Interpretable machine learning approach reveals gene expression biomarkers predicting cancer patient outcomes at early stages

Ting Jin[1], Alisha Kamat[3], So Yeon Min[4], Flaminia Talos[2,5], Jonas Almeida[1], Daifeng Wang[1,5]

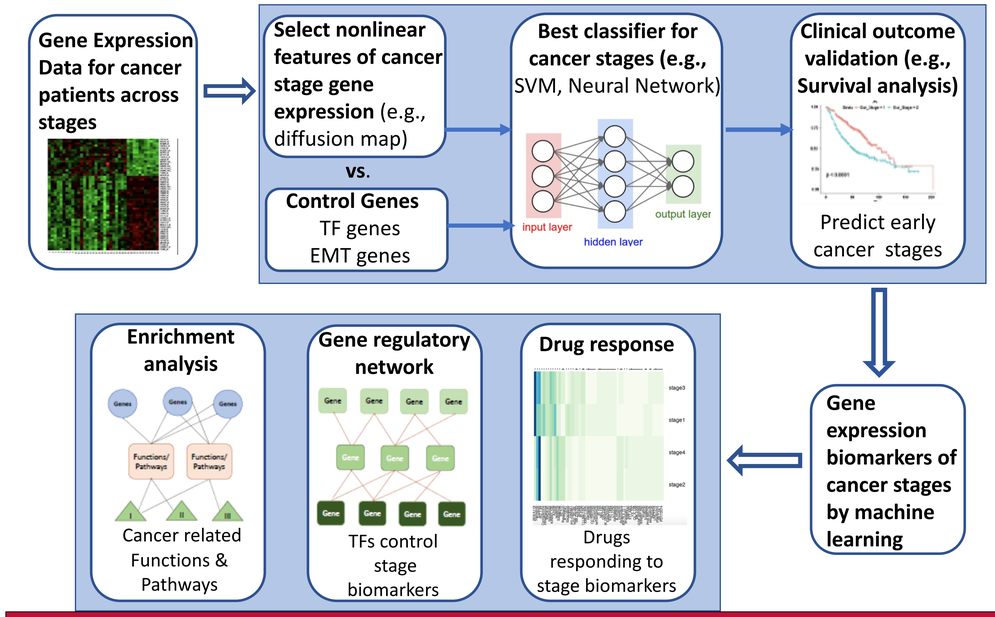
[1]Department of Biomedical Informatics; [2]Departments of Pathology and Urology; [3]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA; [4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA ;[5]Stony Brook Cancer Center, Stony Brook Medicine, Stony Brook, NY, USA

Stony Brook Medicine

Abstract

In order to understand the molecular mechanisms underlying early cancer development, we developed an interpretable, data-driven machine learning approach to identify the gene biomarkers that predict the clinical outcomes of early cancer patients. As a demonstration, we applied this approach into large-scale pan-cancer datasets including TCGA[1] to find out how effective it would be at identifying the developmental gene expression biomarkers across tumor stages for various cancer types. More relevant to the goal of machine learning interpretable classifiers, we found that early cancer patient groups clustered by the biomarkers selected have significantly more survival differences than ones by early TNM stages, suggesting that this method identified novel early cancer molecular biomarkers. Furthermore, using lung cancer as a study case, we leveraged the hierarchical architectures of neural network to identify the developmental regulatory networks controlling the expression of early cancer biomarkers, providing mechanistic insights of functional genomics driving the onset of cancer development. Finally, we reported the drugs targeting early cancer biomarkers, revealing potential genomic medicine affecting ,the early stage cancer development. The resulting computational methods are provided with open source in the public domain.

Interpretable machine learning framework



Pan-cancer

We applied our machine learning approach to multiple cancer types using TCGA datasets, and identified the cancer-type-specific gene expression biomarkers for early stage. We found that the patient groups clustered by these biomarkers (e.g., using Partitioning Around Medoids, Hierarchical Clustering) have significantly different survival rates than the ones grouped by cancer stages (log-rank test).

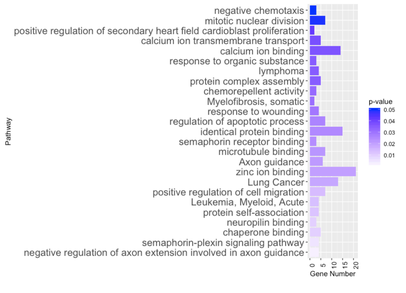
Cancer type	Log-rank test pvalue by cancer stages	Log-rank test pvalue by patients groups clustered from gene expression biomarkers	Num of Genes	Staging information for identifying biomarkers (TMN stages, patient sample size)
Breast Invasive Ductal Carcinoma (BRCA)	0.13	0.025	10	Early (I N=100) Middle (IA,IB N=88)
Kidney Renal Clear Cell Carcinoma (KIRC)	0.45	0.1	50	Early (I, N=256) Middle (II, N=56)
Head and Neck Squamous Cell Carcinoma (HNSC)	0.1	0.075	300	Early (I, N=25) Middle (II, N=75)
Thyroid Carcinoma (THCA)	0.013	0.0052	100	Early (I, N=290) Middle (II, N=51)
Bladder Urothelial Carcinoma (BLCA)	0.042	0.017	50	Early (I, N=243) Middle (II, N=162)

Lung Adenocarcinoma

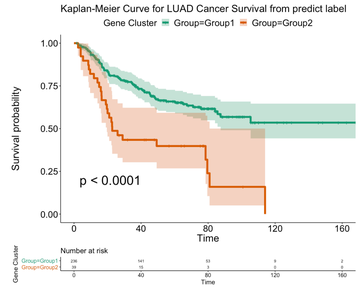
GROUP	TMN stage (number of samples) [2]	Sample size
GROUP 1	Stage I (68)+ IA (344)+IB(354)	412
GROUP 2	Stage II(58)+IIA(36)+IIB(112)+III(19)+IIIA(93)+IIIB(19)+IV(10)	691

	Classifiers(75% training and 25% testing)				
Features\Accuracy	**ANN**	**RF**	**Xgboost**	**SVM**	**LR**
Top 100 diffusion maps	0.67	0.66	0.68	0.71	0.70
Top 100 pca	0.57	0.67	0.69	0.71	0.70
Top100 Ile	0.60	0.63	0.66	0.68	0.66
Autoencoder pca (1000，100，1000)	0.64	0.67	0.70	0.80	0.63
Autoencoder diffusion map (1000，100，1000)	0.64	0.61	0.66	0.69	0.66
Autoencoder genes (11094,100,11094)	0.62	0.66	0.67	0.70	0.67
Lung mutation genes(N=52)	0.67	0.64	0.67	0.70	0.70
TF genes(N=2000)	0.52	0.66	0.69	0.55	0.51
EMT genes(N=10)	0.54	0.65	0.68	0.70	0.69

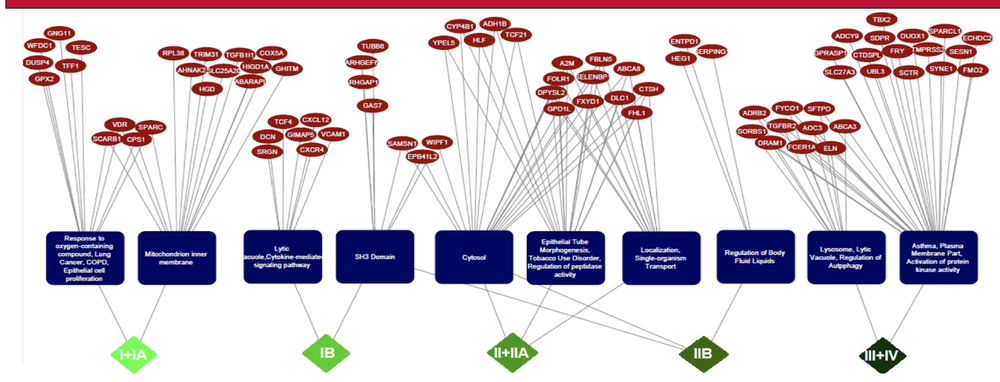
David Enriched Functions&Pathways of Top 200 genes



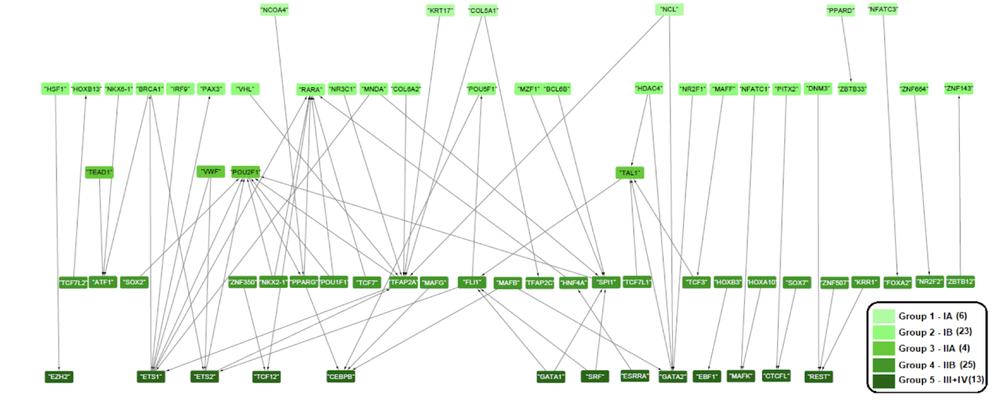
Gene expression features predicting survival rates of early caner patients



Hierarchy in artificial neural network reveals biological functions and pathways across cancer stages for lung adenocarcinoma



Gene regulatory network controlling gene expression biomarkers predicting cancer stages in lung adenocarcinoma



Select drugs for gene expression biomarkers in lung adenocarcinoma

GDSC[3]

$$Y = \beta 0 + \beta iGi + \beta tT + \beta bB$$

where Y denotes the drug sensitivity variable, Gi , T and B denote the expression of gene i, the tissue source and the experimental batch respectively, and βs are the regression coefficients.
The strength of gene-drug association is quantified by βi , above and beyond the relationship between drug sensitivity and tissue source.[4]

Drug	Biomarker Gene	Stage
Crizotinib	ALK	Stage IB-IIIA Non-Small Cell Lung Cancer that has been removed by surgery and ALK mutations.
BI.2536	FAM107A,, CYP4B1, VCAM1	non-small cell lung cancer (Stage IIIB/IV)
Navitoclax erlotinib Cyclopamine	SESN1, ECHDC2, AOC3, SLC27A3,TGFBR3	Early stage cancer stage

Conclusion

We developed an integrated and biologically interpretable machine learning approach along with a computational framework to identify gene expression biomarkers for early cancer stages. Using it, we revealed potential molecular mechanisms relating to cancer development; e.g., gene regulatory networks and functional genomics, and associated drugs. This could potentially guide future experimental validations for cancer mechanisms and treatments.

References

1. The Cancer Genome Atlas https://tcga-data.nci.nih.gov/docs/publications/tcga/?
2. https://www.cancerrxgene.org/
3. Andrew J. Gentles et al. Integrating Tumor and Stromal Gene Expression Signatures With Clinical Indices for Survival Stratification of Early-Stage Non- Small Cell Lung Cancer. Journal of the National Cancer Institute
4. https://bioconductor.org/packages/devel/bioc/vignettes/PharmacoGx/inst/doc/PharmacoGx.pdf

* Network visualization: Cytoscape (http://www.cytoscape.org)
*Enrichment analysis: DAVID (https://david.ncifcrf.gov)