A Machine Learning Approach for Biomarker Discovery using Hepatocellular Carcinoma Transcriptome Data

Michael G. Hirezi, Aditi S. Kuchi, Dr. Jiande Wu, Dr. Chindo Hicks Bioinformatics and Genetics Program, Department of Genetics, LSUHSC, New Orleans, LA

Introduction

Despite remarkable progress in patient management, liver cancer remains the fourth leading cause of cancer mortality worldwide [1]. Within the United States, an estimated 42,810 people were diagnosed with liver cancer in 2019 and an estimated 30,160 died from the disease, according to the American Cancer Society [2]. Sadly, it is one of the few types of cancer with increases in both incidence and mortality rates of about 3% per year in the United States [2]. The prevalent liver cancer is hepatocellular carcinoma (HCC) accounting for 70%–90% of all newly diagnosed liver cancers. Well supported risk factors include hepatitis B virus/hepatitis C virus (HBV/HCV) infection, nonalcoholic steatohepatitis, alcoholism, and smoking [2]. The 5-year survival rate of HCC varies widely across different populations, with an average rate of less than 32% [1]. It is a highly heterogeneous disease entity with a complex etiology, which makes prediction of disease prognosis and clinical outcomes very challenging [2]. This is further complicated by very limited effective therapeutic strategies. Thus a critical unmet medical need pivots around discovery of clinically actionable diagnostic and prognostic and targets for the development of novel therapeutics and risk prediction.

Advances in microarray and next generation sequencing have enabled molecular classification of subtypes of HCC [3], discovery of driver mutations and increased our understanding of the molecular taxonomy of the disease. Large multi-center and multinational studies such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have performed detailed analysis of the liver cancer transcriptome and genomes [3]. These primary analysis provide valuable insights about the molecular basis of liver cancer. However, despite remarkable progress afforded by these primary analyses, significant challenges remain. One of the more significant challenges is development of robust algorithms with specificity and sensitivity to accurately identify molecular diagnostic and prognostic markers and individuals at high risk of dying from the disease. This limited progress must be balanced against the recognition that liver cancer is inherently a heterogeneous disease with complex etiology. Therefore, new and more robust algorithms with sensitivity and specificity to accurately identify novel diagnostic and prognostic markers predictive of clinical outcome are needed to guide therapeutic decisions.

To address this critical unmet medical need, we propose the application of a computational framework using Machine Learning (ML) with application to gene expression and somatic mutation data for molecular classification of liver cancer and predicting clinical outcomes. Our working hypothesis is that genomic alterations in the tumor transcriptome and genome would lead to measurable changes that affect therapeutic decision making and that application of ML would enable development of more accurate algorithms to guide decision making at the point of care. To test this hypothesis, we used publicly available data gene expression and mutation data on 360 patients diagnosed with HCC and 136 control samples from the TCGA. The developed methods were validated on an independent cohort.

Methods

Liver Hepatocellular Carcinoma (LIHC) and normal tissue RNA transcript data was downloaded through the publicly available The Cancer Genome Atlas (TCGA) database and processed using Python. The data afterwards was filtered so that any genes without expression values were removed and balanced so that there were equal tumor and normal sample counts (136 each). Normalization and feature selection were performed using the EdgeR package in R to identify differentially expressed genes between the tumor and control groups. The data was re-filtered to include only the differentially expressed genes identified from the previous step and was subsequently split 75:25 into training and testing sets. The data sets were trained and tested using a machine learning classifier. This step was repeated using different subsets of genes based off increasing p-value thresholds as determined by the differential expression analysis. The classifier was designed using the Weka Python package, and the outline of the final model can be visualized below (Figure 1); different models were created, tuned, and tested against the data to determine the best model in terms of sensitivity and specificity of the classification task. A subset of significant genes were derived from the best performing model and used for pathway analysis. The final subset of genes were used to repeat the methodology on tumor data separated into two groups by outcome (alive or dead).

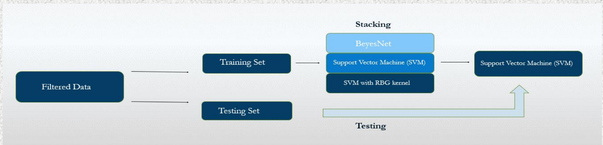
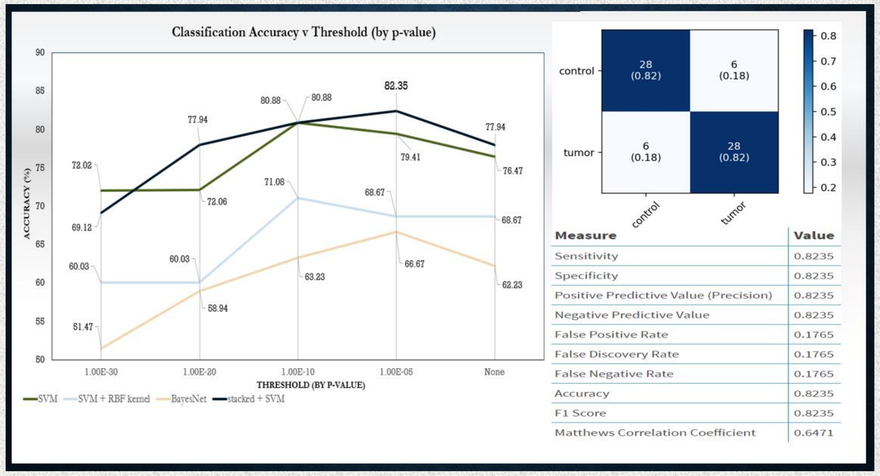


Figure 1. Schema of the machine learning classifier model. After differential expression analysis the data was filtered to include only differentially expressed genes. The data was filtered again per trial based off a p-value threshold. The data was split into training / testing sets and the training set was put into a stacked model followed by a Support Vector Machine classifier. The testing set was then put into the trained model giving prediction accuracy as an output.

Results: Comparison between tumor and control

A

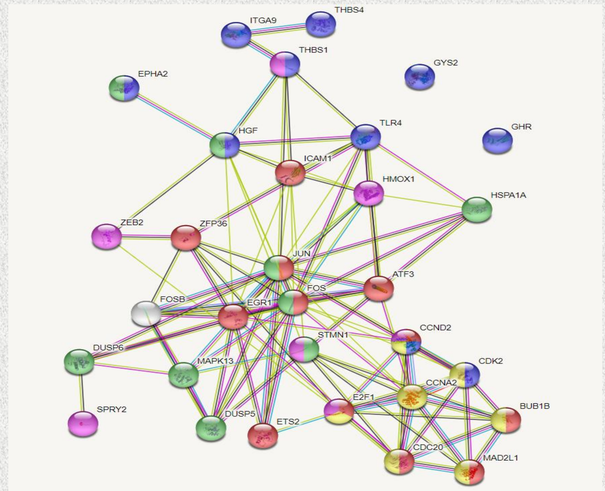


Measure	Value
Sensitivity	0.8235
Specificity	0.8235
Positive Predictive Value (Precision)	0.8235
Negative Predictive Value	0.8235
False Positive Rate	0.1765
False Discovery Rate	0.1765
False Negative Rate	0.1765
Accuracy	0.8235
F1 Score	0.8235
Matthews Correlation Coefficient	0.6471

B

Ensemble id	Gene	Protein Product
ENSG00000185664	PMEL	premelanosome protein
ENSG00000101335	MYL9	myosin light chain 9
ENSG00000214548	MEG3	maternally expressed 3
ENSG00000126838	PZP	PZP alpha-2-macroglobulin like
ENSG00000133800	LYVE1	lymphatic vessel endothelial hyaluronan receptor 1
ENSG00000142748	FCN3	ficolin 3
ENSG00000082196	C1QTNF3	C1q and TNF related 3
ENSG00000004776	HSPB6	heat shock protein family B (small) member 6
ENSG00000114270	COL7A1	collagen type VII alpha 1 chain
ENSG00000184557	SOCS3	suppressor of cytokine signaling 3
ENSG00000173918	C1QTNF1	C1q and TNF related 1
ENSG00000106366	SERPINE1	serpin family E member 1
ENSG00000138315	OIT3	oncoprotein induced transcript 3
ENSG00000126759	CFP	complement factor properdin
ENSG00000148346	LCN2	lipocalin 2
ENSG00000123454	DBH	dopamine beta-hydroxylase
ENSG00000139289	PHLDA1	pleckstrin homology like domain family A member 1
ENSG00000100292	HMOX1	heme oxygenase 1
ENSG00000087237	CETP	cholesteryl ester transfer protein
ENSG00000143387	CTSK	cathepsin K
ENSG00000160678	S100A1	S100 calcium binding protein A1
ENSG00000149716	LTO1	LTO1 maturation factor of ABCE1
ENSG00000143369	ECM1	extracellular matrix protein 1
ENSG00000113594	LIFR	LIF receptor subunit alpha
ENSG00000155659	VSIG4	V-set and immunoglobulin domain containing 4

Figure 2.a. To find the best performing model, the accuracy of 4 different classifiers were compared across 5 p-value thresholds; the p-values were determined through the differential expression analysis and serve as a cutoff for genes included in the classifier task so that all genes with a p-value less than the threshold were included during the trial. The stacked + SVM model achieved the highest accuracy at 82.35% with a p-value threshold of 1.0*10e-6 (total of 384 genes). The confusion matrix for the trial is shown to the right. **Figure 2.b** shows the top 25 differentially expressed genes between the tumor and normal tissue groups. These genes are hypothesized to hold the highest predictive value for the classification task.

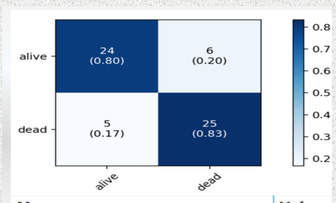


HTLV-I infection	13 of 250	0.00044
Rheumatoid arthritis	7 of 84	0.0033
Cell cycle	8 of 123	0.0033
MicroRNAs in cancer	8 of 149	0.0081
Malaria	5 of 47	0.0081
Mineral absorption	5 of 51	0.0081
Cellular senescence	8 of 156	0.0081
Osteoclast differentiation	7 of 124	0.0083
Progesterone-mediated oocyte maturation	6 of 94	0.0107
Kaposi's sarcoma-associated herpesvirus infection	8 of 183	0.0122
Hepatitis B	7 of 142	0.0122
Leishmaniasis	5 of 70	0.0122
Prolactin signaling pathway	5 of 69	0.0122
Signaling pathways regulating pluripotency of stem cells	7 of 138	0.0122
PI3K-Akt signaling pathway	11 of 348	0.0122
p53 signaling pathway	5 of 68	0.0122
MAPK signaling pathway	10 of 293	0.0122
Jak-STAT signaling pathway	7 of 160	0.0155
Influenza A	7 of 168	0.0192
TGF-beta signaling pathway	5 of 83	0.0192
Fluid shear stress and atherosclerosis	6 of 133	0.0252
IL-17 signaling pathway	5 of 92	0.0256
AGE-RAGE signaling pathway in diabetic complications	5 of 98	0.0317
Proteoglycans in cancer	7 of 195	0.0329
Epstein-Barr virus infection	7 of 194	0.0329
Chagas disease (American trypanosomiasis)	5 of 101	0.0329
Toll-like receptor signaling pathway	5 of 102	0.0329
Focal adhesion	7 of 197	0.0329
Hippo signaling pathway	6 of 152	0.0342
TNF signaling pathway	5 of 108	0.0360
Oocyte meiosis	5 of 116	0.0464
Pertussis	4 of 74	0.0477

Figure 3. The 384 genes determined by the best machine learning model were input into the STRING database for pathway analysis. The significant pathways in which these genes are involved in shown above, ordered by p-value. The KEGG pathway is named in the first column followed by the number of genes in the list per total genes recognized in that pathway. The top five pathways by gene count were visualized showing some overlap in the genes involved. These pathways include HTLV-1 infection, Cell Cycle, MicroRNAs in cancer, PI3K-Akt signaling, and MAPK signaling. The CCND2 gene (Cyclin D2) is found in 4 of 5 top pathways, and E2F1 (E2F Transcription Factor) is found in 3 of 5 top pathways.

Results: Comparison by clinical outcome

A



Measure	Value
Sensitivity	0.8276
Specificity	0.8065
Positive Predictive Value (Precision)	0.8
Negative Predictive Value	0.8333
False Positive Rate	0.1935
False Discovery Rate	0.2
False Negative Rate	0.1724
Accuracy	0.8167
F1 Score	0.8136
Matthews Correlation Coefficient	0.6337

B

Ensemble id	Gene	Protein Product
ENSG00000004776	HSPB6	heat shock protein family B (small) member 6
ENSG00000101335	MYL9	myosin light chain 9
ENSG00000185664	PMEL	premelanosome protein
ENSG00000138061	CYP1B1	cytochrome P450 family 1 subfamily B member 1
ENSG00000142748	FCN3	ficolin 3
ENSG00000149716	LTO1	LTO1 maturation factor of ABCE1
ENSG00000214548	MEG3	maternally expressed 3
ENSG00000173918	C1QTNF1	C1q and TNF related 1
ENSG00000126838	PZP	PZP alpha-2-macroglobulin like
ENSG00000178860	MSC	musculin
ENSG00000166741	NNMT	nicotinamide N-methyltransferase
ENSG00000087237	CETP	cholesteryl ester transfer protein
ENSG00000126759	CFP	complement factor properdin
ENSG00000154153	RETREG1	reticulophagy regulator 1
ENSG00000240583	AQP1	aquaporin 1 (Colton blood group)
ENSG00000100292	HMOX1	heme oxygenase 1
ENSG00000113296	THBS4	thrombospondin 4
ENSG00000082196	C1QTNF3	C1q and TNF related 3
ENSG00000109846	CRYAB	crystallin alpha B
ENSG00000125378	BMP4	bone morphogenetic protein 4
ENSG00000168077	SCARA3	scavenger receptor class A member 3
ENSG00000184557	SOCS3	suppressor of cytokine signaling 3
ENSG00000109576	AADAT	aminoadipate aminotransferase
ENSG00000163687	DNASE1L3	deoxyribonuclease 1 like 3
ENSG00000187479	C11orf96	chromosome 11 open reading frame 96

Figure 4.a. A confusion matrix of a classification task comparing clinical outcome of disease (alive vs dead). The model was able to correctly classify 81.67% of 60 samples as either alive or dead. The genes used for this task were based off the genes yielded from the best performing model from the tumor vs normal tissue comparison. A second differential expression analysis on this core list reorders the genes based off their presumptive predictive power in the new comparison by outcome. **Figure 4.b** shows the top differentially expressed genes between the alive and dead groups. The genes highlighted in yellow are those which also found in the list which demonstrates the top 25 differentially expressed genes between normal and tumor groups (see *Figure 2.b*).

Conclusions

A list of differentially expressed genes were derived from raw transcriptome data and were demonstrated to be good predictors for tumorigenicity through machine learning methods. Pathway analysis of the final list reveals genes involved in historically defined anti-apoptotic and pro-proliferative pathways, as well as those involved in cytokine signaling, cellular senescence and differentiation, hepatic disease, and viral infection. 8 genes are known miRNAs involved with cancer, and 7 are proteoglycans in cancer. The usefulness of the same method is illustrated again by comparing clinical outcome of disease within the tumor group; the list of genes produced in this step are hypothesized to explicate the genes most predictive of poor prognosis of disease. Many of the same genes are significant predictors of cancer compared to normal tissue. The overall method denotes genes which are apt candidates for biomarker analysis. It is important to reiterate that cancer is an extremely heterogenous disease, even within the same cancer type; the method doesn't currently account for subclasses of Hepatocellular Carcinoma, but HCC tumorgenicity as a whole. Future improvements of the model should aim to cluster the single tumor group into appropriate subclasses. One way of achieving this goal is to include more robust predictive data such as methylation, histone modification, mutation, and copy number variant information to increase sensitivity between subgroups.. Nevertheless, this work serves as a formidable outline for any prospective pipeline.

Citations:

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 68(6):394–424

[2] American Cancer Society. Facts & Figures 2020. American Cancer Society. Atlanta, Ga. 2020.

[3] The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.