CIRCA- CircularRNA for Cancer Active Immunotherapy: A Machine Learning Model to Predict Liver Cancer and Top Genes for Cancer Vaccine

Aditya Kiran Koushik [a*]

[a)] La Cueva High School, Albuquerque, NM, USA

ABSTRACT

Circular RNAs (circRNAs) are long non-coding RNAs with excellent prognostic and diagnostic biomarker properties for many diseases including cancer. By using liver tissues of Hepatocellular Carcinoma (HCC) patient dataset, this study designed and tested a robust machine learning pipeline to predict HCC and circRNA targeted hub gene immunogenicity for immunotherapy. First, a publicly available circRNA microarray dataset was analyzed in Python for the top twelve deregulated circRNAs in tumor tissue compared to healthy tissue. Next, classification models were trained and tested on the circRNA data. microRNA (miRNA) and gene targets (mRNA) of deregulated circRNAs were predicted and top hub genes were found from gene interaction network analysis in Cytoscape. Finally, an immunogenicity predictor in Python was built with a T-cell epitope prediction framework. This study found: 1) hsa_circ_0005284 is strongly upregulated in tumor tissue and hsa_circRNA_089372 is strongly downregulated, 2) the Logistic Regression and Naive Bayes classification models most accurately predicted tumors from circRNA data, 3) the *TMED10* and *RAB1A* hub genes were most immunogenic based on Python predictions. In conclusion, this project identifies hsa_circ_0005284 and hsa_circRNA_089372 as well as their linked immunogenic hub gene peptides as biological candidates for a liver cancer vaccine.

KEYWORDS: Liver cancer, circRNA, miRNA, T-cell epitope, Machine Learning, Artificial Intelligence, Bioinformatics, Immunogenicity, Oncogenes and Cancer Vaccine.

1 INTRODUCTION

1.1 Liver Cancer Pathophysiology

Liver cancer is a deadly disease where abnormal cell division occurs in the liver. Liver cancer remains a global health challenge as it is a leading cause of cancer-related death worldwide. Hepatocellular carcinoma (HCC) is the most common form of liver cancer, and it accounts for over 90% of cancer cases. Viral infection from hepatitis b and c are the most common causes of development, although non-alcoholic steatohepatitis associated with diabetes mellitus is also a frequent risk factor now rising in the west (Llovet et al., 2021). A few known mechanisms of HCC development/progression include genetic predisposition, viral and non-viral risk factor interaction, and cellular microenvironment alteration. Genes that play oncogenic (cancer-causing)/tumor suppressive roles can create changes in the cellular

microenvironment. With high throughput next-generation sequencing and artificial intelligence (AI) analysis, these genes can be identified and targeted (Balogh et al., 2016).

1.2 Screening, Diagnosis, and Prevention

Liver cancer can be diagnosed with a detection of a liver nodule in an abnormal ultrasonography test and with high serum α-fetoprotein levels (>20 ng/ml). Lesions less than 1 cm in diameter can be detected with ultrasonography, but for bigger lesions, a quadruple-phase CT scan or MRI is required. More and more methods to diagnose and screen HCC are being developed, which includes circular RNA (circRNA) biomarkers–which is being further explored in this study. Currently, HCC can be prevented from viral hepatitis vaccines, but there is no cure or vaccine for non-viral HCC (Llovet et al., 2021).

1.3 CircRNAs Functions and Mechanisms

CircRNAs are non-coding RNAs that form closed, continuous loops that and regulate genes in mammals. CircRNAs are generated via the back splicing of exons and introns to form exonic or intronic circRNAs (Greene et al., 2017). CircRNAs lack 5'-3' ends and poly a tails. The image in Figure 1 demonstrates the mechanisms and functions of circRNAs (Conn et al., 2015). Recent evidence implicated circRNA-mediated mechanisms in many cancers, including liver cancer (Liu, Zhang, Yan, & Li, 2020; Shen et al., 2021; Su et al., 2019; Zhang et al., 2019).

The present study is focused on using publicly available circRNAs expression data in HCC tumor tissue versus healthy tissue and identifying optimum machine learning model that correlate with tumor occurrence and likely predict tumor occurrence based on circRNAs expression levels. The study also extends to the test circRNAs -miRNA-mRNA/protein network for circRNAs target hub gene (protein) for potential cancer immunotherapy.

1.4 miRNA Sponging

Micro RNA (miRNAs) are small noncoding RNAs. miRNAs control gene expression by binding to the 3'-unstranslated region (3'-utr) in mRNAs and inhibit/suppress messenger RNA (mRNA) and translational processes. Many studies demonstrate that circRNAs contain miRNA response elements (MRES), which serve as miRNA sponges. CircRNAs can regulate gene expression via releasing miRNA to target genes (mRNAs). These mRNAs that are targeted/regulated by circRNAs via miRNA sponging will be referred to as circRNAs-targeted genes. CircRNAs that have a superior ability to bind to circRNAs are called "super sponges." Figure 2 explains the process in which circRNAs can cause cancer. Since miRNAs in turn regulate their target mRNAs, which can affect the protein expression, some of which may be oncogenes and others can act as tumor-suppressors; therefore, it is important to understand circRNAs -miRNA-mRNA-protein expression relationships for developing cancer therapies at different stages. Because proteins are the ultimate drivers of the disease process, it is also important to test if they can be directly targeted by immunotherapy via T-cell mediated blocking of such oncoproteins.

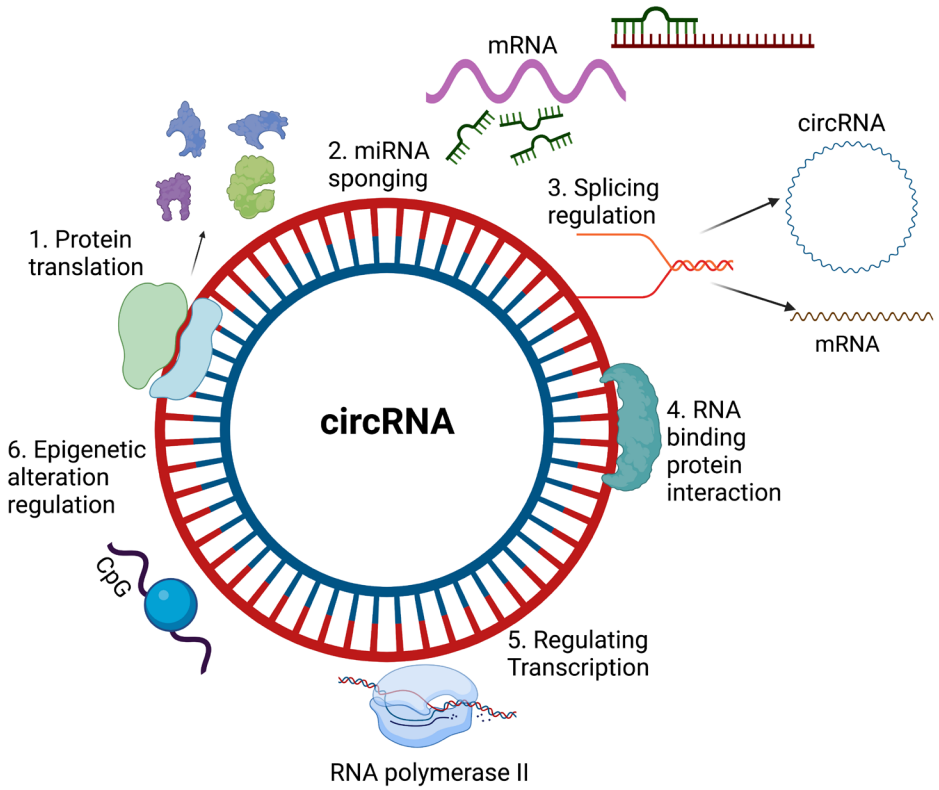

Figure 1. CircRNA interactions. 1. Protein translation–classic translational machinery can occur on circRNA producing proteins; 2. miRNA sponging–circRNAs can act to bind to miRNA and deploy when needed for anti-sense inhibition of complementary mRNAs; 3. Splicing regulation–to lead to production of circRNAs and mRNAs; 4. Interact with RNA binding proteins–circRNAs can interact with RNA binding proteins and can regulate post-transcriptional processes; 5. Regulation of transcription–circRNAs can regulate transcription of mRNAs. 6. Epigenetic alteration regulation–circRNA can also alter epigenetics. Created with BioRender.com.

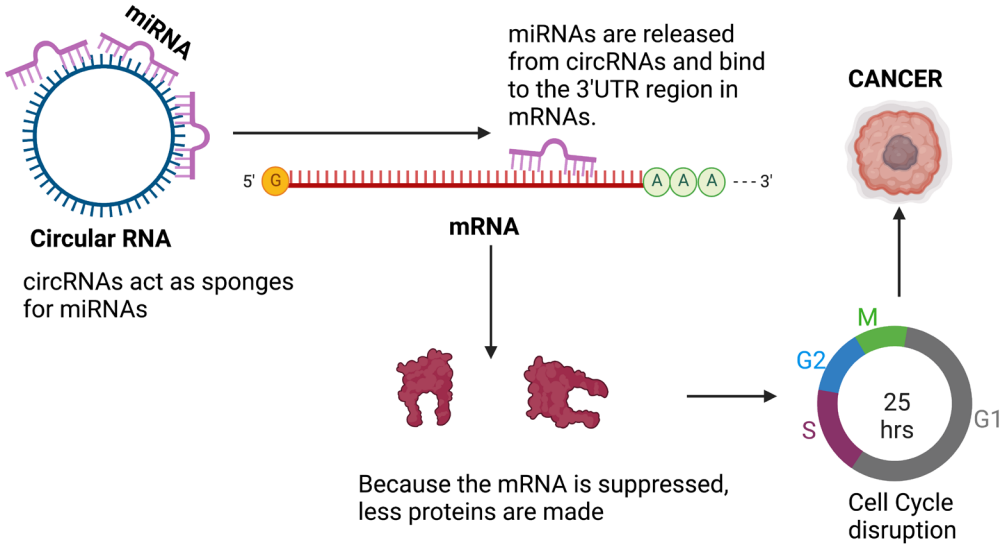

Figure 2. Schematic showing how circRNA may play a role in cancer. Created with BioRender.com.

1.5 T-cell activation and killing

T-cells are a type of leukocytic cell that play an essential role in the immune system (Henderson, 2021). T-cells originate in the bone marrow and are matured in the thymus. T-cell activation can involve antigen specific simulation. The T-cell receptor (TCR) binds to the antigen on the major histocompatibility (MHC) complex present on the surface of the antigen-presenting cell (APC) such as dendritic cells. This results in the activation of a killer T-cell (cytotoxic or CD8+ T-cell) which can kill tumor cells as shown by Figure 3 (Cavanagh, (n.d.)). After antigen simulation, CD8+ cytotoxic T lymphocytes kill tumor cells by secreting granzymes and perforins. Perforins allow granzymes to enter the cell by holes in the cell membrane, and granzymes activate the protease caspases, resulting in apoptosis. CD4+ helper T-cells perform several different functions (De candia, Prattichizzo, Garavelli, & Matarese, 2021). As an example, they produce cytokines which enhance effectiveness of cytotoxic T-cells and respond to MHC class II antigen stimulation and are required to produce antibodies from B-cells (Garnelo et al., 2017). T-cell epitopes are peptide sequences that are presented by the MHC class II receptors on APCs, which result in stimulation of CD4+ helper T-cells and cytotoxic lymphocytes, resulting in immune system activation (Figure 3).

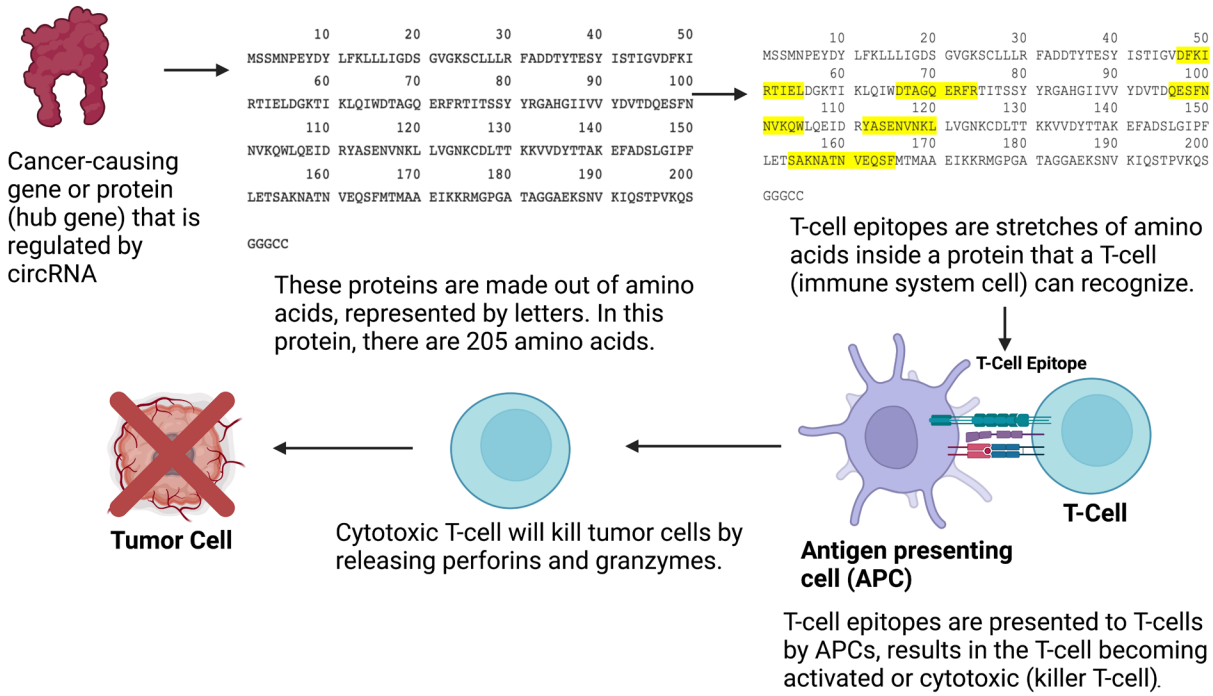

Figure 3. Potential tumor therapy strategy to identify oncogenes/oncoproteins in the hub-genes coded by mRNAs that are in turn regulated by circRNAs. Created with BioRender.com.

1.6 Immunogenicity

Immunogenicity is the ability of a protein to elicit an effective immune response. In any oncogene(s), a high T-cell epitope count can mean a higher immunogenicity and more effective immune response, and thus cytotoxic T-cell killing of tumor cells as shown by Figure 3 (De Groot et al., 2020). Peptides from proteins with high immunogenicity are great candidates

for peptide vaccines, which offer a significant alternative to whole cell antigens because of protective immune response, specificity and recognition to specific antigen, and fewer side effects (Muhammad et al., 2020).

Based on the above background, the present study used publicly available circRNAs expression data and identified optimum machine learning model that could predict, i.e. highly correlate with, the occurrence of hepatocellular carcinoma in patients and to predict the immunogenicity of deregulated circRNAs gene targets (called "hub genes") associated with tumor growth.

2 MATERIALS AND METHODS

2.1 Materials

The data for the present study was obtained from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155949 for data on ~60,000 circRNA expression levels in tumor (n=49) and adjacent healthy tissue (n=49) from n=49 liver cancer (HCC) patients (total n=98 HCC patient samples) (publicly available from GEO – gene expression omnibus). Additional materials/resources used for this study are as follows: Jupyter notebook running Python 3, Cytoscape (Shannon et al., 2003) for circRNA-miRNA-gene network, Protein-protein interaction network, and hub gene analysis, circInteractome to predict circRNA-miRNA pairs from https://circinteractome.nia.nih.gov/mirna_target_sites.html, miRWalk (Sticht, De La Torre, Parveen, & Gretz, 2018) to predict miR-gene pairs from http://mirwalk.umm.uni-heidelberg.de/, fully licensed version of BioRender to prepare model of the concept, Microsoft Excel/GraphPad Prism for data analysis, Human protein atlas for validation (https://www.proteinatlas.org/), GitHub repository for T-cell epitope data https://github.com/pirl-unc/cd8-tcell-epitope-prediction-data, Python modules used include SciPy, Pandas, NumPy, Matplotlib, Sklearn and Seaborn. Graphics included in Figure 1-4 are generated using fully licensed version of BioRender®–an online graphical tool.

2.2 Methods

First, all necessary modules were installed on Python/Anaconda using “pip install”. Next, data publicly available from GEO was downloaded. Using Pearson correlation, the top 12 circRNAs most correlated with tumor occurrence ($|r| > 0.68$ & $p < 0.01$) were identified. In this study, Pearson correlation can be used for categorical variable data. To validate the top 12 circRNAs, the mutual information (information gain) algorithm was used. Next, data on the top 12 circRNAs found previously was loaded into a new Jupyter notebook and the data was shuffled to prevent bias. Next, the data was split into training and testing with 70% – 30% split (70 for training ML models, 30 for testing). Machine learning analysis was performed for the data with seven models (K-nearest neighbor (KNN), Random Forest, Decision Tree, Support Vector Machine (SVM), Naive Bayes, Gradient Boosting, and Logistic Regression (described in methods). The predictive ability of these models was measured with classification accuracy (percent of correctly predicted tumor occurrences in patients), area under ROC curve (True positive vs False positive rate), and confusion matrix. To eliminate biases from the train-test split, Stratified K-fold Cross Validation was performed in Python, and model accuracies were compared with train-test split (normal machine learning). The confusion matrix was

performed on 12 circRNA features. Top two circRNAs were only selected for further analysis involving miRNA and mRNA expression. The confusion matrix was generated with true positive (model predicted positive and the actual label is positive), false positive (model predicted positive and the actual label was negative), true negative (model predicted negative and the actual label is negative) and false negative (model predicted negative and the actual label was positive). The second part of the project is to find circRNA- regulated hub genes (mRNAs) and their immunogenicity. To do this, another circRNA filter was created but made more statistically stringent to only get the top two circRNAs (strongest correlated with tumor occurrence) with criterion ($|r| > 0.7$ & $p < 0.01$). Using the circInteractome (Dudekula et al., 2016) and mirWalk algorithms and Microsoft Excel, datasheet was compiled of circRNA–miRNA-gene (mRNA) interactions for the top two circRNAs. Next, the interactions were graphed in Cytoscape as well as interaction strength/probability. Using the STRING function in Cytoscape, a gene interaction network (PPI) was created to model the circRNA-targeted gene interactions. Using the MCODE algorithm (finds clusters in the gene networks), the top hub genes (strongest correlation) for each circRNA were determined. Next, the Human Protein Atlas website was referenced for validation of the genes (check to see if the predicted hub genes are known oncogenes). To predict the immunogenicity of the hub genes, a T-cell epitope dataset was downloaded from GitHub. Using epitopes found in this dataset, a Python program was written to determine the number of T-cells epitopes as well as clusters of epitopes that are found in each hub gene protein sequence. More T-cell epitopes equates to higher immunogenicity since more T-cell epitopes can elicit more effective immune response by CD8+ and CD4 T-cells. To compare predicted epitopes to a negative control, a random "fake" protein sequence was generated in Python that has the same length as the hub gene, and the number of epitopes found was recorded. Finally, the number of T-cells epitopes found for each hub gene was normalized to the peptide length of each hub gene, and all immunogenicity scores and results were recorded. A brief overview of the methodology is shown in Figure 4.

Additional statistical methods, Python code and machine learning algorithms are described below.

Code snippets: All code can be found at this GitHub repository: https://github.com/adityakoushikk/ml-bio. About 1730 lines of code in total was needed.

Data Analysis and Filtering: The code (in Figure 5) filters the top 12 circRNAs using Pearson correlation. CircRNAs that have correlation coefficient $|r|>0.68$ and p-value (significance) < 0.05 are shortlisted. Pearson correlation is given by the equation (1):

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2(y-\bar{y})^2}} \quad (1)$$

Where r is the correlation coefficient, x is the values of the independent variables (each of the 12 circRNAs), $\bar{x}$ is the mean of the independent variable sample, y is the values of the target variable (tumor occurrence represented in binary 0-No tumor, 1-Tumor), $\bar{y}$ is the mean of the values in the target variable sample. Pearson correlation on categorical data works the same as point biserial correlation.

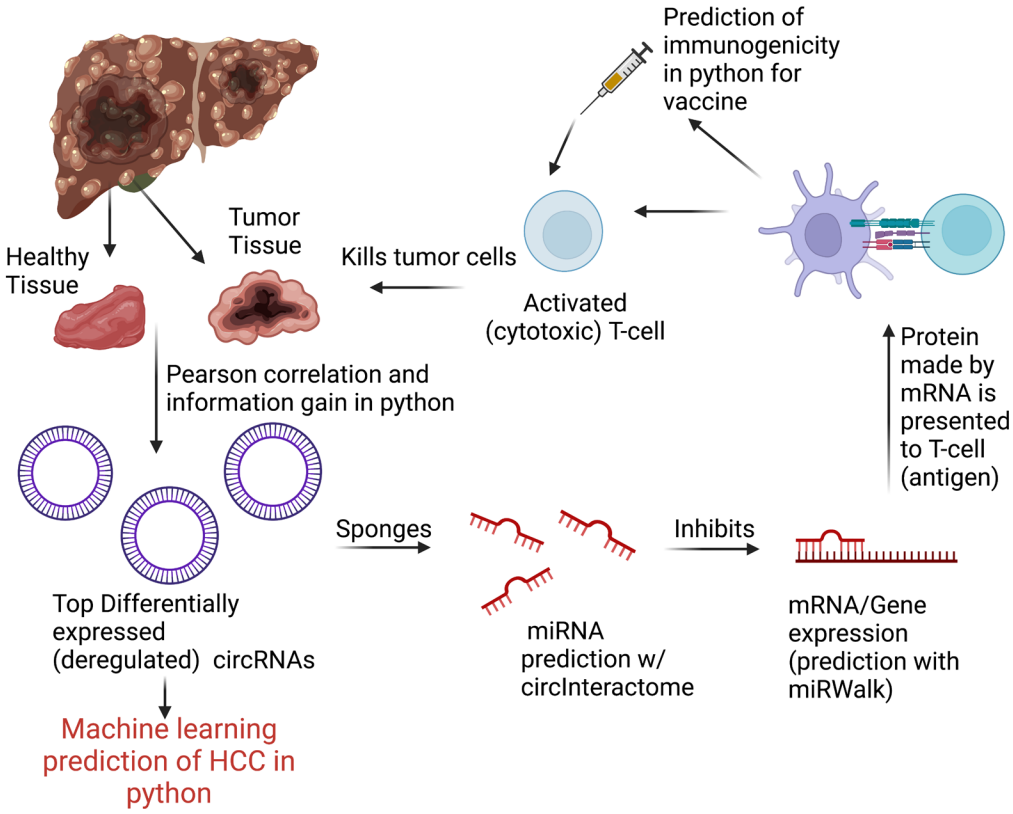

Figure 4. Experimental outline showing the analysis pipeline. Created with BioRender.com.

Machine Learning in Python & Examples: KNN

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$
$$x = y \Rightarrow D = 0$$
$$x \neq y \Rightarrow D = 1 \qquad (2)$$

An example of building a machine learning model in Python is shown in Figure 5. This classification model is K-Nearest Neighbor. First, data is split into 70% for training the model and 30% for testing using the train-test split from SKlearn. The KNN model from SKlearn is imported, and the model is fit (learns patterns in data using the KNN algorithm) for the training data. The trained model is tested on the other 30% of the data, and the prediction accuracy is displayed. A confusion matrix is displayed (explained in detail in results section). The K-nearest neighbor classifies data by looking at similarity based on Hamming distance (equation 2), or the neighboring points (if other points near the point at question fall under a certain class, then the point at question will most likely fall into that class). The Hamming Distance function is shown on the left (used for categorical prediction -https://www.saed-sayad.com/k_nearest_neighbors.htm), where K determines the number of neighbors. The code for building the other machine learning models is repeated in a similar way.

```
#create a dict to keep track of the accuracy of each model
accuracies = {}
#X will contain circRNA values, y will contain target variable - tumor
X = circ.values
y = circ['Tumor'].values
X = np.delete(X,0,axis=1)
#split the dataset into 70-30 for training and testing
x_train, x_test, y_train, y_test = train_test_split(X,y,test_size = 0.3,random_state=1)
scoreList = []
# KNN Model
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors = 2)
knn.fit(x_train, y_train)

print("{} NN Score: {:.2f}%".format(2, knn.score(x_test, y_test)*100))
knn_predictions = knn.predict(x_test)
scoreList.append(knn.score(x_test, y_test))
acc = max(scoreList)*100
accuracies['KNN'] = acc
confusion_matrix(y_test, knn_predictions)
```

```
2 NN Score: 83.33%

array([[15,  1],
       [ 4, 10]])
```

```
knn_predictions
```

```
array([1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0,
       0, 1, 0, 0, 0, 1, 1, 0])
```

Figure 5. K-Nearest Neighbor (KNN) classification model in Python.

Random Forest & Decision Tree: Decision trees learn by splitting dataset into smaller subsets to predict a target value (each condition is called a node, and possible outcomes are called "branches"), hence forming a tree. Random forest consists of many individual decision trees that operate as an ensemble (multiple learning algorithms). Decision Trees usually classify using Gini impurity, which gives a probability of misclassifying an observation by randomly picking a datapoint and randomly classifying it according to the distribution of the dataset (equation 3). The equation is shown on below, where C is the total possible classes, p(i) is the probability of picking a datapoint with class i (in this study, C is 2 for Tumor/No-tumor prediction, and p(i) would be 0.5).

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i)) \tag{3}$$

Logistic Regression: Logistic Regression solves binary classification problems although it is a regression function. The basis of logistic regression is the sigmoid function shown in equation 4:

$$y = \frac{1}{1+e^{-x}} \tag{4}$$

The function can take any real value number and map it to a value between 0 and 1 (continuous circRNA expression values can be mapped in this way). The sigmoid curve can be graphed as in Figure 6 (https://towardsdatascience.com/logistic-regression-explained-9ee73cede081). With this method, the location of the point on the curve can determine the classification of that point. In this study, the x-axis would be the circRNA expression, and the y-axis would be the probability of tumor vs no-tumor.

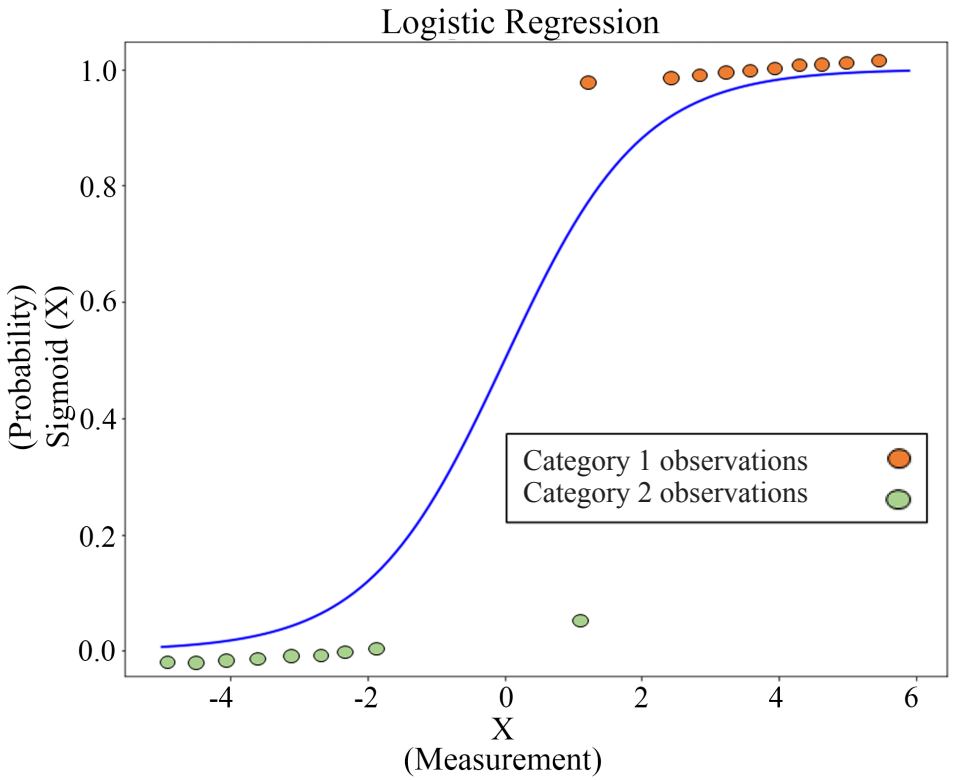

Figure 6. Sigmoid curve based on Logistic Regression showing CircRNA expression values (0 and 1; 0 being no tumor and 1 being tumor).

CircRNA-miRNA-Gene network

First, the top 2 circRNAs from Pearson correlation filtering ($|r| > 0.7$) were found (further explained below). These circRNAs were entered into the circInteractome website to predict the miRNA sponges for each of the two circRNAs. Figure 7 (based on circInteractome analyses - https://circinteractome.nia.nih.gov/mirna_target_sites.html) shows the result for one of the circRNAs (hsa_circ_0005284). Only the top 3 miRNAs were selected based on the context+ score (smallest numbers are ideal). miRNAs selected for hsa_circ_0005284 were hsa-miR-558, hsa-miR-639, hsa-miR-626. For hsa_circRNA_089372, miRNAs selected were hsa-miR-1247-5p, hsa-miR-1289, and hsa-miR-1184. These miRNAs were fed into the mirWalk algorithm to predict the gene (mRNA) targets that these miRNAs inhibit. Figure 8 shows an example for gene target search with mirWalk for hsa-miR-1184. These analyses were done based on mirWalk (http://mirwalk.umm.uni-heidelberg.de/). Genes were filtered for a score higher than 0.85, only CDS (coding sequence genes), and finally validation from miRTarBase.

miRNA Predictions

miRNA	#Sites
hsa-miR-1265	1
hsa-miR-495	1
hsa-miR-526b	1
hsa-miR-558	1
hsa-miR-582-3p	1
hsa-miR-587	1
hsa-miR-622	1
hsa-miR-626	1
hsa-miR-639	1

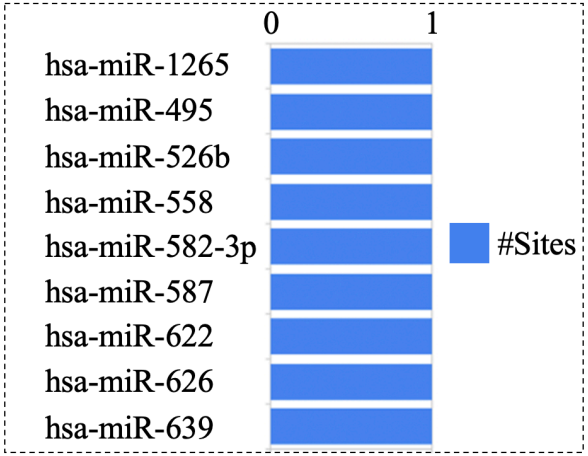

Figure 7. CircRNA-miRNA-mRNA (gene) network for top two significantly altered CircRNA based on circInteractome analyses.

TargetScan miRNA predictions

CircRNA Mirbase ID	CircRNA (Top) - miRNA (Bottom) pairing	Site Type	circRNA Start	circRNA End	3' pairing	Local AU	Position	TA	SPS	Context+ Score	Context+ Score+%
hsa_circ_0005284 (5' ... 3') hsa-miR-1265 (3' ... 5')	GGAUACAGCCUGUGCACAUCCUG UUGUUGUGAACUGGUGUAGGAC	7mer-m8	213	219	0.003	0.064	-0.061	-0.011	-0.002	-0.127	86
hsa_circ_0005284 (5' ... 3') hsa-miR-495 (3' ... 5')	UCUGAGAAGGGACCAGUUUGUUG UUCUUCACGUGGUACAAACAAA	7mer-m8	44	50	-0.007	0.053	-0.056	0.038	0.071	-0.021	93
hsa_circ_0005284 (5' ... 3') hsa-miR-526b (3' ... 5')	AUGCGACUGAGACAGCUCAAGAG UGUCUUUCACGAAGGGAGUUCUC	7mer-m8	92	98	0.003	0.047	-0.049	0.003	0.016	-0.1	86
hsa_circ_0005284 (5' ... 3') hsa-miR-558 (3' ... 5')	UACAUGCGACUGAGACAGCUCAA UAAAACCAUGUCGUCGAGU	7mer-1a	89	95	0.004	0.020	-0.040	0.010	-0.052	-0.132	80
hsa_circ_0005284 (5' ... 3') hsa-miR-582-3p (3' ... 5')	GAAAUCUGAGAAGGGACCAGUUU CCAAGUCAACAAGUUGGUCAAU	7mer-m8	40	46	0.012	0.057	-0.056	-0.009	0.004	-0.112	82
hsa_circ_0005284 (5' ... 3') hsa-miR-587 (3' ... 5')	AAUCUUAAACCAAGAAUGGAAAC CACUGAGUAGUGGAUACCUUU	7mer-1a	184	190	0.004	0.024	-0.046	0.023	0.021	-0.048	81
hsa_circ_0005284 (5' ... 3') hsa-miR-622 (3' ... 5')	NNNAAUAAUCAUGGGCCAGACUGG CGAGGUUGGAGUCGUCUGACA	7mer-m8	14	20	too_close	too_close	too_close	too_close	too_close	too_close	NA
hsa_circ_0005284 (5' ... 3') hsa-miR-626 (3' ... 5')	GAGUACAUGCGACUGAGACAGCU UUCUGUAAAAG---UCUGUCGA	7mer-m8	86	92	-0.025	0.023	-0.050	-0.005	-0.037	-0.214	91
hsa_circ_0005284 (5' ... 3') hsa-miR-639 (3' ... 5')	AACCAAGAAUGGAAACAGCGAAG UGUCGCGAGCGUUGGCGUCGCUA	7mer-1a	191	197	0.008	0.010	-0.047	-0.080	-0.048	-0.231	73

miRNAID | GeneSymbol | 0.85 | CDS | | | miRTarBase | set filter

miRNA	RefseqID	GeneSymbol	Duplex	Score	Position	Binding Site	Au	Me	N Pairings	Targetscan	Mirdb	Mirtarbase
hsa-miR-1184	NM_001256695	PRDM11	details	0.92	CDS	3292,3330	0.37	-5.18	22	—	—	MIRT790030
hsa-miR-1184	NM_001271282	DICER1	details	1.00	CDS	5623,5675	0.52	-5.495	20	—	Link	MIRT558376
hsa-miR-1184	NM_001271608	LASP1	details	0.92	CDS	596,619	0.34	-6.828	18	—	—	MIRT735893
hsa-miR-1184	NM_001282878	LAX1	details	1.00	CDS	824,846	0.48	-6.372	18	—	—	MIRT643447
hsa-miR-1184	NM_001284390	VGLL4	details	1.00	CDS	878,898	0.27	-8.492	18	—	Link	MIRT464214
hsa-miR-1184	NM_001284391	VGLL4	details	1.00	CDS	780,800	0.27	-8.492	18	—	Link	MIRT464214
hsa-miR-1184	NM_001286265	MRS2	details	0.92	CDS	978,1001	0.53	-6.352	18	—	—	MIRT640711
hsa-miR-1184	NM_001287489	OTOF	details	1.00	CDS	3440,3474	0.34	-5.269	19	—	—	MIRT718119
hsa-miR-1184	NM_001287489	OTOF	details	1.00	CDS	5364,5392	0.46	-5.763	20	—	—	MIRT718119
hsa-miR-1184	NM_001287489	OTOF	details	1.00	CDS	2559,2586	0.35	-6.484	21	—	—	MIRT718119
hsa-miR-1184	NM_001290145	POLDIP2	details	1.00	CDS	954,999	0.37	-10.845	17	—	Link	MIRT566361
hsa-miR-1184	NM_001297576	PEA15	details	0.92	CDS	512,534	0.48	-4.171	18	—	Link	MIRT735898
hsa-miR-1184	NM_031459	SESN2	details	1.00	CDS	1077,1139	0.31	-8.382	18	—	—	MIRT505717
hsa-miR-1184	NM_032206	NLRC5	details	0.92	CDS	4773,4799	0.47	-8.948	19	—	—	MIRT725310
hsa-miR-1184	NM_032206	NLRC5	details	1.00	CDS	1929,1953	0.43	-5.771	19	—	—	MIRT725310
hsa-miR-1184	NM_032313	NOA1	details	1.00	CDS	1872,1908	0.56	-6.251	21	—	—	MIRT719531
hsa-miR-1184	NM_032816	CEP89	details	0.92	CDS	950,979	0.53	-4.957	19	—	—	MIRT626489
hsa-miR-1184	NM_080550	SYNRG	details	1.00	CDS	1030,1062	0.4	-6.484	18	—	—	MIRT725121
hsa-miR-1184	NM_133635	POFUT2	details	0.92	CDS	182,205	0.43	-10.139	18	—	—	MIRT761200
hsa-miR-1184	NM_180976	PPP2R5D	details	0.92	CDS	1753,1772	0.38	-10.526	15	—	—	MIRT735899

Figure 8. mirWalk analyses showing mRNA (hub gene) targets for the top miRNA from Figure 7. Based on this, it is expected that these mRNAs are inhibited by microRNAs that are in turn sponged by CircRNA (from Figure 6 above). The identified miRNA-mRNA pairs were validated by miRTarBase.

A.

	allele	peptide	measurement_value	measurement_inequality	measurement_type	measurement_kind	measurement_source	original_allele
0	BoLA-1*21:01	AENDTLVVSV	7817.0	=	quantitative	affinity	Barlow - purified MHC/competitive/fluorescence	BoLA-1*02101
1	BoLA-1*21:01	NQFNGGCLLV	1086.0	=	quantitative	affinity	Barlow - purified MHC/direct/fluorescence	BoLA-1*02101
2	BoLA-2*08:01	AAHCIHAEW	21.0	=	quantitative	affinity	Barlow - purified MHC/direct/fluorescence	BoLA-2*00801
3	BoLA-2*08:01	AAKHMSNTY	1299.0	=	quantitative	affinity	Barlow - purified MHC/direct/fluorescence	BoLA-2*00801
4	BoLA-2*08:01	DSYAYMRNGW	2.0	=	quantitative	affinity	Barlow - purified MHC/direct/fluorescence	BoLA-2*00801

B.

```
total = 0
lists = []
cluster = []
for index, row in peps.iterrows():
    lists.append(row["peptide"])
for x in lists:
    if x in e:
        total+=1
        clus = e.find(x)
        cluster.append(clus)
print(total)
```

C.

```
         10         20         30         40         50
MNLFRFLGDL SHLLAIILLL LKIWKSRSCA GISGKSQVLF AVVFTARYLD
         60         70         80         90        100
LFTNYISLYN TCMKVVYIAC SFTTVWLIYS KFKATYDGNH DTFRVEFLVV
        110        120        130        140        150
PTAILAFLVN HDFTPLEILW TFSIYLESVA ILPQLFMVSK TGEAETITSH
        160        170        180        190        200
YLFALGVYRT LYLFNWIWRY HFEGFFDLIA IVAGLVQTVL YCDFFYLYIT
        210
KVLKGKKLSL PA
```

Figure 9. (A) Potential T-cell epitope in hub genes (mRNAs identified from Figure 6 and 7) were compared against known T-cell epitopes from global T-cell epitope dataset. The table shows some few examples of T-cell epitopes (peptides), their score and the assays used to confirm immunogenicity. (B) Python program used to find the immunogenicity of each protein sequence. (C) Epitope clustering is determined by the location of epitopes in the protein sequence. Note that the actual number (and not normalized) of T-cell epitopes in each candidates gene targets are identified.

Immunogenicity Prediction

Figure 9A shows the head of the T-cell epitope dataset, which contained about 200,000 T-cell epitopes. First, epitopes were filtered for only human alleles of MHC proteins (HLA). Next, they were filtered for the best nanomolar affinity between each epitope and peptide (most often an IC50 or inhibit concentration value). The smaller value indicates a higher MHC binding affinity and therefore a better epitope. Each IC50 value was normalized to a value between 0 and 1 using 1 – log (min(IC50, 50000))/log(50000). The epitopes that scored less than 0.35 on the normalized scale were discarded. Another dataset on eluted-MHC ligands was downloaded, which contained epitopes that are identified to bind MHCs via immuno-precipitation. Finally, an overlap between epitopes from each dataset was created, and the epitopes (Figure 9A) would be checked in the hub gene peptide sequences.

Figure 9B was the program used to find the immunogenicity of each protein sequence. "e" was the peptide sequence, and x was each T-cell epitope. The location of each epitope found was also predicted. Location is given by the position of the epitope amino acid in the full proteins sequence. For example, the full proteins sequence for KDELR1 hub gene is shown in Figure 9C. The epitope clustering in Figure 9C was determined by the location of epitopes in the protein sequence.

CircRNA correlation (Supplementary Finding): Some circRNAs are highly correlated with each other, rather than with tumor occurrence. The researcher hypothesized that circRNAs that are highly correlated with each other come from the same gene. For example, ASCRP3008985 (hsa_circ_0008539) and ASCRP3009102 (hsa_circ_0031027) have a strong correlation (r = 0.95) on the heatmap. Arraystar data on the two circs show they come from the same gene (TMCO3). Example graphs are shown below (Figure 10), and actual correlations are in Figure 11.

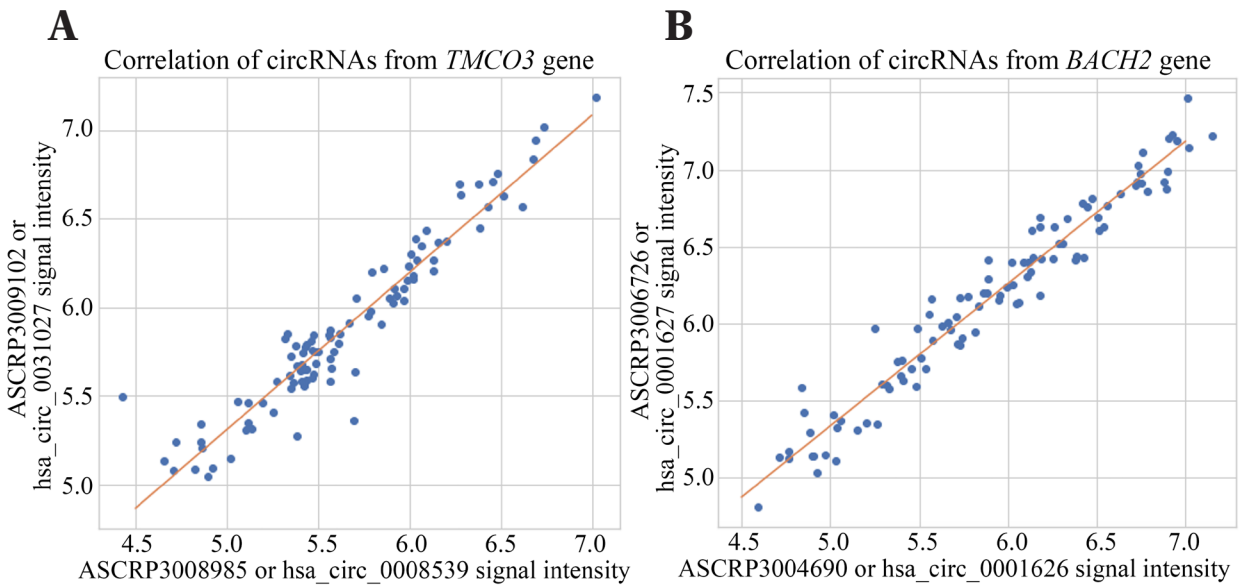

Figure 10. Pearson correlation for hsa_circ_0008539 and hsa_circ_0008539 for hub gene TMCO3 (A); and has_circ_0001627 and has_circ_0001626 for hub gene BACH2 (B). Pearson correlation coefficient for panel A – 0.95; Pearson correlation coefficient for panel B – 0.97.

3 RESULTS

The results are divided into following sections: CircRNA & Tumor Correlation, Machine Learning, Cytoscape Network Analysis, Validation & Survival Analysis, and Immunogenicity Prediction.

CircRNA & Tumor Correlation

Figure 11 shows the Pearson correlation heatmap generated in Python of top 12 deregulated circRNAs. Purple regions indicate strong negative correlation, and green areas indicate strong positive correlation. The most significant column is the first column, as it shows the correlation with tumor occurrence, which is this study's purpose. CircRNAs that are strongly negatively correlated ($r<0$) with tumor occurrence are downregulated (less expression in tumor tissue compared to healthy tissue), and circRNAs that are positively correlated ($r>0$) are upregulated (more expression in tumor tissue compared to healthy tissue). Upregulated circRNAs are associated with tumor growth, while downregulated circRNAs are not associated with tumor growth. The top 12 deregulated circRNAs can also be visualized similarly with violin plots below (Figure 12).

Violin Plots of Top 12 Deregulated CircRNAs

Figure 12 shows the violin plots for the top 12 deregulated circRNAs. The violin plot graphs the general distribution of the data as well as the median signal intensities for both Tumor and No-Tumor samples (0 indicates No-Tumor, 1 indicates Tumor). The shape of each plot follows a probability density function, where the widest point indicates a high probability that each sample will have the given circRNA expression value. A uniform shape shows a normal distribution, where data points are concentrated around the median. Dots shown in the middle of the graph represent each of the 98 tissue samples from the dataset, and more concentrated/clustered data points will widen the probability density function/distribution curve. Upregulated circRNAs show a blue violin plot higher than the red (signal intensity/expression for tumor samples is higher than for non-tumor samples), and the opposite for downregulated circRNAs. Distributions of both Tumor and Non-Tumor plots show generally normal distributions with some plots having an irregular distribution.

After restricting the Pearson correlation values to $[r] > 0.7$, the top 2 circRNAs were ASCRP3001251 (formal alias is hsa_circ_0005284) and ASCRP3001458 or hsa_circRNA_089372. The hsa_circ_0005284 is upregulated in tumor with a correlation value of 0.71 and hsa_circ_089372 is downregulated with correlation value of -0.82 (strongest out of all circRNAs). Figure 13A & 13B are graphs of a sigmoid function fit for the top 2 circRNAs.

From the graphs, high hsa_circ_0005284 expression values are mostly associated with tumor occurrence, while high hsa_circRNA_089372 expression values are indicative of no tumor. This shows that the circRNAs are opposites of each other, and the sigmoid curves are also pointing opposite ways for each circRNA. hsa_circRNA_089372 (Figure 13A) has a more uniform sigmoid curve and higher Pearson value than hsa_circ_0005284 (outliers at the top of the graph). The sigmoid curve in Figure 13A is close to 1, which shows a near 100% tumor probability for low circRNA expression. In Figure 13B, the probability curve is not as certain and even high signal intensity will only guarantee ~90% tumor probability. Overall,

from sigmoid curve analysis and Pearson correlation, hsa_circ_089372 is more correlated with lower tumor probability and is relatively a better biomarker than hsa_circ_0005284 albeit both show strong, yet opposite, correlations (hsa_circ_089372 – negative correlation and hsa_circ_0005284 – positive correlation) with tumor probability.

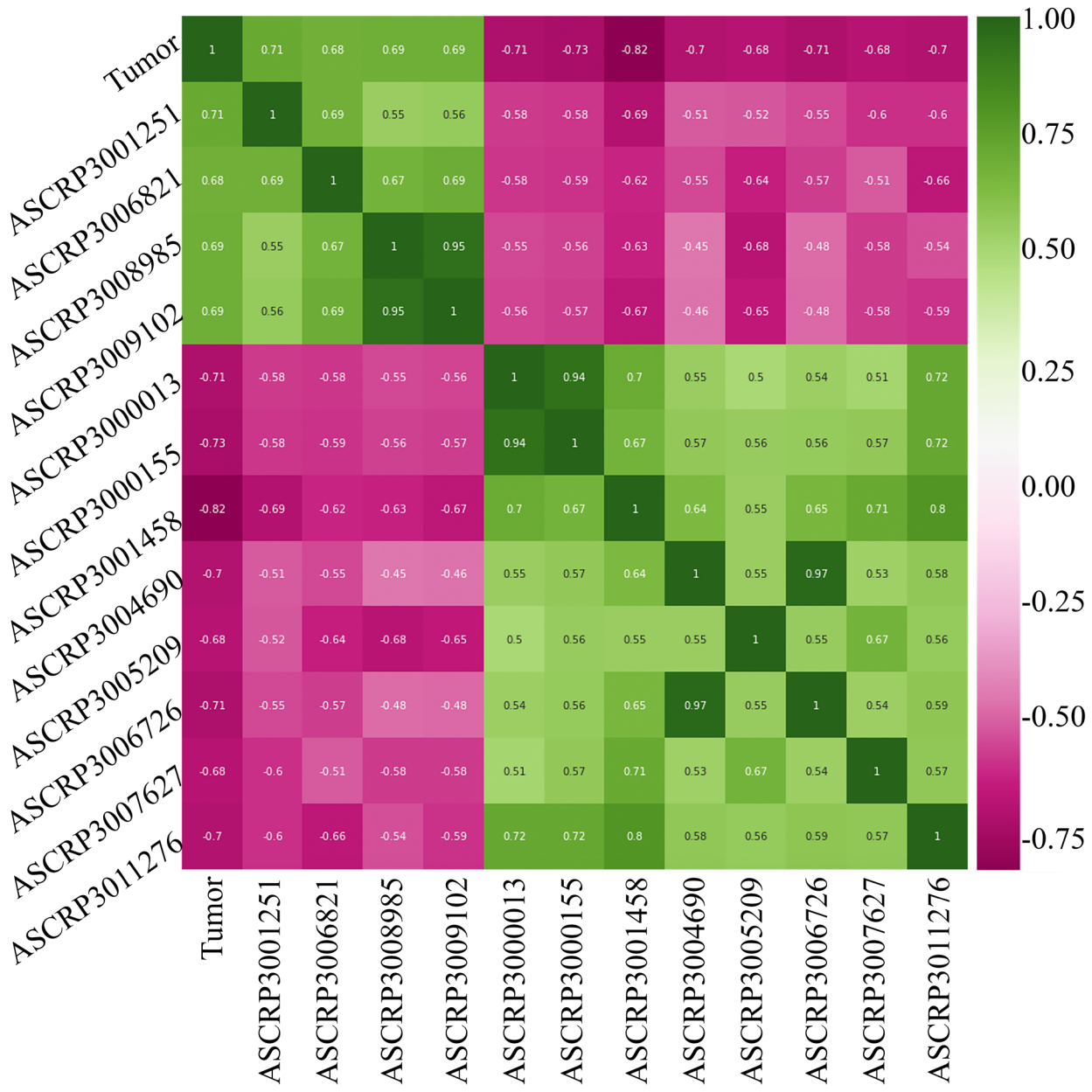

Figure 11. Pearson correlation heatmap generated with intensity of expression of top 12 deregulated circRNAs in healthy liver/non-tumor and tumor regions. X and Y-axis – circRNA type, and first column depicts Pearson correlation value for each of top 12 circRNA for tumor. Other columns correlations between circRNAs based on their expression intensity. Data generated in Python of top 12 deregulated circRNAs. Purple regions indicate strong negative correlation, and green areas indicate strong positive correlation. The most significant column is the first column, as it shows the correlation with tumor occurrence.

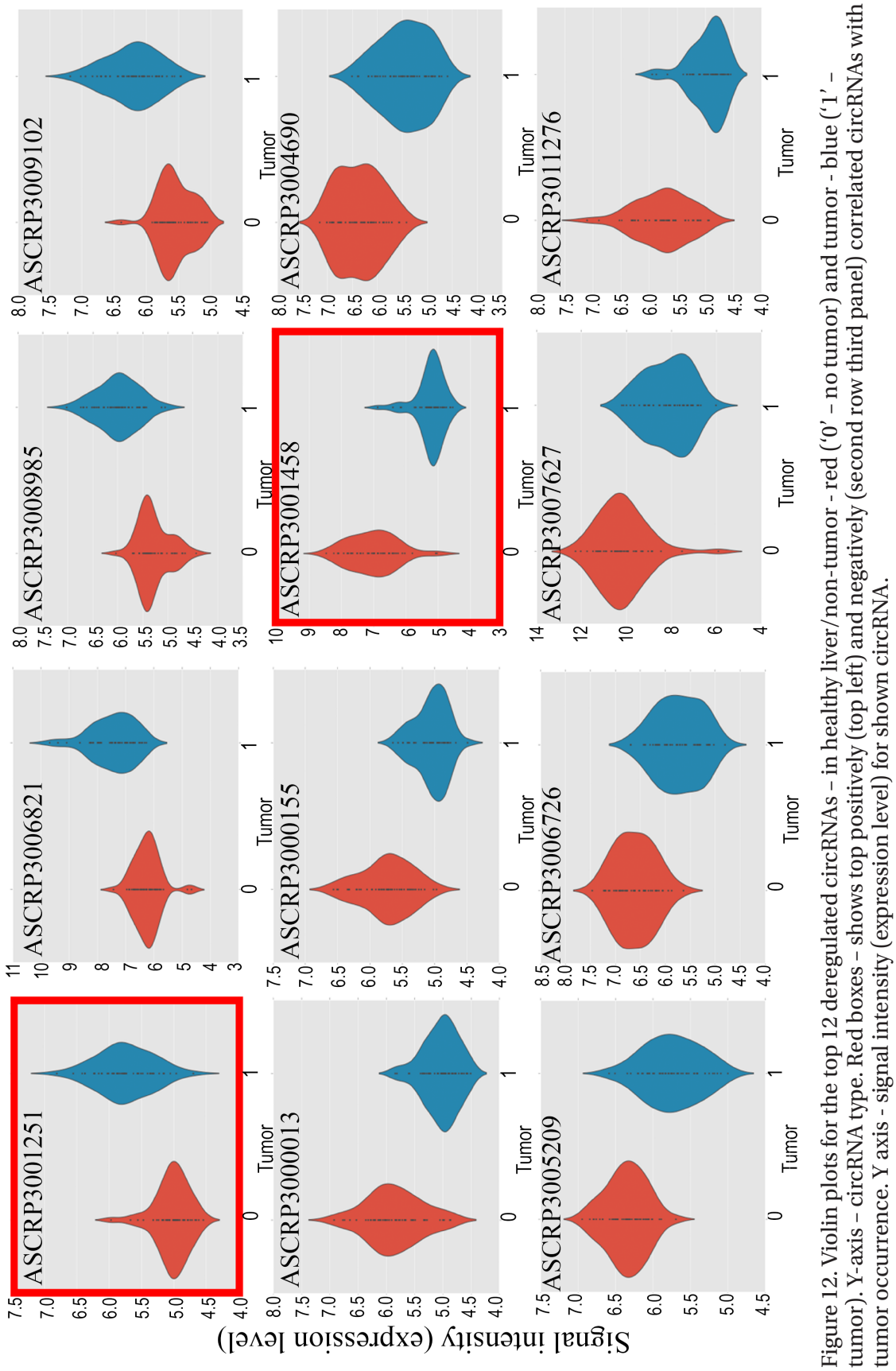

Figure 12. Violin plots for the top 12 deregulated circRNAs – in healthy liver/non-tumor - red ('0' – no tumor) and tumor - blue ('1' – tumor). Y-axis – circRNA type. Red boxes – shows top positively (top left) and negatively (second row third panel) correlated circRNAs with tumor occurrence. Y axis - signal intensity (expression level) for shown circRNA.

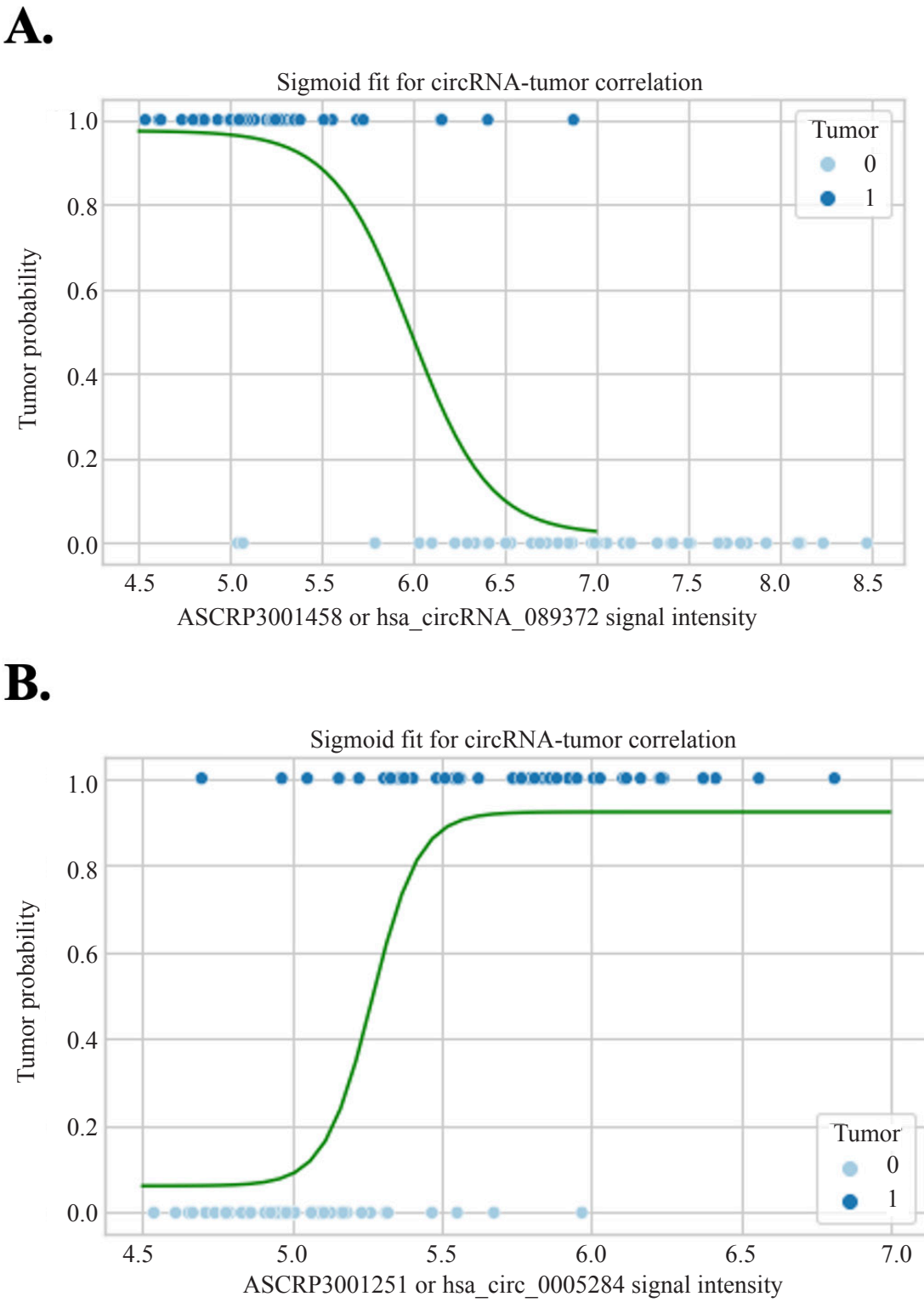

Figure 13. Sigmoid curve based on Pearson correlation for circRNA and tumor occurrence. (A) has_circRNA_089372 vs tumor probability. (B) has_circRNA_00005284 vs tumor probability.

Machine learning and model accuracies

Figure 14A shows the accuracies of seven machine learning classification models after trained by the dataset. The logistic regression scored the highest with 100% accuracy. Model accuracies for other machine learning algorithms were 'KNN': 83.33, 'Random Forest': 96.67, 'Decision Tree': 73.33, 'SVM': 96.67, 'Naive Bayes': 96.67, 'Gradient Boosting': 90.0 (Figure 14). There is a chance that the model might be overfitting (generalizing patterns in training data but performing poorly in test data), but through cross-validation this can be prevented. This was further confirmed by confusion matrix analyses (Figure 14B).

Figure 14B shows the confusion matrix for each of the seven machine learning models. Confusion matrices evaluate model performance by showing the number of predictions of each class. Table 1 shows what each number means in each quadrant of the confusion matrix:

	Predicted: Tumor	Predicted: No Tumor
Actual: Tumor	True positive	False Positive
Actual: No Tumor	False Negative	True negative

Table 1.

True Positive: # of correctly predicted tumor occurrences
False Positive: # of incorrectly predicted tumor occurrences
False Negative: # of incorrect predictions of no tumor (healthy)
True Negative: # of correct predictions of no tumor (healthy)

To better visualize the confusion matrix, a receiver operating characteristic (ROC) curve was plotted for all seven machine learning algorithms. An ROC curve illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Overall, many of the machine learning algorithm tested showed >0.9 sensitivity compared to random prediction, whose area under the curve (AOC) was at 0.5 (Figure 15). Model performances were analyzed using ROC curves and AUROC (Area under ROC curve). The ROC curve plots the True positive rate (TPR) vs False positive rate (FPR) (equations 5 and 6).

$$TPR\ or\ Recall\ or\ Sensity = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{5}$$

TPR = (# of True positives)/(True positives + False Negatives)

(False positive rate = (# of False Positives)/(True Negatives + False Positives)

$$FPR\ = 1 - specificity = \frac{\mathrm{FP}}{\mathrm{TN+FP}} \tag{6}$$

The AUROC is the area under the ROC curve. A good AUROC is near 1, which shows a good measure of separability (Tumor vs No Tumor). A poor model will have an AUROC near 0, which indicates the model gets every prediction incorrect. A model with AUROC of 0.5 cannot make any separation (cannot differentiate between Tumor and Non-Tumor circRNA samples). Figure 15 shows the ROC curve for all models. Most models had an AUROC near 1 and are therefore not visible (cross into the border). Logistic Regression, Support Vector Machine, and Random Forest had the highest AUROC scores.

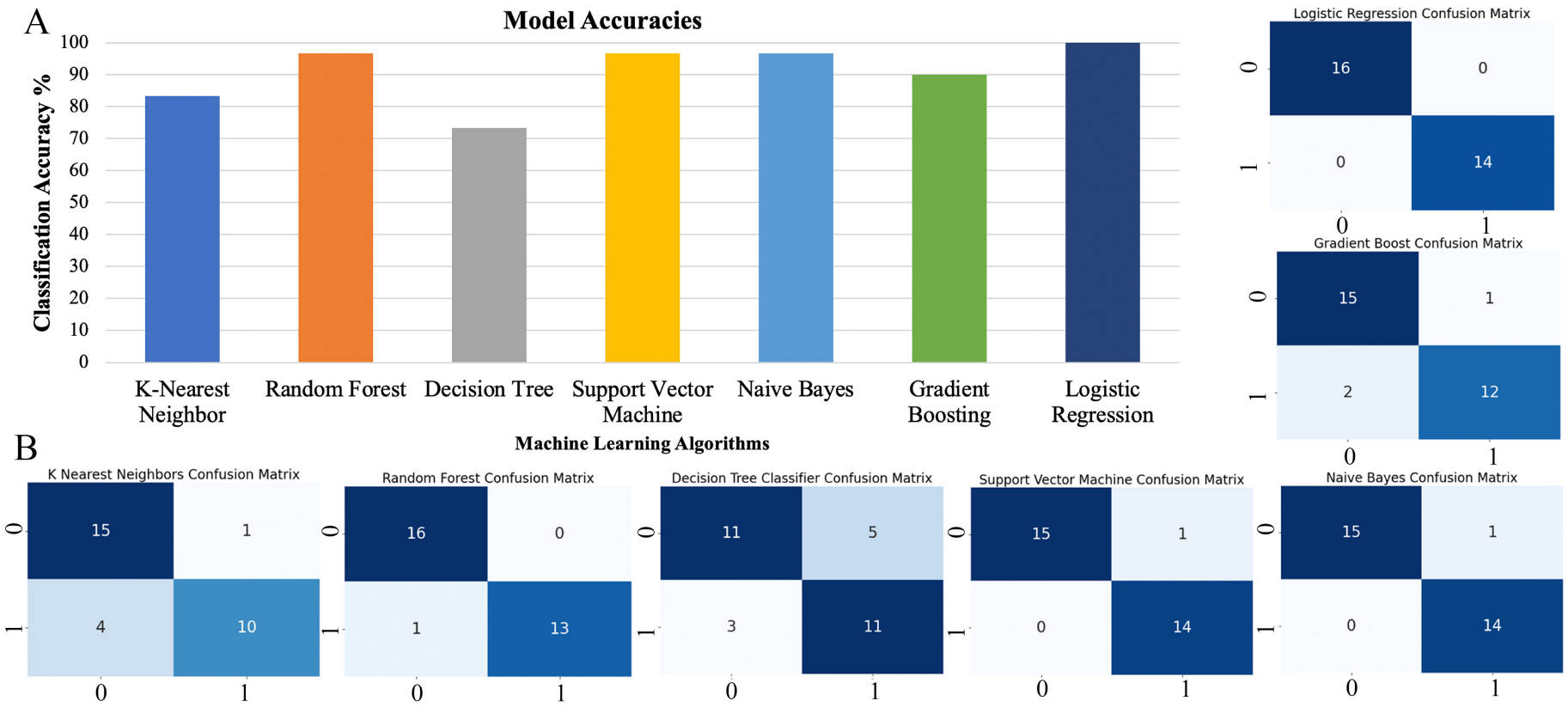

Figure 14. Graph showing ML model accuracies and confusion matrix. (A) Shows the accuracies of 7 machine learning classification models after training. Note that majority of the models (except decision tree) showed >80% accuracy with logistic regression showing 100% classification accuracy. (B) shows the confusion matrix for each of the 7 machine learning models. Left label (Y-axis) – predictive value; bottom (X-axis) – actual values. Positive value (1) and negative value (0). True positive are bottom right, false positive are bottom left, true negative are top left and false negative are top right.

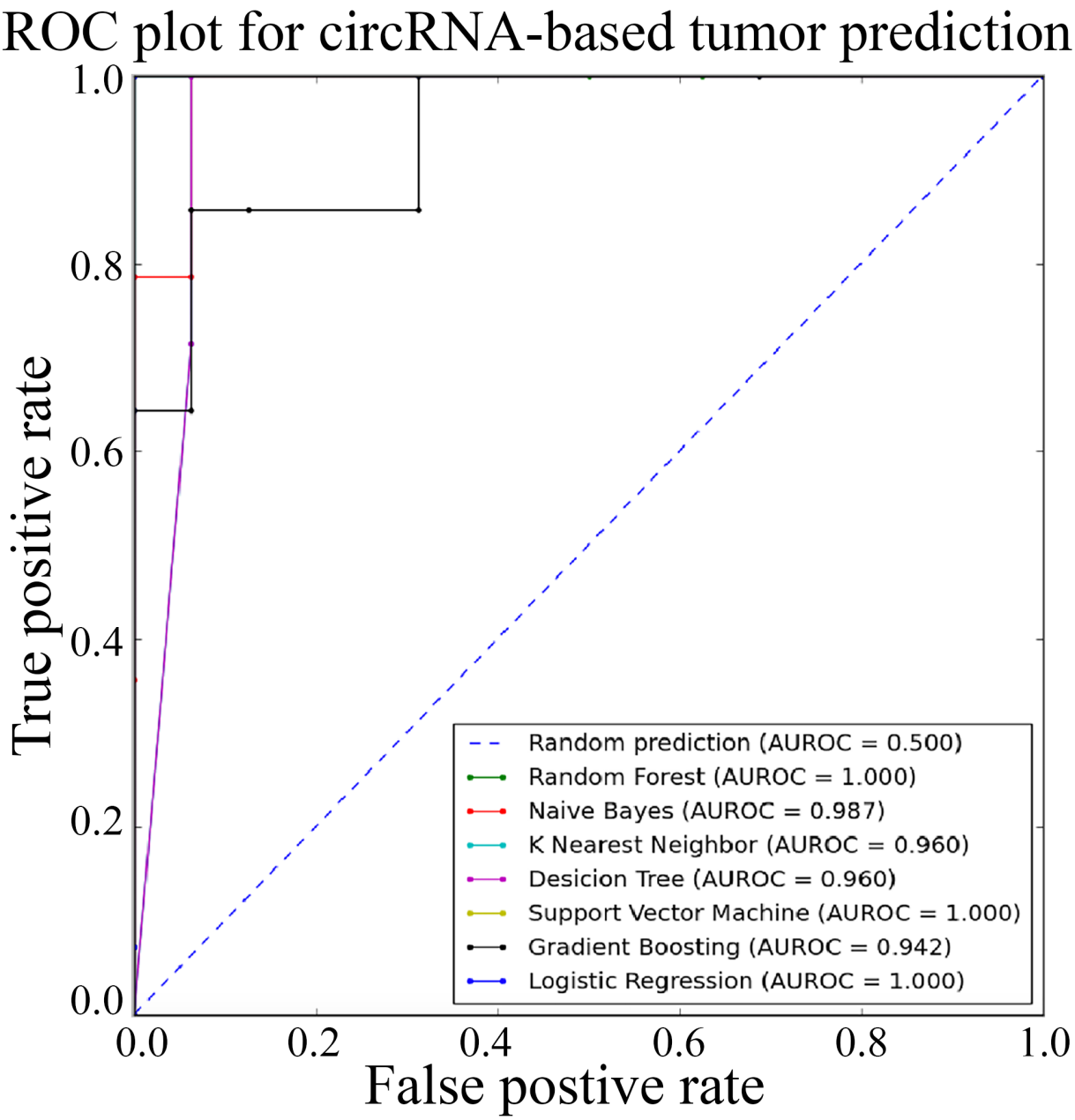

Figure 15. ROC curve for 7 different machine learning algorithms.

Stratified K-Fold Cross Validation: Next, Stratified K-fold Cross validation was performed in Python. In regular machine learning, the data was split using the train test split (70% of the data used to train models, and the other 30% used to test the model accuracy). Splitting the data with a train-test split has more bias and can lead to overfitting, so Stratified K-fold splits the data in a different manner (Figure 16A). K-fold cross validation involves splitting data into k equal folds (Figure 16). The first k-1 folds are used for training, and the remaining are used for testing. This is repeated for all k-folds, and the mean of the accuracies of each k-fold is returned. Stratified k-fold is similar, but involves splitting the data into folds not randomly, but based on the number of each class (if fold 1 has 15 tumor samples and 15 non-tumor samples, then fold 2 should have a roughly equal amount). This helps eliminate biases that

come with randomly splitting the data. Model accuracies for Stratified K-Fold: KNN: 87.784, Random Forest: 95.928, Decision Tree: 88.826, Support Vector Machine: 95.928, Naïve Bayes: 95.928, Gradient Boosting: 90.846, Logistic Regression: 95.928

Figure 16B shows the model accuracies of each model after performing Stratified-K-Fold Cross Validation. Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest had the highest accuracy of nearly 96%.

Next, original accuracy prior to cross validation vs the accuracy after cross validation was plotted for all machine learning models (Figure 17). After Stratified K-Fold cross validation, the KNN, Decision Tree, and Gradient boosting algorithm accuracies increased while the Random Forest, Support Vector Machine, Naive Bayes, and Logistic Regression models

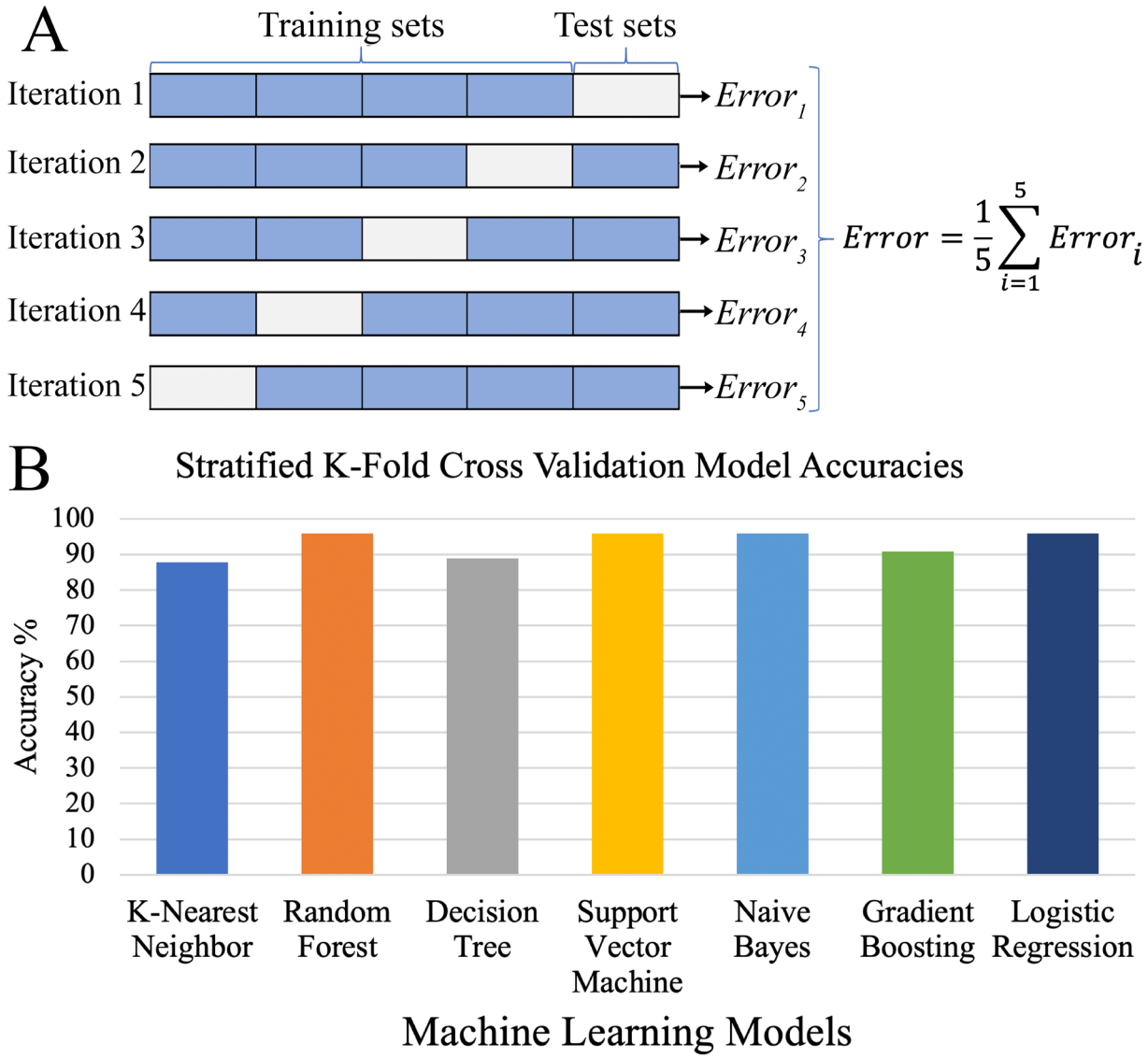

Figure 16. Cross-validation of ML algorithm (https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b). (A) Stratified K-fold Cross validation was performed in Python. (B) Shows the cross validation of 7 machine learning classification models after training.

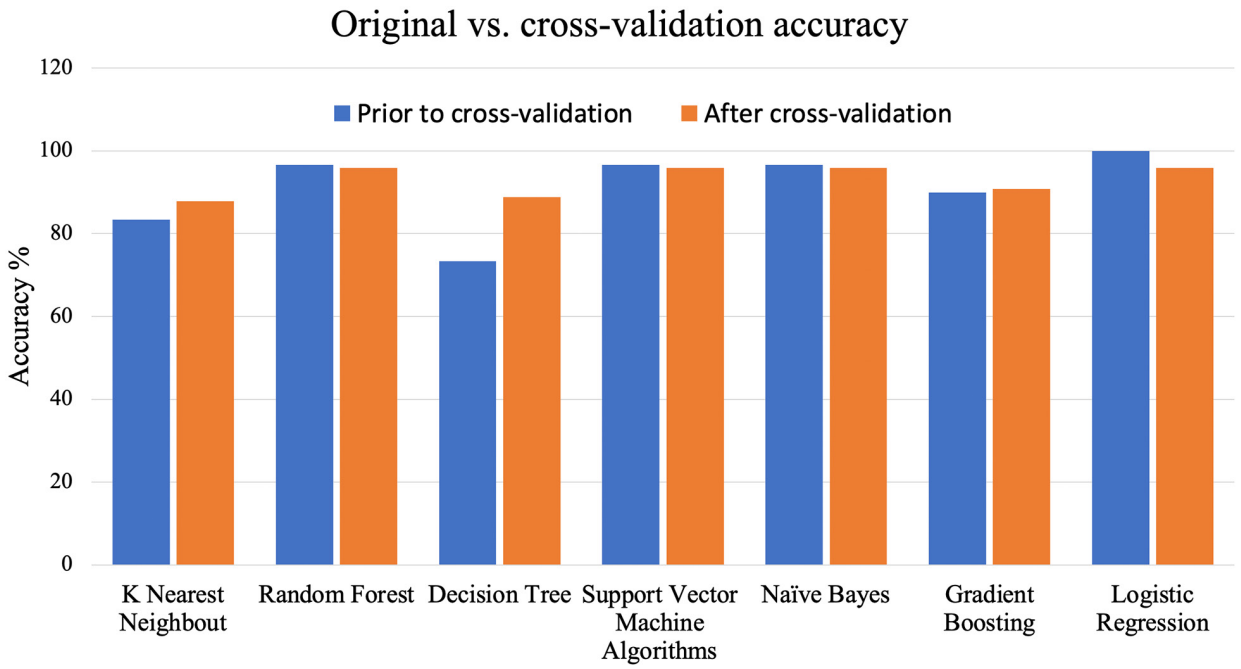

Figure 17. Comparison between original accuracy and validation accuracy of ML algorithm.

decreased. It is likely that the algorithms that decreased in accuracy were slightly overfitting prior to cross validation.

Finally, re-plotting ROC and AUROC scores for each model for Cross Validation showed that the Naive Bayes and Logistic Regression had the largest AUROC scores of 0.992 (Figure 18).

Cytoscape Network Analysis: Cytoscape is an open-source bioinformatics analyses software tool/program that helps in visualization of gene and protein interaction networks. Using Cytoscape, the entire network of two circRNAs that are downregulated (hsa_circRNA_089372) or upregulated (hsa_circ_0005284) in tumor was plotted along with their corresponding miRNAs and mRNAs or the "hub genes".

Figure 19 is a circRNA-miRNA-gene network created in Cytoscape. “Bindingp” is a metric predicted by miRWalk which shows the probability/score of an interaction between a miRNA and mRNA/gene. Top two circRNAs are shown in two orange octagons, miRNA is shown in purple circles, and genes/mRNAs are shown in squares. Fewer validated and high-scoring gene targets were found for hsa-mir-626 and hsa-mir-639 as shown. Because of the vast quantity of genes predicted by mirWalk, the researcher wanted to filter all predicted genes to the top 4 for each circRNA. These are hub genes and usually play critical roles in HCC progression and development.

To find the hub genes, first a PPI (Protein-Protein Interaction) or gene interaction network was constructed using the STRING algorithm in Cytoscape for each circRNA (Figure 20). Note that genes in hsa_circ_0005284 (upregulated in tumor as shown in Figure 20A) have no significant interaction with other genes and were not investigated in this study.

On the other hand, the PPI network of hsa_circRNA_089372 showed far more gene inter-

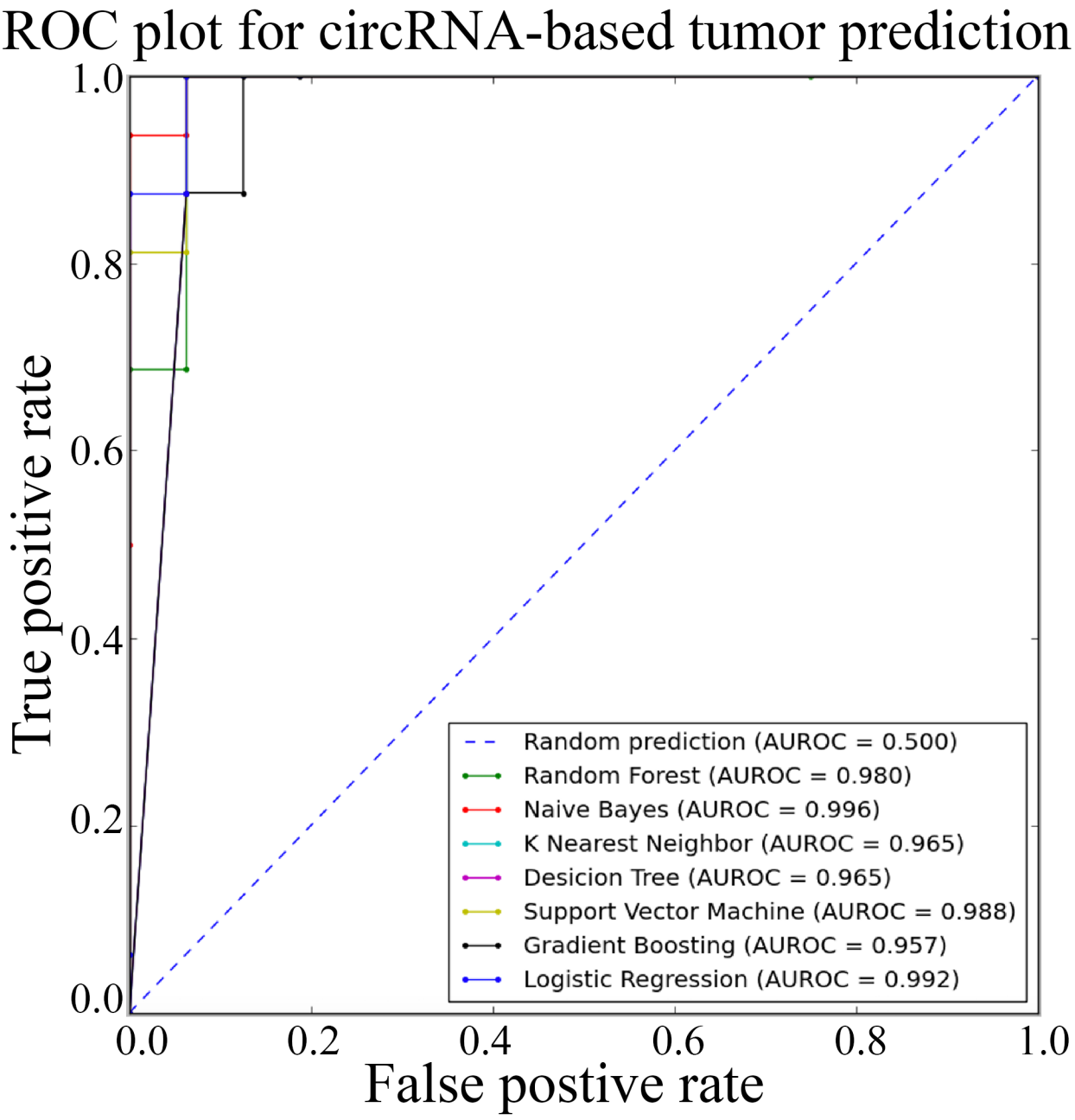

Figure 18. Shows the ROC plot and AUROC scores for each model for Cross Validation.

actions compared to hsa_circ_0005284 (Figure 20B), which could mean that it likely plays a more critical role in HCC progression and development (also proven by higher correlation value w/ tumor). Next, the MCODE algorithm (finds clusters of genes in the network) was performed for both PPI networks to determine the final hub gene. These clusters contain hub genes, which was used for immunogenicity analysis. After MCODE algorithm, the resulting hub genes involvement in cancer were validated using Human Protein Atlas (https://www.proteinatlas.org/).

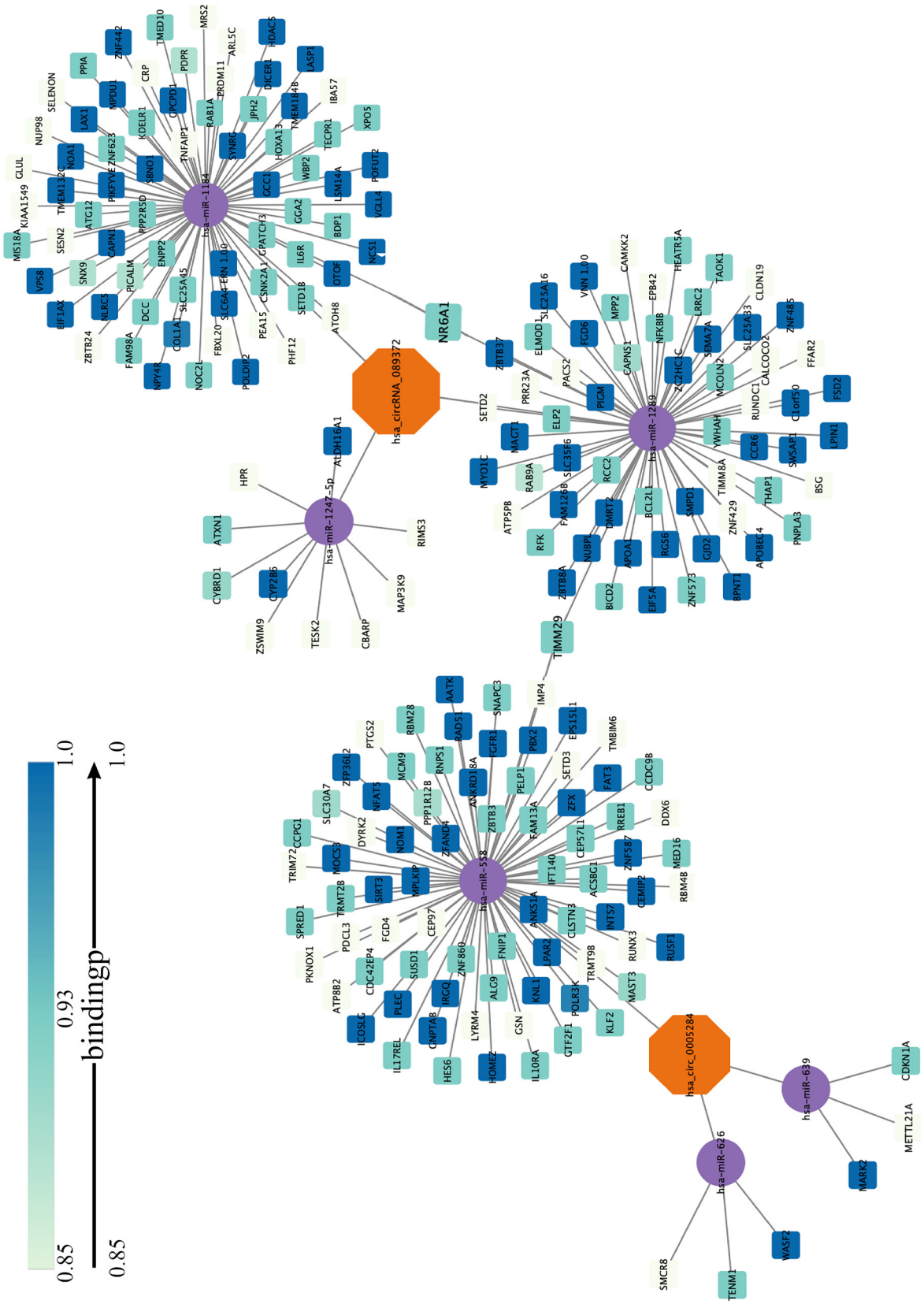

Figure 19. CircRNA-miRNA-gene network created in Cytoscape. CirRNA (orange octagon); miRNA (purple circles; mRNA (hub genes) (blue or cyan squares).

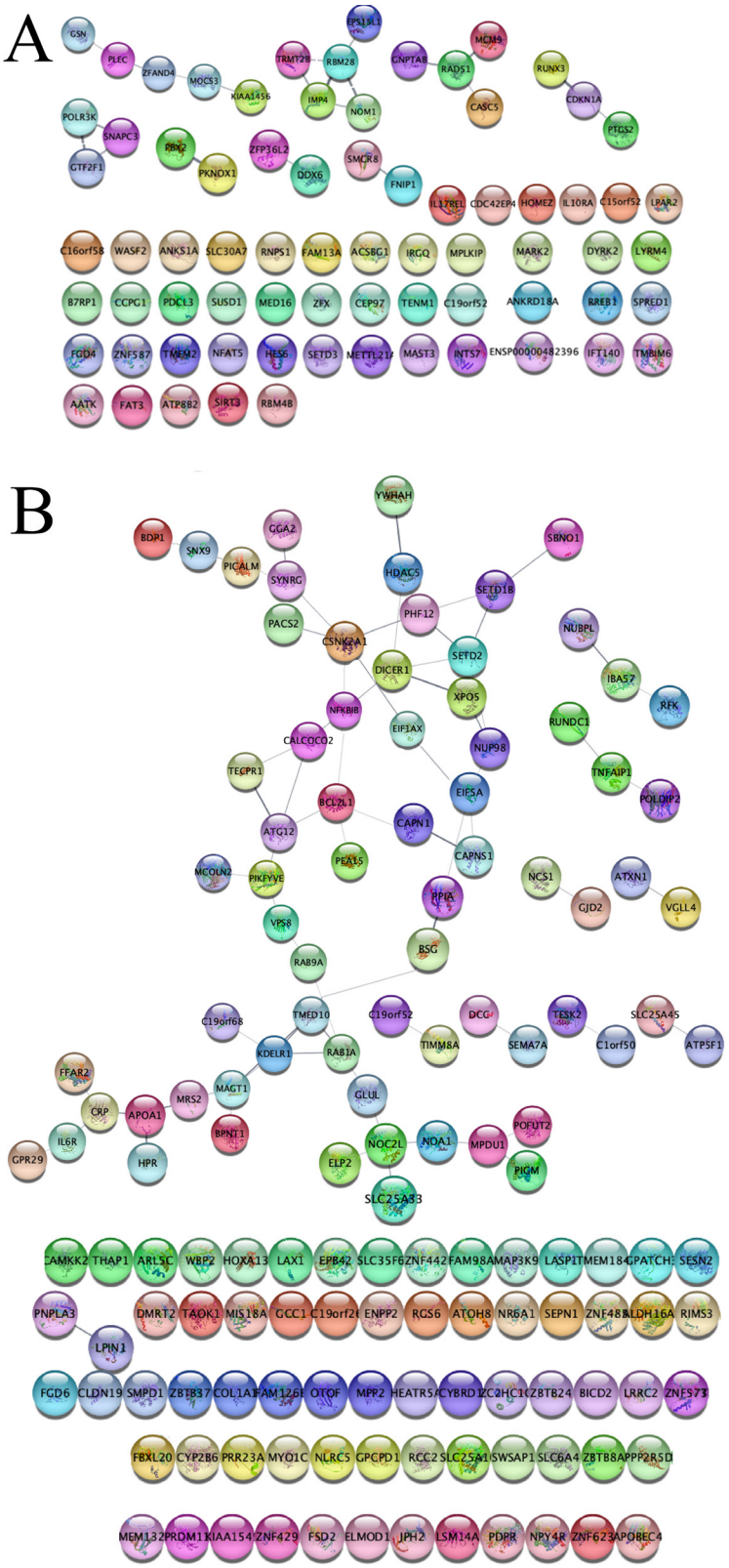

Figure 20. Protein-protein interaction (PPI) network of hub genes regulated by (A) hsa_circ_0005284 (upregulated in tumor) and (B) hsa_circRNA_089372 (downregulated in tumor) based on STRING algorithm in Cytoscape.

Validation and Survival Analysis: As seen in Table 2 (below), 7/8 of the hub genes are known to play an unfavorable or favorable prognostic markers in a variety of cancers based on the human protein atlas. Further, 3/8 of the hub genes are known to show an unfavorable prognosis in liver cancer (HCC), although additional papers outside of the Human Protein Atlas were found that validated hub genes other than these 3 for liver cancer (for example *RAB1A*). Next, human protein atlas survival analysis was performed for these 3 hub genes validated for liver cancer, and the graphs are shown below.

hsa-circ-0005284 hub genes(below)	**Prognosis and Cancer type (from Human Protein Atlas)**
TRMT2B	Prognostic marker in renal cancer (favorable) and breast cancer (unfavorable)
RBM28	Prognostic marker in liver cancer (unfavorable), endometrial cancer (unfavorable) and melanoma (unfavorable)
IMP4	Prognostic marker in liver cancer (unfavorable)
NOM1	Not prognostic
hsa-circ-089372 hub genes (below)	
MAGT1	Prognostic marker in breast cancer (unfavorable), pancreatic cancer (unfavorable) and melanoma (unfavorable)
KDELR1	Prognostic marker in liver cancer (unfavorable) and head and neck cancer (unfavorable)
TMED10	Prognostic marker in thyroid cancer (unfavorable)
RAB1A	Prognostic marker in head and neck cancer (unfavorable)

Table 2. Hub genes predicted by MCODE and Human Protein Atlas validation.

Survival Analysis: Table 3 shows the survival analysis graphs for the 3 circRNA targeted hub genes that were validated for liver cancer. Out of the three hub genes, *RBM28* expression levels can most significantly determine survival probability, then *KDELR1* and finally *IMP4* (*RMB28* had lowest p-value). In all three graphs, the survival probability for high gene expression is lower than low gene expression, which shows unfavorable prognosis for high gene expression. The expression cutoff is highest in *KDELR1* (highest *FPKM*), which could mean that *KDELR1* has higher median gene expression levels. It can also be noted that *RBM28* has the lowest 5-year survival rate for high expression (0% survival rate).

Hub gene	**Survival analyses score**
RBM28	Median follow up time (years) :1.63 P score: 2.7e-11 5-year survival for high expression: 0% 5-year survival low expression: 54% High vs low expression is determined by the cutoff 1.52 Fragments Per Kilobase of transcript per Million mapped reads (FPKM). High expression is above this cutoff, low expression is below this cutoff.

IMP4	Median follow up time (years) :1.63 P score: 0.00047 5-year survival for high expression: 35% 5-year survival low expression: 57% High/Low expression cutoff: 15.27 FPKM
KDELR1	Median follow up time (years) :1.63 P score: 0.00016 5-year survival for high expression: 40% 5-year survival low expression: 54% High/Low expression cutoff: 78.45 FPKM

Table 3. Survival analyses values for all three genes.

Immunogenicity Predictions: Immunogenicity predication can be done by a variety of ways. Table 4 is the graph for the number of HLA T-cell epitopes (detailed in appendices) found in the peptide sequences for each of the circRNA targeted hub genes. Hub genes targeted by hsa_circ_089372 have a higher number of T-cell epitopes in general. Because each hub gene has a different length (number of amino acids), the number of T-cell epitopes were normalized to a length of 1000 amino acids using (equation 7):

$$\frac{\#\ of\ T{-}cell\ epitopes}{(length\ of\ protein\ sequence) * 1000} \tag{7}$$

The number of T-cell epitopes predicted for each hub genes is shown in Table 4 (below). For a negative control, a random, fake protein sequence was created the same size of each hub gene to see how many epitopes were predicted. In all random protein sequences, 0 epitopes were found.

Gene	# Of T-cell epitopes predicted for each hub gene
TRMT2B	2
RBM28	18
IMP4	78
NOM1	20
MAGT1	110
KDELR1	52
TMED10	151
RAB1A	151

Table 4. # of T-cell epitopes predicted for each hub gene (protein)

Next, epitope clustering histograms for the top 4 genes were created. Regions in the protein sequences where there are a high number of T-cell epitopes have higher immunogenicity and are great targets for a cancer vaccine. These graphs show the raw number of epitopes (without normalization). Figure 21A-D shows the number of T-cell epitopes in each region of the protein sequences for the top 4 hub genes with the most T-cell epitopes. Peptide regions where there is a high epitope concentration have high immunogenicity.

The figure below shows the highest immunogenicity regions for each of the 4 hub genes, and the peptide sequence in this region of the protein are great targets for a peptide cancer vaccine.

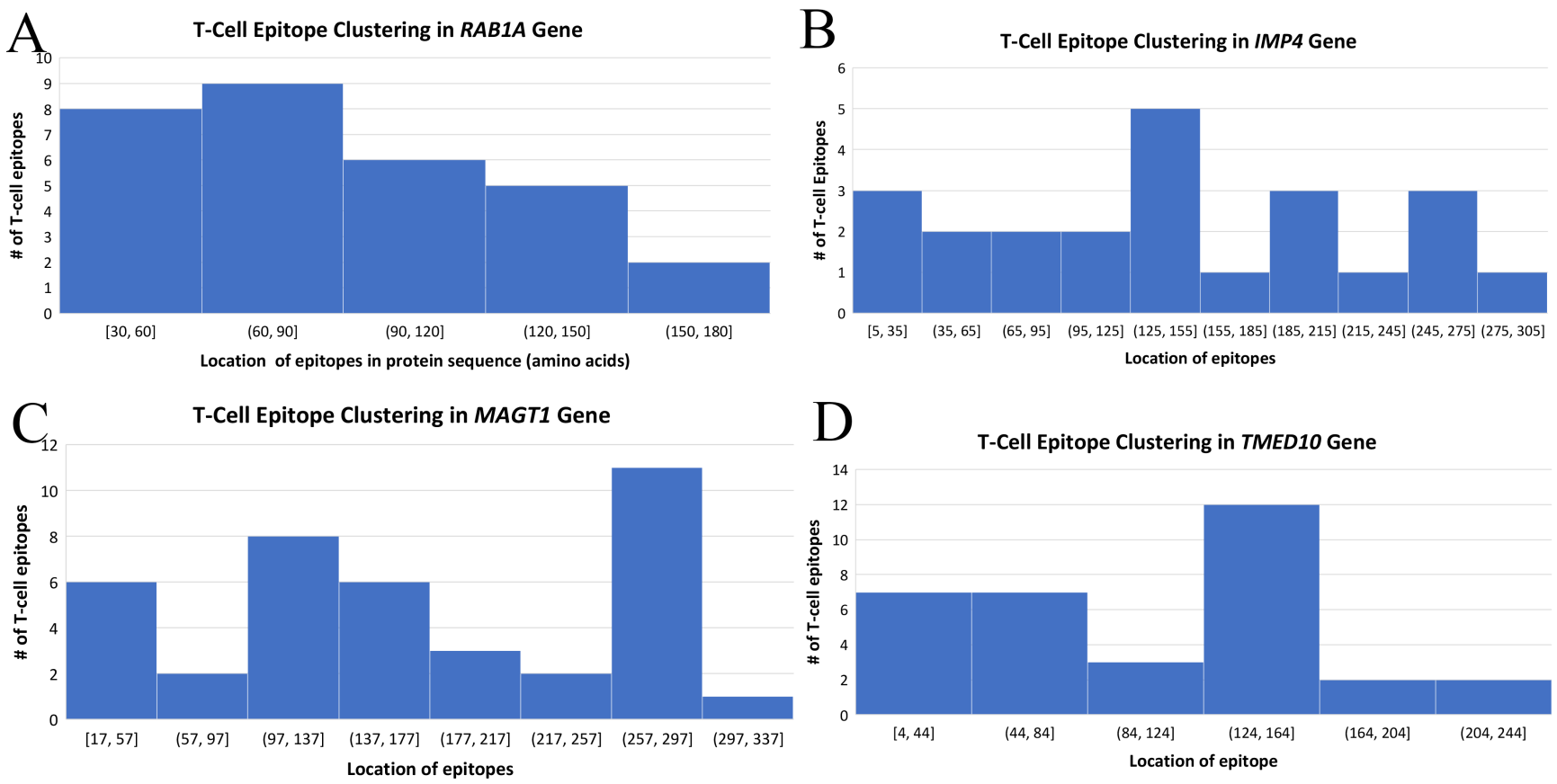

Figure 21. Shows the number of T-cell epitopes in each region of the protein sequences for the top 4 hub genes with the most T-cell epitopes. Peptide regions where there is a high epitope concentration have high immunogenicity. (A) *RAB1A;* (B) *IMP4* (C) *MAGT1* (D) *TMED10*. Note that the actual number (and not normalized) of T cell epitopes in each candidates genes shown.

4. DISCUSSION

This study aimed to design and build multiple machine learning models to predict the occurrence of hepatocellular carcinoma in patients and to predict the immunogenicity of deregulated circRNA gene targets for a potential immunotherapy or cancer vaccine. With Pearson correlation and information gain statistical methods, the top deregulated circRNAs in tumor tissue were determined. All machine learning models used could predict the occurrence of tumors with circRNA expression data with >85% accuracy and many reached nearly 100% classification accuracy. The top machine learning models upon stratified k-fold cross validation were logistic regression and naive bayes. Correlation analysis for the top 2 deregulated circRNAs showed that hsa_circ_0005284 is strongly upregulated in tumor, and hsa_circrna_089372 is downregulated in tumor. From Cytoscape network analysis, miRWalk, and circInteractome, the top hub genes were determined, and 7 of the 8 hub genes were validated to be cancer-prognosis biomarkers with the human protein atlas. Finally, with t-cell epitope analysis in hub gene peptide sequences, it was determined that IMP4, MAGT1, TMED10, and RAB1A have immunogenic potential due to high epitope count and concentration and are good candidates for vaccine targeting. IMP4 was the most immunogenic out of the hub genes validated to be a prognostic marker in liver cancer (the rest were RBM28, IMP4, and KDELR1), but RAB1A and TMED10 showed the highest immunogenicity and can still be potential immunotherapeutic candidates for their respective cancers. However, a study by Yang, et al. (2015) showed and validated RAB1A to be prognostic in hepatocellular carcinoma, although not from the human protein atlas. This indicates some genes can still be prognostic of liver cancer even if they are not validated from the protein atlas. Additionally, novel, unexplored oncogenes for other cancers could be predicted with the pipeline developed in this study since it was validated to work and predict oncogenes by the human protein atlas. Recent reviews provide a comprehensive review on the role of circRNAs in liver and other cancers (Liu et al., 2020; Shen et al., 2021; Su et al., 2019).

In future studies, other methods of machine learning could be explored, including neural networks, hierarchical clustering, and more. These methods could find new trends in circRNA expression data which could not be picked up by classification models. Additionally, pharmacogenomic therapies could be explored such as drug targeting and drug repurposing for the hub genes found in this study. More immunotherapeutic options can also be explored such as immune checkpoint inhibitors and monoclonal antibodies. CircRNA expression data can also be explored for other cancers and the same machine learning pipeline created in this study can be applied to all other cancers.

One of the biggest limitations to this project is data size. The circRNA expression data was for 98 tissue samples. In the future, the researcher might try to find larger datasets on circRNA expression and compare model accuracies. Another limitation is immunogenicity prediction. Multiple methods can be used for immunogenicity prediction and T-cell epitope concentration is only one of them. It is possible that epitope concentration may not be able to fully gauge the immunogenicity of hub gene peptide sequences, so more methods could be explored in the future.

The main future prospects of this project, if validated in biological system, is potential development of active immunotherapy (vaccine) /passive immunotherapy (monoclonal anti-

bodies) therapies for any cancer if circRNA data is given. The novel methodology developed in this project (circRNA and immunogenicity) can help find immunotherapy for any cancer with circRNA data. The machine learning models can be a valuable tool for healthcare professionals because they were over 90% accurate in predicting/diagnosing cancer using circRNA data, potentially replacing existing diagnosing methods such as expensive MRI/CT scans/machines).

AUTHOR INFORMATION

***Corresponding Author**

Aditya K. Koushik
La Cueva High School
7801 Wilshire Ave NE
Albuquerque, NM 87122
Email: adityakoushik1234@gmail.com

ACKNOWLEDGMENT

I would like to thank Dr. Nikolaos Mellios Associate Professor at the University of New Mexico for answering my questions on circRNA data, and my parents for their encouragement and helping me with my project.

REFERENCES

Balogh, J., Victor, D., 3rd, Asham, E. H., Burroughs, S. G., Boktour, M., Saharia, A., . . . Monsour, H. P., Jr. (2016). Hepatocellular carcinoma: a review. *J Hepatocell Carcinoma, 3*, 41-53. doi:10.2147/JHC.S61146

Cavanagh, M., and Findlay, E. M. . ((n.d.)). T-cell activation. *British Society for Immunology*. Retrieved from https://www.immunology.org/public-information/bitesized-immunology/systems-processes/t-cell-activation

Conn, S. J., Pillman, K. A., Toubia, J., Conn, V. M., Salmanidis, M., Phillips, C. A., . . . Goodall, G. J. (2015). The RNA binding protein quaking regulates formation of circRNAs. *Cell, 160*(6), 1125-1134. doi:10.1016/j.cell.2015.02.014

de Candia, P., Prattichizzo, F., Garavelli, S., & Matarese, G. (2021). T Cells: Warriors of SARS-CoV-2 Infection. *Trends Immunol, 42*(1), 18-30. doi:10.1016/j.it.2020.11.002

De Groot, A. S., Moise, L., Terry, F., Gutierrez, A. H., Hindocha, P., Richard, G., . . . Martin, W. D. (2020). Better Epitope Discovery, Precision Immune Engineering, and Accelerated Vaccine Design Using Immunoinformatics Tools. *Front Immunol, 11*, 442. doi:10.3389/fimmu.2020.00442

Dudekula, D. B., Panda, A. C., Grammatikakis, I., De, S., Abdelmohsen, K., & Gorospe, M. (2016). CircInteractome: A web tool for exploring circular RNAs and their interacting proteins and microRNAs. *RNA Biol, 13*(1), 34-42. doi:10.1080/15476286.2015.1128065

Garnelo, M., Tan, A., Her, Z., Yeong, J., Lim, C. J., Chen, J., . . . Chew, V. (2017). Interaction between tumour-infiltrating B cells and T cells controls the progression of hepatocellular carcinoma. *Gut, 66*(2), 342-351. doi:10.1136/gutjnl-2015-310814

Greene, J., Baird, A. M., Brady, L., Lim, M., Gray, S. G., McDermott, R., & Finn, S. P. (2017). Circular RNAs: Biogenesis, Function and Role in Human Diseases. *Front Mol Biosci, 4*, 38. doi:10.3389/fmolb.2017.00038

Henderson, S. R. R. (2021). What are T cells? Retrieved from https://www.news-medical.net/health/What-are-T-Cells.aspx

Liu, J., Zhang, X., Yan, M., & Li, H. (2020). Emerging Role of Circular RNAs in Cancer. *Front Oncol, 10*, 663. doi:10.3389/fonc.2020.00663

Llovet, J. M., Kelley, R. K., Villanueva, A., Singal, A. G., Pikarsky, E., Roayaie, S., . . . Finn, R. S. (2021). Hepatocellular carcinoma. *Nat Rev Dis Primers, 7*(1), 6. doi:10.1038/s41572-020-00240-3

Muhammad, S. A., Zafar, S., Rizvi, S. Z., Imran, I., Munir, F., Jamshed, M. B., . . . Zhang, Q. (2020). Experimental analysis of T cell epitopes for designing liver cancer vaccine predicted by system-level immunoinformatics approach. *Am J Physiol Gastrointest Liver Physiol, 318*(6), G1055-G1069. doi:10.1152/ajpgi.00068.2020

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res, 13*(11), 2498-2504. doi:10.1101/gr.1239303

Shen, H., Liu, B., Xu, J., Zhang, B., Wang, Y., Shi, L., & Cai, X. (2021). Circular RNAs: characteristics, biogenesis, mechanisms and functions in liver cancer. *J Hematol Oncol, 14*(1), 134. doi:10.1186/s13045-021-01145-8

Sticht, C., De La Torre, C., Parveen, A., & Gretz, N. (2018). miRWalk: An online resource for prediction of microRNA binding sites. *PLoS One, 13*(10), e0206239. doi:10.1371/journal.pone.0206239

Su, M., Xiao, Y., Ma, J., Tang, Y., Tian, B., Zhang, Y., . . . Wang, W. (2019). Circular RNAs in Cancer: emerging functions in hallmarks, stemness, resistance and roles as potential biomarkers. *Mol Cancer, 18*(1), 90. doi:10.1186/s12943-019-1002-6

Yang, Y., Hou, N., Wang, X., Wang, L., Chang, S., He, K., . . . Huang, C. (2015). miR-15b-5p induces endoplasmic reticulum stress and apoptosis in human hepatocellular carcinoma, both in vitro and in vivo, by suppressing Rab1A. *Oncotarget, 6*(18), 16227-16238. doi:10.18632/oncotarget.3970

Zhang, X., Wang, S., Wang, H., Cao, J., Huang, X., Chen, Z., . . . Xu, Z. (2019). Circular RNA circNRIP1 acts as a microRNA-149-5p sponge to promote gastric cancer progression via the AKT1/mTOR pathway. *Mol Cancer, 18*(1), 20. doi:10.1186/s12943-018-0935-5