%	98%
	KNN	96.70%	97%	96%	96%

prominent for a cancer type, correlation analysis is performed between the expressions of identified cancer-related lncRNA and each cancer type. Based on results obtained from this analysis ,we narrowed the most prominent lncRNAs for each cancer. Top 10 related lncRNAs for each cancer are pointed out using correlation matrix . The corresponding gene names for these lncRNAs are gathered and are mentioned in Table IV.

VI. Conclusion

In recent studies, several lncRNAs have been discovered. Nonetheless, the majority of the lncRNAs lack functional



(a) Accuracy of Trained Models

(b) Mean Accuracy of 10-fold Cross Validation

TABLE IV: Gene Names of identified lncRNAs

Cancer Type	Gene names for corresponding "cancer-relaed" lncRNAs
BLCA	GATA3-AS1, lnc-ACER2-1:1, RP11-379F12.4, GATA3-AS1:5, LINC02888, TBX2-AS1, lnc-GRM7-7:1, KRT7-AS:3, LINC01410, lnc-MARCH8
COAD	GAU1, MIR194-2HG, lnc-BRF2-7, LINC02086, PKD1-AS1, lnc-NTHL1-1:1, SMIM31, lnc-TRMT112-1:1, HOXB-AS4, TMEM238L
CESC	MIR9-3HG, RP11-323C15.2, FOXD3-AS1, AC018470.4, AC005082.12, LOC100996842, EPIST , MELTF-AS1, CRNDE, Lnc-SLTM-1
LIHC	SLC38A3, Lnc-APOB-1, ADORA2A-AS1, LINC02532, HULC, lnc-C2orf72-4, lnc-ONECUT1-1:6, MTUS1-DT , LINC01767
HNSC	Lnc-LRRC38-1, CT69, LINC00958, C5orf34-AS1, lnc-NDUFS6-5:6, Lnc-KHNYN-2, LINC00519, PTCSC2:1, MYOSLID, LOC101927189

characterization. In this study, we have developed an effective method to detect crucial lncRNAs for multi-cancer treatment by implementing the proposed computational framework with underlying ML algorithms fine tuned to multiclass cancer classification by using only the lncRNA expression values of cancer patients. This study led to the identification of 234 vital lncRNAs that can differentiate 5 cancer types with resultant accuracy ranging from —90.0% to 98.2%. This study also reduced the computational cost of dealing the lncRNA expression profiles by identifying the crucial lncRNAs.The t-SNE projection of these lncRNAs endorses that the identified are proficient in differentiating multiple cancers. Hence, the five cancers (Bladder, Colon, Cervical, Liver, Head and Neck) that were part of this research can both be diagnosed and prognosed using the newly found "cancer-related" lncRNAs that are responsible for cancer development.

REFERENCES

[1] Armstrong, Gregory & Kawashima, Toana & Leisenring, Wendy & Stratton, Kayla & Stovall, Marilyn & Hudson, Melissa & Sklar, Charles & Robison, Leslie & Oeffinger, Kevin. (2014). Aging and Risk of Severe, Disabling, Life-Threatening, and Fatal Events in the Childhood Cancer Survivor Study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 32. 10.1200/JCO.2013.51.1055.

[2] Seyfried, T. N., & Shelton, L. M. (2010). Cancer as a metabolic disease. Nutrition & metabolism, 7(1), 1-22.

[3] Chekuri, S., Panjala, S., & Anupalli, R. R. (2017). Cytotoxic activity of Acalypha indica L. hexane extract on breast cancer cell lines (MCF-7). The Journal of Phytopharmacology, 6(5), 264-268.

[4] Statello, L., Guo, C. J., Chen, L. L., & Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nature Reviews Molecular Cell Biology, 22(2), 96-118.

[5] Gutschner, T., & Diederichs, S. (2012). The hallmarks of cancer: a long non-coding RNA point of view. RNA biology, 9(6), 703-719.

[6] Jiang, M. C., Ni, J., Cui, W. Y., Wang, B. Y., & Zhuo, W. (2019). Emerging roles of lncRNA in cancer and therapeutic opportunities. American journal of cancer research, 9(7), 1354.

[7] Groskopf, J., Aubin, S. M., Deras, I. L., Blase, A., Bodrug, S., Clark, C.,& Rittenhouse, H. (2006). APTIMA PCA3 molecular urine test: development of a method to aid in the diagnosis of prostate cancer. Clinical chemistry, 52(6), 1089-1095.

[8] Zhang, X., Wang, J., Li, J., Chen, W., & Liu, C. (2018). CRlncRC: a machine learning-based method for cancer-related long noncoding RNA identification using integrated features. BMC medical genomics, 11(6), 99-112.

[9] Lorenzi L, Avila Cobos F, Decock A, Everaert C, Helsmoortel H, Lefever S, Verboom K, Volders PJ, Speleman F, Vandesompele J, Mestdagh P. Long noncoding RNA expression profiling in cancer: Challenges and opportunities. Genes Chromosomes Cancer. 2019 Apr;58(4):191-199. doi: 10.1002/gcc.22709. Epub 2019 Jan 20. PMID: 30461116.

[10] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016, January 26). A survey of best practices for RNA-seq data analysis - Genome Biology.

[11] He, X., Ou, C., Xiao, Y., Han, Q., Li, H., & Zhou, S. (2017). LncRNAs: key players and novel insights into diabetes mellitus. Oncotarget, 8(41), 71325.

[12] Pandey, R. R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., & Kanduri, C. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Molecular cell, 32(2), 232-246.

[13] Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. Cold Spring Harbor Protocols, 2015(11), pdb-top084970.

[14] Han, L., Yuan, Y., Zheng, S., Yang, Y., Li, J., Edgerton, M. E., & Liang, H. (2014). The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. Nature communications, 5(1), 1-9.

[15] Flippot, R., Mouawad, R., Spano, JP. et al. Expression of long non-coding RNA MFI2-AS1 is a strong predictor of recurrence in sporadic localized clear-cell renal cell carcinoma. Sci Rep 7, 8540 (2017).

[16] Bolha, L., Ravnik-Glavac, M. and Glavac, D. (2017) 'Long non-coding RNAs as biomarkers in cancer', Disease Markers, p.14, doi: 10.1155/2017/7243968.

[17] A. Al Mamun and A. M. Mondal, "Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers," 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 2825-2831, doi: 10.1109/BIBM47256.2019.8983413.

[18] Kumar, P. S., Manju, M., & Gopakumar, G. (2018). Function prediction of cancer-related LncRNAs using heterogeneous information network model. International Journal of Data Mining and Bioinformatics, 21(4), 315-338.

[19] Yang, Y., Wen, L. and Zhu, H. (2015) 'Unveiling the hidden function of long non-coding RNA by identifying its major partner-protein', Cell Bioscience, Vol. 5, No. 1, pp.1–10.

[20] Fan Zhang, Howard L Kaufman, Youping Deng, and Renee Drabier. Recursive svm biomarker selection for early detection of breast cancer in peripheral blood. BMC medical genomics, 6(1):S4, 2013.

[21] Ying Zhang, Qingchun Deng, Wenbin Liang, and Xianchun Zou. An efficient feature selection strategy based on multiple support vector machine technology with gene expression data. BioMed research international, 2018, 2018.

[22] Qian Y, Shi L, Luo Z. Long Non-coding RNAs in Cancer: Implications for Diagnosis, Prognosis, and Therapy. Front Med (Lausanne). 2020 Nov 30;7:612393. doi: 10.3389/fmed.2020.612393. PMID: 33330574; PMCID: PMC7734181.

[23] Chen, J., Lu, C. E., Wang, X., Wang, L., Chen, J., & Ji, F. (2022). LncRNA NONRATT009773. 2 promotes bone cancer pain progression through the miR-708-5p/CXCL13 axis. European Journal of Neuroscience, 55(3), 661-674.

[24] Zhang, Lin & Zhang, Dan & Qin, Zhen-Ying & Li, Jing & Shen, Zi-Yang. (2020). The role and possible mechanism of long noncoding RNA PVT1 in modulating 3T3-L1 preadipocyte proliferation and differentiation. IUBMB Life. 72. 10.1002/iub.2269.

[25] Cheetham SW, Gruhl F, Mattick JS, Dinger ME. Long noncoding RNAs and the genetics of cancer. British Journal of Cancer. 2013 Jun;108(12):2419-2425. DOI: 10.1038/bjc.2013.233. PMID: 23660942; PMCID: PMC3694235.