International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

System Biology and Machine Learning Framework for Prostate Cancer Survival Prediction

Utpala Nanda Chowdhury[a], A. F. M. Mahbubur Rahman[b], Md. Omar Faruqe[c]*, M. Babul Islam[d], Shamim Ahmad[e]

[a,b,c,e]Department of Computer Science and Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh
[d]Department of Electrical and Electronic Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh
[a]Email: unchowdhury@ru.ac.bd
[b]Email: mmr@ru.ac.bd
[c]Email: faruqe@ru.ac.bd
[d]Email: babul.apee@ru.ac.bd
[e]Email: shamim_cst@ru.ac.bd

Abstract

Prostate cancer (PC) is the most commonly diagnosed and the second most lethal malignancy in men. Proper understanding about the factors influencing the disease mechanism, response to the treatment and long term survival could facilitate effective disease management, treatment planning and decision making. Previous research initiatives reported a number of genes having impact on PC development but their genetic influence on the overall survival of the patients is still obscure. In this study, we fist identified PC related signature genes by analysing the RNA-seq transcriptomic data. Then we investigated the influence of those genes on the survival of PC patients using the clinical and transcriptomic data from the Cancer Genome Atlas (TCGA). Considering the univariate and multivariate analysis using the Cox proportional-hazards (CoxPH) model, we evidenced notable variation in the survival period between the altered and normal groups for two genes (APLN, and DUOXA1). We also identified ten hub genes such as CAV1, RHOU, TUBB4A, RRAS, EFNB1, ZWINT, MYL9, PPP3CA, FGFR2 and GATA3 in protein-protein interaction analysis that could be the source of potential therapeutic intervention. Moreover, several significant molecular pathways through functional enrichment analysis was obtained. After verification through functional studies, the identified genetic determinants could serve as therapeutic target for prolonged PC survival.

Keywords: Prostate Cancer; Gene expression; RNA-Seq; Survival analysis; Biomarker.

----

* Corresponding author.

1. Introduction

Prostate cancer (PC) is the malignancy in the prostate, the reproductive system gland of men. In 2018, it caused the second highest number of cancer incidences in men among all age groups and covered 13.5% of all cancer cases (over 1.3 million) all over the world [1]. It was the most frequent diagnosed cancer in men in Caribbean islands, Australia/New Zealand, Northern and Western Europe, and Northern America [2]. Nearly 99% of the PC occurred men are above 50 years of age [2]. The incidence rate is getting higher and higher in developed countries [3]. PC was the fifth leading death causing malignancy accounting 358,989 cancer deaths in men globally in 2018 [1]. The mortality rate notably fluctuate among different regions of the world. In 2018, the highest rates were reported in Central America, Australia, New Zealand and Western Europe which was around one tenth [1]. This rate reduced to half in Asia and North Africa [1]. Overall, PC causes the second highest death causing cancer among males in the USA [4]. However, age is considered to be the most influential factor in PC mortality where one among every two PC patients die if the patient is over 65 years old [1]. Other than age, genetic factors, ethnicity and previous history of PC occurrence in family are considered to be the most influential risk factors for PC [5–7].

Now a days, availability of high-throughput sequence (RNA-seq) data has enabled us analyzing gene expression profiles to identify genes with altered expression during cancer progression [8,9]. Such genes are the primary goals for many research activities as they pose prognostic ability and could be potential drug targets. So, we can discover putative prognostic biomarkers based on these gene expression and clinical data [10]. Several studies have suggested that integrated bioinformatics methodologies can facilitate the identification of diagnostic and prognostic biomarkers for PC [11,12]. In this study, we first identified the genes associated with PC progression by finding the overlapping differentially expressed genes (DEGs) between three RNA-seq data from the Gene Expression Omnibus (GEO) database of the National center for Biotechnology Information (NCBI). Then, we analyzed these DEGs with large clinical data obtained from the Cancer Genome Atlas (TCGA) using the Cox Proportional Hazards (CoxPH) regression model to identify the genes associated with PC survival. For this, we modelled the survival function of each DEG individually through univariate analysis and simultaneously through multivariate analysis to filter out the genes having notable difference in expression levels between altered and normal groups. The functional enrichment of the identified biomarker genes was determined by gene ontology (GO) and signaling pathway. Protein-protein interactions (PPI) were also mapped out in order to facilitate hub node identification. The methodology incorporated in this study is shown in Figure 1.

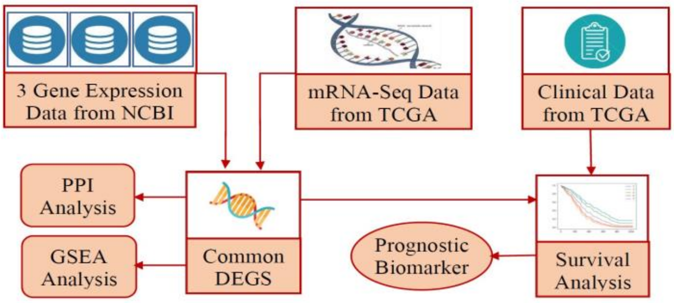

Figure 1: Block diagram for the multi-stage methodology of the study

2. Materials and Method

2.1. Dataset

In this study, we retrieved three independent RNA-seq gene expression data having accesion number GSE29155, GSE104131 and GSE75035 from the NCBI's GEO repository. GSE29155 is prepared through next generation sequencing of gene expression using the RNA-Seq technology from prostate cancer cell line and normal cell line [13]. GSE104131 is obtained by transcriptomic comparison between prostate tumor and adjacent normal cell from 16 PC patients (8 African American men and 8 European American men) using Illumina HiSeq 2500 [14]. GSE75035 is generated through Expression profiling by high throughput sequencing of prostate cancer cell line (LNCaP) and cell line representing normal prostate epithelium using Illumina HiSeq 2000 [15]. For survival analysis of PC patients integrating clinical and genetic factors, we collected the RNA-seq datasets with clinical information for PC (Prostate Adenocarcinoma TCGA, PanCancer Atlas 2018) from cBioPortal [16]. The clinical data includes 38 features for 494 patients and the RNA-seq gene expression data contains 493 cases with 12140 genes. Among the clinical features we considered the censor status indicating whether the patient survived during the observation period or not. Average age of the patients at the time of their diagnoses was 61.02 years, with a range between 41 and 78 years old. Age distribution of patients is shown in Figure 2.

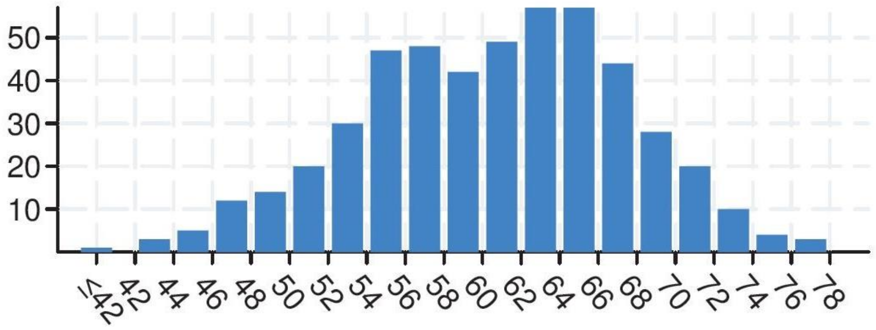

Figure 2: Age distribution at which disease was first diagnosed

2.2. Functional Enrichment Analysis

We performed the functional enrichment analysis using GO and molecular pathway analysis for the common DEGs using the web-based tool EnrichR [19]. For this, we considered Biological process (BP), Cellular component (CC) and Molecular function (MF) for GO analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) database for pathway analyses. For statistical significance of the enriched GO and KEGG pathways, manual curation was performed using the threshold of adjusted P-value < 0.05.

2.3. Protein-Protein Interactions Network Analysis

Proteins exhibit physical contacts with each other indicating some biochemical events, typically functions as some molecular processes within a cell, and thereby forms a PPI network [20]. A PPI network for the overlapping DEGs was constructed using the STRING interactome database [21]. The confidence score cutoff was set to 900 and the minimum degree was set to above 10 for the detection of highly interacted proteins (i.e., hub proteins) using topological analysis. We constructed and visualized the PPI network using the web-based visual analytical platform Network Analyst [22].

2.4. Cox proportional hazards model construction

Survival analysis estimates the expected time duration for a event, such as a death in cancer, to happen through some statistical measurements. For survival analysis of PC patients considering the overlapping DEGs, we defined the survival function using the product limit (PL) estimator. We then examined whether there is statistically considerable variation in the survival function of patients with altered gene expression and patients with normal gene expression. Then the CoxPH regression model was built determine the significant genes using the survival package in R [23]. Finally, we performed functional analyses for the obtained genes. We labeled the gene expression z-score value of each gene as altered or normal by comparing with the threshold value 2 (i.e., altered for z < 2 and normal otherwise). We performed univariate and multivariate regression which is CoxPH regression for every gene individually and simultaneously, respectively.

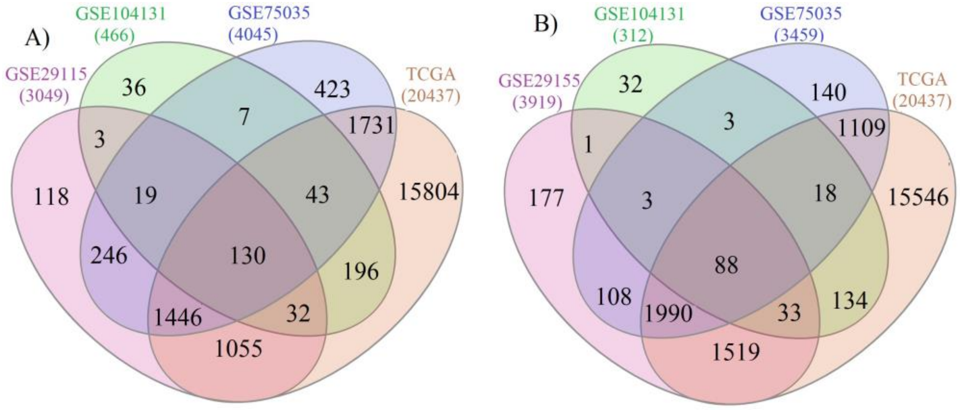

Figure 3: The Venn diagram of overlapping a) up-regulated and b) down-regulating genes among the three datasets and TCGA data

3. Results

3.1. Differentially Expressed Genes (DEGs) Identification

Comparing the tumor tissue with normal tissue we identified 149 DEGs to be commonly over-expressed and 91 DEGs to be commonly under-expressed among the three RNA-seq gene expression datasets. Among them respectively 130 and 88 (total 218) DEGs were common with the TCGA RNA-seq dataset. Figure 3 depicts the DEGs sharing among the four datasets through a venn diagram.

3.2. Functional Enrichment Analysis

Functional enrichment analyses of these overlapping DEGs identified total 1768 GO terms (1380 BP, 112 CC and 276 MF) and 198 KEGG pathways. The 5 most significant GO terms of each category and pathways according to their corresponding p-value revealing the functional mechanisms of the DEGs are summarized in

Figure 4.

3.3. Analysis of the PPI network

The PPI subnetwork considering the DEGs as proteins consists of 521 nodes and 571 interactions among the nodes (Fig. 5). Topological analysis of the network employing degree and betweenness revealed 10 hub proteins: CAV1, RHOU, TUBB4A, RRAS, EFNB1, ZWINT, MYL9, PPP3CA, FGFR2 and GATA3 (Table 1). These hub genes could be the target for therapeutic development.

3.4. Identification of survival DEGs

We applied univariate analysis on each of the common DEGs individually and multivariate analysis considering all DEGs at a time using CoxPH modelling to predict their survival function.

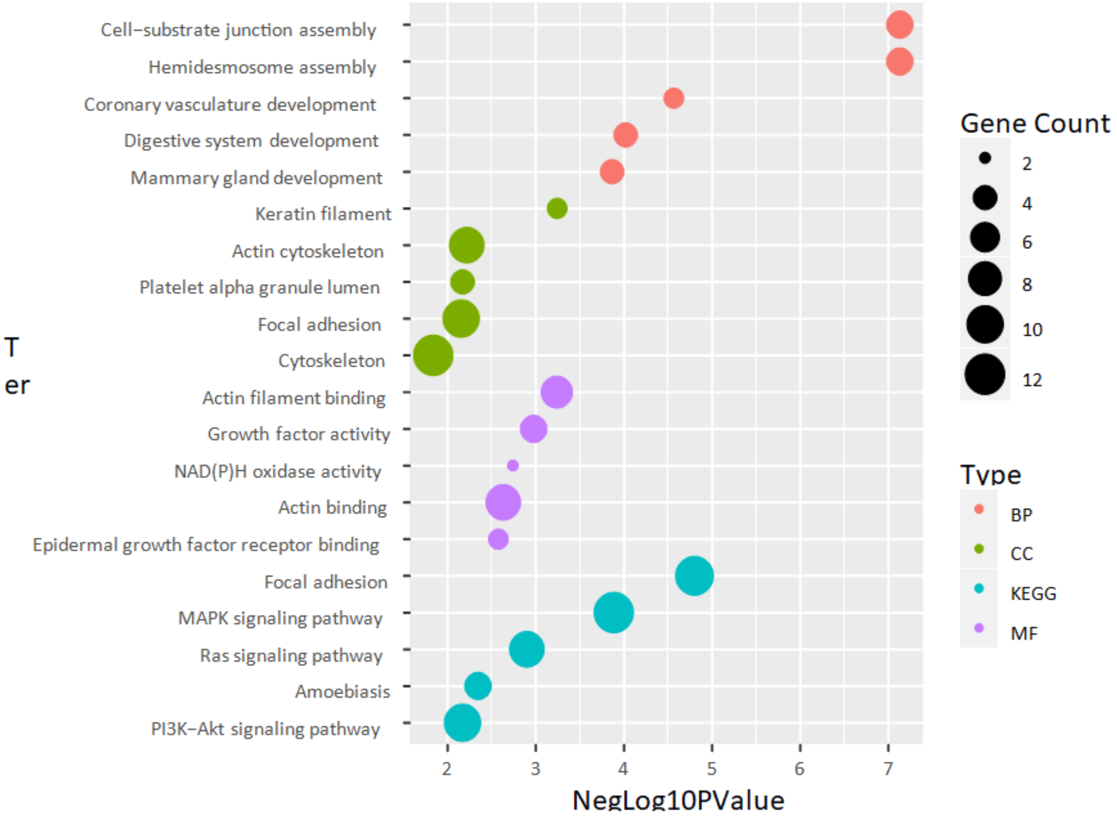

Figure 4: 5 topmost enriched GO terms of each category and KEGG pathways

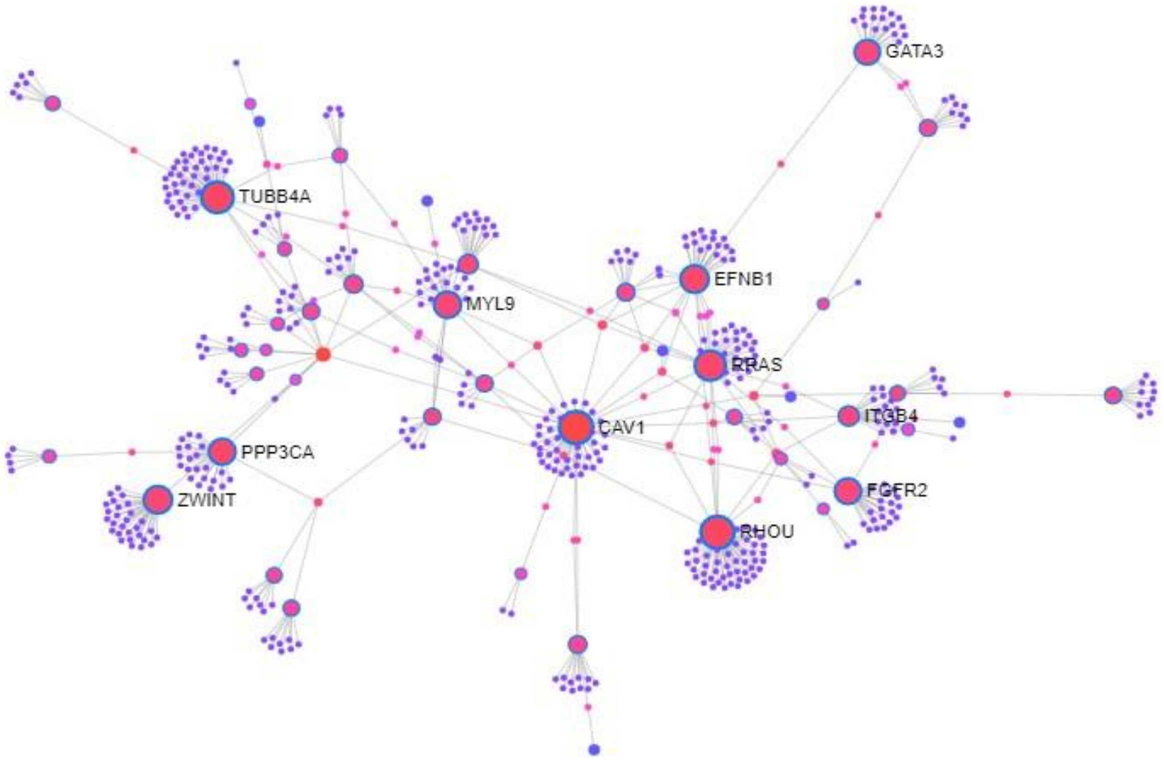

Figure 5: The PPI network of the overlapping DEGs highlighting the hub genes

In this process, the survival function of each gene was compared in altered and normal patient group. The genes having statistically significant difference in their survival function were then identified by selecting p-value less than 0.05. Thus, we found 19 such significant genes in univariate analysis and 31 genes in multivariate analysis. Among them, 2 genes APLN and DUOXA1 resulted notable difference in the survival period in both analysis (Figure. 6A). The survival pattern of these two genes in patients with altered and normal expression level are shown in Figure. 6 (B-C). It is evident from the figure that the survival probability of patients having altered expression is much less compared to the normal group for these genes.

Table 1: Particulars for the hub genes in the PPI network

Gene Symbol	Name	Expression	Degree	Betweenness
CAV1	Caveolin 1	Down	54	72,255.68
RHOU	Ras homolog family member U	Up	53	25,905.21
TUBB4A	Tubulin beta 4A class IVa	Up	45	24,823.67
RRAS	RAS related	Down	39	18,542.99
EFNB1	Ephrin B1	Down	29	20,814.17
ZWINT	ZW10 interacting kinetochore protein	Up	29	14,154
MYL9	Myosin light chain 9	Down	25	13,266.02
PPP3CA	Protein phosphatase 3 catalytic subunit alpha	Down	24	20,243.16
FGFR2	Fibroblast growth factor receptor 2	Down	23	10,995.37
GATA3	GATA binding protein 3	Down	20	10,376.01

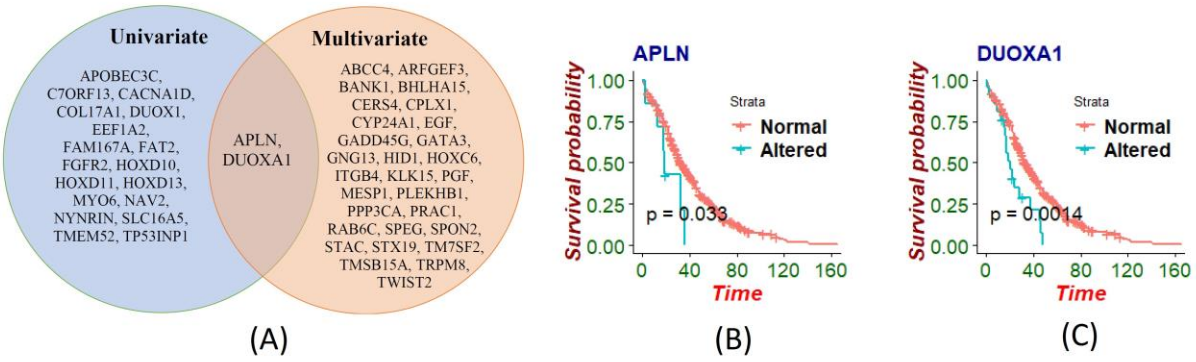

Figure 6: Prognostic biomarkers obtained in survival analysis. (A) Venn diagram shows the genes having significant influence on the survival period. (B - C) Survival Curve of genes showing significant impact of PC survival

4. Discussion

The principal goal of this study was set to identify the prognostic biomarker with an intention to mitigate the information gap regarding the progression and survival of PC. For this, we first determined the candidate signature genes by cross comparing the DEGs obtained through gene expression analysis for three transcriptomic datasets of PC. Thus, we found 218 DEGs being common between the GEO and TCGA datasets considering the expression pattern. These DEGs were the basis for the subsequent course of actions including functional enrichments in terms of GO and molecular pathways, protein-protein interactome and survival analysis. Functional enrichment analysis for the common DEGs derived significant GO terms and molecular pathways related to the disease under consideration. PPI analysis promisingly identified ten genes (CAV1, RHOU, TUBB4A, RRAS, EFNB1, ZWINT, MYL9, PPP3CA, FGFR2 and GATA3) exhibiting high degree of interactions. Among them, Liu, Yu and his colleagues previously reported TUBB4A having significant survival probability and elevated expression level in PC [24]. Again, CAV1 has been repeatedly presented as a potential biomarker and therapeutic target for PC [25]. Therefore, the obtained hub genes could be further investigated for their biological involvement and prospect as source of therapeutic targe in PC. We also estimated the survival of PC patients by univariate and multivariate analysis using the PL estimator of the CoxPH modeling. APLN and DUOXA1 showed significant variation between the altered and normal group in both studies. Hua, Wei and his colleagues evidenced that aberrant expression pattern of Apelin (APLN) in PC tissue was associated with the tumor formation and its progression towards malignancy [27]. Interestingly, APLN was reported as a putative prognostic biomarker in cervical cancer provided its impact on the disease progression [28]. Again, in a previous study, Dual Oxidase Maturation Factor 1 (DOUXA1) showed response to oxidative stress and was associated with cytokine-mediated signaling pathway, cuticle development and hormone biosynthetic process [29]. Therefore, these two genes could be the candidate for prospective prognostic biomarker for PC. Overall, identification of the candidate DEGs along with the prognostic genes for PC will favor future research and effective clinical perspective. However, the identified signature genes can further be assessed for their contribution in the survival of PC.

5. Conclusion

To conclude, the gene expression analysis of the TCGA data with three GEO datasets revealed 218 DEGs to be critical with PC development and progression. The survival analysis of these DEGs for 494 patients observed that two of them significantly reduced PC survival. In addition to this, PPI analysis resulted ten hub genes which could be of great therapeutic interest. However, the prospect of the gained results should be validated and verified through extended functional experiments. Altogether, this study offers useful knowledge and direction into clinical therapies and potential prognostic biomarkers of PC.

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), pp.394-424.

[2] American Cancer Society. Global Cancer Facts & Figures 4th Edition. Atlanta: American Cancer Society; 2018.

[3] Baade, P.D., Youlden, D.R. and Krnjacki, L.J., 2009. International epidemiology of prostate cancer: geographical distribution and secular trends. Molecular nutrition & food research, 53(2), pp.171-184. [4] Prostate Cancer Foundation. Prostate Cancer Survival Rates. (Accessed on 2021, June 25).

[5] Bostwick, D.G., Burke, H.B., Djakiew, D., Euling, S., Ho, S.M., Landolph, J., Morrison, H., Sonawane, B., Shifflett, T., Waters, D.J. and Timms, B., 2004. Human prostate cancer risk factors. Cancer: Interdisciplinary International Journal of the American Cancer Society, 101(S10), pp.23712490.

[6] Dagnelie, P.C., Schuurman, A.G., Goldbohm, R.A. and Van den Brandt, P.A., 2004. Diet, anthropometric measures and prostate cancer risk: a review of prospective cohort and intervention studies. BJU international, 93(8), pp.1139-1150.

[7] Pienta, K.J. and Esper, P.S., 1993. Risk factors for prostate cancer. Annals of internal medicine,118(10), pp.793-803.

[8] Hossain, M.A., Islam, S.M.S., Quinn, J.M., Huq, F. and Moni, M.A., 2019. Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. Journal of biomedical informatics, 100, p.103313.

[9] Hossain, M.J., Chowdhury, U.N., Islam, M.B., Uddin, S., Ahmed, M.B., Quinn, J.M. and Moni, M.A., 2021. Machine Learning and Network-Based Models to Identify Genetic Risk Factors to the Progression and Survival of Colorectal Cancer. Computers in Biology and Medicine, p.104539.

[10] Huang, Z., Yang, Q. and Huang, Z., 2018. Identification of critical genes and five prognostic biomarkers associated with colorectal cancer. Medical science monitor: international medical journal of experimental and clinical research, 24, p.4625.

[11] Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.A. and Chinnaiyan, A.M., 2001. Delineation of prognostic biomarkers in prostate cancer. Nature, 412(6849), pp.822-826.

[12] Sardana, G., Dowell, B. and Diamandis, E.P., 2008. Emerging biomarkers for the diagnosis and prognosis of prostate cancer. Clinical chemistry, 54(12), pp.1951-1960.

[13] Kim, J.H., Dhanasekaran, S.M., Prensner, J.R., Cao, X., Robinson, D., Kalyana-Sundaram, S., Huang, C., Shankar, S., Jing, X., Iyer, M. and Hu, M., 2011. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome research, 21(7), pp.1028-1041.

[14] Teslow, E.A., Bao, B., Dyson, G., Legendre, C., Mitrea, C., Sakr, W., Carpten, J.D., Powell, I. and Bollig-Fischer, A., 2018. Exogenous IL-6 induces mRNA splice variant MBD2 v2 to promote stemness in TP53 wild-type, African American PCa cells. Molecular oncology, 12(7), pp.1138-1152.

[15] Itkonen, H.M., Brown, M., Urbanucci, A., Tredwell, G., Lau, C.H., Barfeld, S., Hart, C., Guldvik, I.J., Takhar, M., Heemers, H.V. and Erho, N., 2017. Lipid degradation promotes prostate cancer cell survival. Oncotarget, 8(24), p.38264.

[16] Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E. and Antipin, Y., 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.

[17] Al Mahi, N., Najafabadi, M.F., Pilarczyk, M., Kouril, M. and Medvedovic, M., 2019. GREIN: An interactive web platform for re-analyzing GEO RNA-seq data. Scientific reports, 9(1), pp.1-9.

[18] Robinson, M.D., McCarthy, D.J. and Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), pp.139-140.

[19] Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R. and Ma'ayan, A., 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics, 14(1), pp.1-14.

[20] De Las Rivas, J. and Fontanillo, C., 2010. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol, 6(6), p.e1000807.

[21] Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P. and Jensen, L.J., 2016. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research, p.gkw937.

[22] Zhou, G., Soufan, O., Ewald, J., Hancock, REW, Basu, N. and Xia, J., 2019. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Research 47 (W1): W234-W241.

[23] Aalen, O.O., 1989. A linear regression model for the analysis of life times. Statistics in medicine, 8(8), pp.907-925.

[24] Liu, Y., Hu, C., Zhang, Q., Liu, W., Li, G., Tang, Q., Li, P., Lai, W., Zhou, M., Liu, Y. and Sheng, F., 2020. Identification of Key Genes and Molecular Mechanisms Associated With Docetaxel Resistance in Castration Resistant Prostate Cancer Based on Bioinformatics Analysis.

[25] Thompson, T.C., Tahir, S.A., Li, L., Watanabe, M., Naruishi, K., Yang, G., Kadmon, D., Logothetis, C.J., Troncoso, P., Ren, C. and Goltsov, A., 2010. The role of caveolin-1 in prostate cancer: clinical implications. Prostate cancer and prostatic diseases, 13(1), pp.6-11.

[26] Wang, R., Wu, Y., Yu, J., Yang, G., Yi, H. and Xu, B., 2020. Plasma messenger RNAs identified through bioinformatics analysis are novel, non-invasive prostate cancer biomarkers. OncoTargets and therapy, 13, p.541.

[27] Hua, W., Zhong, W., Jiang, M., Ming, X.I., Wan, S., Jiang, F. and Wan, Y., 2018. Expression of Apelin in prostate cancer tissue and its correlation with clinical prognosis. Chinese Journal of Primary Medicine and Pharmacy, 25(12), pp.1545-1548.

[28] Chen, Y., Lin, X., Zheng, J., Chen, J., Xue, H. and Zheng, X., 2021. APLN: A potential novel biomarker for cervical cancer. Science Progress, 104(2), p.00368504211011341.

[29] Chen, J.H., He, H.C., Jiang, F.N., Militar, J., Ran, P.Y., Qin, G.Q., Cai, C., Chen, X.B., Zhao, J., Mo, Z.Y. and Chen, Y.R., 2012. Analysis of the specific pathways and networks of prostate cancer for gene expression profiles in the Chinese population. Medical Oncology, 29(3), pp.1972-1984.