International Journal of Smart Computing and Artificial Intelligence
International Institute of Applied Informatics
2022, Vol. 6, No. 1, IJSCAI628

Extraction of Genes and Transcripts Associated with Liver Cancer Using Machine Learning

Koshiro Sekine [*], Teruhisa Hochin [*],
Hiroki Nomiya [*], Hideki Yoshida [*]

Abstract

The rapid development of large-scale genome analysis technology in recent years has facilitated the acquisition of genome data, but it is difficult to extract effective information from a large amount of data. To solve this problem, machine learning has been attracting attention. In this paper, we used machine learning to extract genes and transcripts associated with liver cancer. Liver cancer is difficult to cure completely and is generally treated by surgery, making it difficult to treat elderly people or those with reduced physical strength. As an overview of the method, using the liver cancer dataset of NBDC, genes and transcripts extracted by using statistical hypothesis tests were used as input to the machine learning to create a classifier. Then, genes and transcripts with high contribution rate to the classification were extracted from the classifier. As a result, we obtained genes and transcripts that were considered to be associated with liver cancer from the created classifier. The results of this paper are expected to contribute to the development of gene therapy for liver cancer. In addition, since the method in this paper is not specialized for liver cancer, it can be expected to be applied to other cancers.

Keywords: Differential Gene Expression, Differential Transcript Expression, Feature Extraction, Machine Learning

1 Introduction

Genome research has made great strides since the completion of the Human Genome Project in 2003. In 2005, the Next Generation Sequencing (NGS) appeared, making it possible to read the entire genome in a few months. In 2007, NGS identified the first whole living human genome in about two months. In recent years, rapid advances in large-scale genome analysis technology, RNA-seq, have made it possible to decode whole genomes in a few days and to analyze many types of genes in a single experiment.

In these circumstances, considerable progress has been made in cancer genomic medicine. In particular, rapid advances in large-scale genome analysis technology have resulted in a comprehensive gene analysis. Analysis of the human genome can provide fundamental treatment for cancers caused by genetic mutations. In June 2019, insurance

[*] Kyoto Institute of Technology, Kyoto, Japan

coverage for genetic testing began in Japan, and genomic medicine has entered the practical stage for some cancers. However, due to insufficient understanding of human genes, the detection rate of gene mutations useful for treatment is said to be about 50% and the possibility of compatible drugs is about 10%. In addition, it is said that it takes several weeks for experts to extract mutations with significant differences from tens or hundreds of thousands of characteristics by analysis, and to identify the etiology from experiments and researches on papers.

At the moment, human resources are insufficient to analyze the huge amount of data. Although genetic analysis technology has advanced, it is difficult to find useful data from the huge amount of data obtained through experiments, and it takes time to elucidate genes. Machine learning draws attention as an automated approach because it can break through this circumstance.

The purpose of this paper is to extract genes and transcripts, which are thought to be associated with cancer, by using machine learning. We particularly focus on a liver cancer, because the cancer is difficult to cure completely and is generally treated by surgery, making it difficult to treat elderly people or those with reduced physical strength. It also has a lower five-year survival rate than other cancers. In addition, as of September 2019, it has been excluded from cancer genetic testing in Japan. In light of these facts, there is a need for genomic therapy that could provide medical and fundamental treatment for liver cancer.

In our previous work [1], we have used genes and transcripts, extracted by statistical hypothesis test, to classify liver data into non-tumor tissues and tumor tissues. And we enumerated the genes and transcripts that have a high contribution to the classification by a random forest, and extracted the genes which are considered to be pathogenic. As a result, we have succeeded in extracting eight candidate causal genes which have not been predicted to associate with liver cancer previously. However, more candidates could be obtained by reviewing the methods of machine learning and statistical hypothesis test.

Therefore, in this paper, in addition to the previous methods [1], we use a Light Gradient Boosting Machine (LightGBM), a Support Vector Machine Classification (SVC), and a Linear SVC for machine learning, and sleuth for statistical hypothesis test. As a result, we succeeded in extracting fifteen candidate causal genes that have not been previously predicted to be associated with liver cancer.

The remainder of this paper is structured as follows. Section 2 presents related work. Section 3 describes our method. Section 4 shows the experimental resulting genes and transcripts which have a high contribution to the classification. Section 5 discusses the results based on the experimental results. Section 6 concludes this paper.

2 Related Work

Ide et al. [2] successfully extracted non-coding RNAs (ncRNAs) associated with acute lung injury by using random forest. As data, they used microarray data containing messenger RNAs (mRNAs) and ncRNAs that recorded genetic changes associated with inflammation occurring in the brain during acute lung injury in mice. As the number of genes in microarray data is much larger than the number of samples, it is necessary to select the genes used for the random forest. Therefore, after normalization of gene expression levels in microarray data, they extracted only the genes whose gene expression levels fluctuated between affected and unaffected mice. They used the extracted genes as training data for the random forests, and extracted ncRNAs with a high classification contribution from the generated

models.

In this paper, the threshold of the accuracy of the generated model was set to 0.92 or more, which is the average accuracy of their models.

Díaz-Uriarte et al. [3] used the random forest and other methods to extract genes suitable for cancer classification. As data, they used gene expression levels obtained from microarray datasets of various types of cancers. They compared the random forest with no info, Support Vector Machine (SVM), k Nearest Neighbor (KNN), Diagonal Linear Discriminant Analysis (DLDA), Shrunken centroids (SC), and Nearest Neighbor (NN) by varying error rate.

As a result, the random forest had the highest accuracy of all. In addition, the evaluation was performed according to the difference of the number of features, the number of weak learners, and the node size, which are parameters of the random forest.

They concluded that:

- As for the number of features, the square root of the number of genes is good.
- As for the number of weak learners, the importance value of each feature becomes slightly stable by increasing the number. In the dataset used in their study, 2,000 or 5,000 is sufficient, and even if it increases, it is considered to be a negligible effect.
- As for the node size, there is no major change.

In this paper, we compared the random forest, LightGBM and SVC. Then, we used the one with the highest rating as the extraction model.

Okun et al. [4] investigated the problems of researches using random forests, such as [3], to extract genes suitable for cancer classification. They identified the following two problems with random forests for cancer classification based on gene expression.

- The complexity of the dataset reduces accuracy.
- The contribution of the random forest may not be reliable.

Gene extraction without considering them would result in incorrect conclusions on the biological relevance of the genes.

First, the solution to the problem of reducing accuracy due to the complexity of the dataset is to avoid datasets that contain many types of cancers. The experimental results showed that the Out-Of-Bag (OOB) error rate was about 0.4 for datasets containing more than nine cancers and about 0.2 for datasets containing one cancer. Therefore, it is desirable to use a random forest for a single cancer dataset.

Next, the solution to the possibility that the contribution of the random forest may not be reliable is to combine feature selection and random forests and to use Area Under the Curve (AUC) for evaluation of accuracy. Random forests generate a training set by bootstrapping, which results in duplication of data. This duplication makes the multiplicity and non-uniqueness of each tree unavoidable, and the contribution is not always a reliable indicator. Therefore, it is necessary to extract in advance the genes that are considered suitable for cancer classification by feature selection. The reason for using AUC is that it has been shown to be an accurate measure of accuracy even when the number of features changes.

In this paper, we only used a dataset of liver cancers and also used AUC to evaluate the machine learning methods.

Bray et al. [5] presented kallisto, an RNA-seq quantification program. In this program, fast quantification was achieved by improving the k-mer analysis, which is commonly used

in sequence search. As a result, the program is more than two orders of magnitude faster than conventional methods and can achieve the same level of accuracy. Using kallisto, they analyzed 30 million unaligned paired-end RNA-seq reads in less than 10 minutes on a standard laptop computer. This has eliminated a major computational bottleneck in RNA-seq analysis.

In this paper, we used kallisto to quantify RNA-seq expression.

Pimentel et al. [6] described a new program, sleuth, for differential analysis of RNA-seq data. This program is optimized for the output of kallisto. The method of sleuth utilizes bootstrapping in conjunction with response error linear modeling to decouple biological variance from inferential variance. Biological variance represents variance in transcript abundance of biological and experimental origin. Inferential variance represents variance caused by sequencing and quantification. In terms of the accuracy of the differential analysis, sleuth is able to achieve higher accuracy than conventional methods.

In this paper, we used sleuth for the differential analysis of RNA-seq.

3 Method

There have been many studies on machine learning for cancer classification. One of the major problems is the large number of features to be used as explanatory variables. If the number of features is too large, the number of combinations that can be expressed will increase dramatically, and the curse of dimension will occur. When the curse of dimension occurs, it will be impossible to obtain proper learning results. To avoid this, it is necessary to select features in advance. Various feature selection methods have been tried in various cancer classification studies, but have not been established yet. In this paper, feature selection was performed by the statistical hypothesis test described later. Details are described in Section 3.1 to 3.6. Figure 1 shows the flow of the entire method.

3.1 Quantification of gene and transcript expression

As the dataset used was base sequences, we performed pseudo-alignment to the reference genome and quantified transcript and gene expression. In this paper, transcript expression level means the expression level that distinguishes the splicing variants, and gene expression level means the expression level that summarizes the splicing variants. GRCh38 cDNA [7] downloaded from Ensembl was used as a reference genome, and quantification was performed by Transcripts Per Kilobase Million (TPM). Pseudo-alignment and quantification were performed by kallisto. The procedure performed is as follows:

1. The index used for realignment was created from the cDNA of GRCh38 by kallisto.

2. The expression levels of transcripts were obtained by kallisto from the created index and liver dataset.

3. The gene expression levels were obtained by tximport [8] from the transcript expression levels obtained in Step 2.

The correspondence table between transcripts and genes used in tximport was created by Ensembl Biomart (Ensembl release 99) [9].

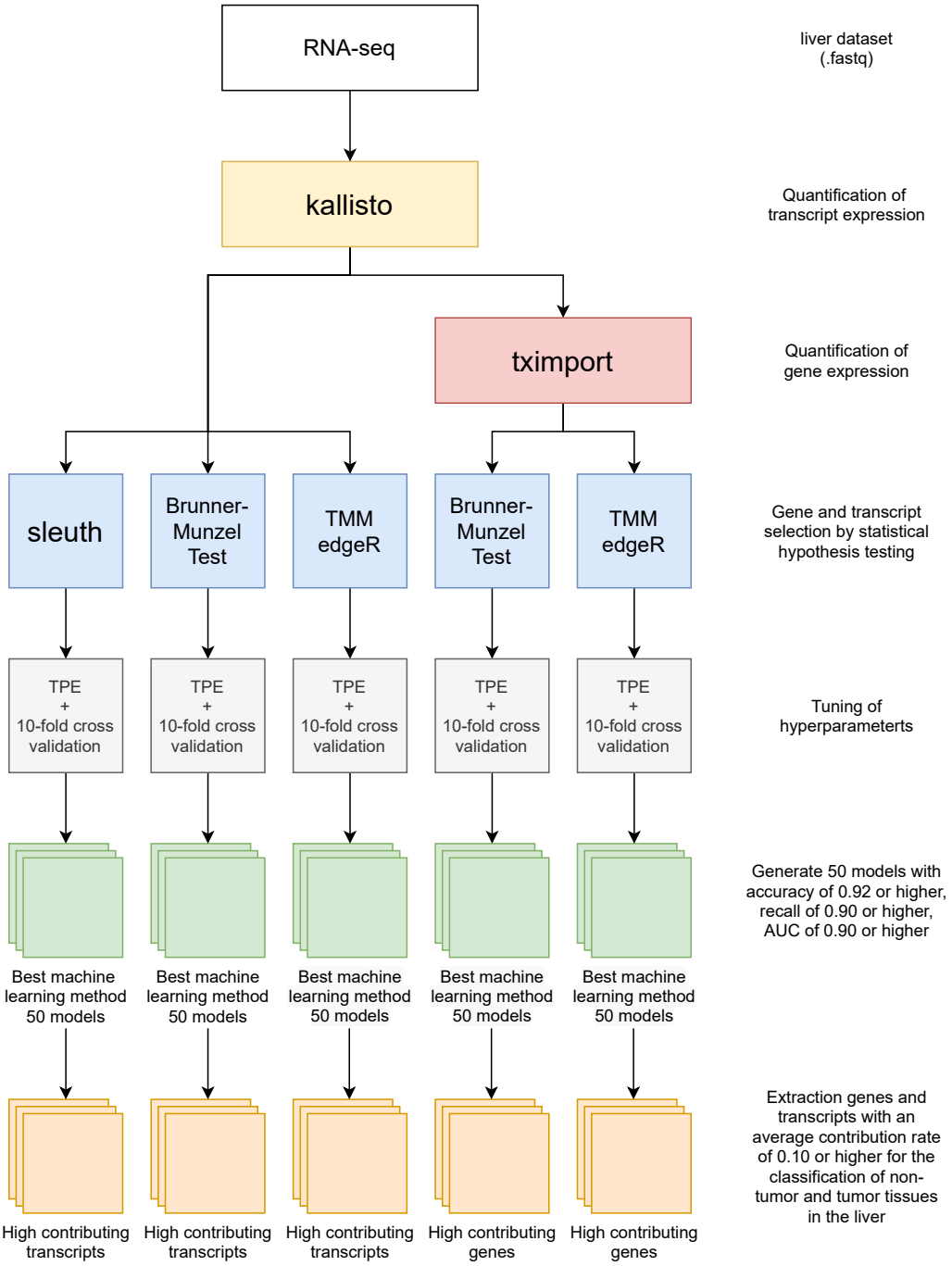


Figure 1: Flow of the entire method

3.2 Selection of genes and transcripts having differences between non-tumor tissues and tumor tissues

Selection was performed by statistical hypothesis tests: the Brunner-Munzel test, the Trimmed mean of M values (TMM) normalized edgeR, and sleuth. Divide the data obtained in the previous procedure into two groups, non-tumor tissues and tumor tissues, and test them by using the expression levels of each gene and transcript.

3.3 Selection by Brunner-Munzel test

It is considered that there are genes and transcripts that show unequal variance between non-tumor tissues and tumor tissues, since diseases are caused by differences in the genes expressed. The Bartlett's test was also performed on the dataset used in this paper, and it was found that about 83% of the genes and about 86% of the transcripts were unequally expressed between non-tumor tissues and tumor tissues. Therefore, it is appropriate to use the Brunner-Munzel test, which can perform the test with high accuracy even if the data are unequal variances. The Brunner-Munzel test does not assume that each distribution of the expression levels of the gene and transcript is the same, but tests the null hypothesis that when the expression levels of the gene or the transcripts are extracted one by one from both groups, the probability of which is the higher is the same. The transcripts and genes of non-tumor tissues and tumor tissues obtained by kallisto and tximport were examined for significant differences in expression level by the Brunner-Munzel (hereinafter BM) test. The significance level was set to 5%.

3.4 Selection by TMM-normalized edgeR

Genes expressed in cells often do not fluctuate in expression, like housekeeping genes. By removing those genes, we can expect to obtain genes and transcripts that are significantly different between non-tumor tissues and tumor tissues. Therefore, it is appropriate to use the TMM-normalized edgeR test. The test assumes that the probability of obtaining each read of gene or transcript between the two groups is equal, we test the null hypothesis that the same number of trials is obtained. As described in Section 3.3, the transcripts and genes of non-tumor tissues and tumor tissues obtained by kallisto and tximport were examined for significant differences in expression level by using the TMM-normalized edgeR (hereinafter edgeR) of TCC-GUI [10] [11]. The significance level was set to 5%.

3.5 Selection by sleuth

The program for analysis of RNA-Seq experiments for which transcript abundances have been quantified with kallisto is sleuth. The program provides tools for exploratory data analysis utilizing Shiny by RStudio, and implements statistical algorithms for differential analysis that leverage the boostrap estimates of kallisto. Since sleuth is optimized for the quantification of transcript expression by kallisto, it is used only for transcript analysis. As described in Section 3.3, the transcripts of non-tumor tissues and tumor tissues obtained by kallisto were examined for significant differences in expression level by using sleuth. The significance level was set to 5%.

3.6 Machine Learning

Machine learning was used to extract genes and transcripts which are considered to be associated with liver cancer.

The extraction method is as follows. First, a random forest, a lightGBM, an SVC, and a linear SVC classifier were created in which the input was the expression level of the gene or transcript extracted in Section 3.3 to 3.5, and the output was non-tumor tissue or tumor tissue. The dataset was divided into 70% of the training data and 30% of the test data. Next, using the training data, hyperparameter tuning of those classifiers was performed using Optuna [12] and 10-fold cross-validation. Table 1 to 4 show the search range of hyperparameters of the classifier of a random forest, a lightGBM, an SVC, and a linear SVC. After that, 50 classifiers with an accuracy of 0.92 or higher, a recall of 0.90 or higher, and an AUC of 0.90 or higher were created by each machine learning method, and compared using test data. Finally, genes and transcripts that contributed significantly to the classification between non-tumor tissues and tumor tissues of liver cancer were extracted using classifiers that are highly evaluated overall.

Table 1: Hyperparameter Search Range of Random Forest

hyperparameter	search range
n_estimators	2,000 - 6,000
max_depth	1 - 1,000

Table 2: Hyperparameter Search Range of LightGBM

hyperparameter	search range
lambda_l1	1e-8 - 10
lambda_l2	1e-8 - 10
num_leaves	2 - 128
feature_fraction	0.4 - 1.0
bagging_fraction	0.4 - 1.0
bagging_freq	1 - 7
min_child_samples	5 - 100

Table 3: Hyperparameter Search Range of SVC

hyperparameter	search range
C	1e-5 - 1e+5
gamma	scale, auto
kernel	poly, rbf, sigmoid

Table 4: Hyperparameter Search Range of Linear SVC

hyperparameter	search range
C	1e-5 - 1e+5
penalty	l1, l2
loss	hinge, squarred_hinge

4 Experiment

4.1 Dataset

In this paper, 227 tumor and 201 non-tumor tissues from the liver cancer dataset JGAD000229 [13] of the NBDC human database were used. These data were RNA extracted from non-tumor tissue (normal tissue) or blood and tumor tissue in resected liver cancer sections. After creating an NGS library (Paired-end) from the RNA, the nucleotide sequence was determined by using Illumina HiSeq, Genome Analyzer.

4.2 Genes and transcripts

There were 190,069 transcripts and 40,480 genes in the data.

As the result of the BM test, transcripts and genes with significant differences were obtained in 103,544 out of 190,069 transcripts and 24,407 out of 40,480 genes.

As the result of the edgeR, transcripts and genes with significant differences were obtained in 80,375 out of 190,069 transcripts and 26,804 out of 40,480 genes.

As the result of the sleuth, transcripts with significant differences were obtained in 6,220 out of 190,069 transcripts.

The transcript overlap rate of the BM and the edgeR was 91.7%, and the gene overlap rate of the BM and edgeR was 87.2%. The transcript overlap rate of the BM and the sleuth was 76.0%, and the transcript overlap rate of the edgeR and the sleuth was 77.1%.

4.3 Comparison of machine learning methods

Table 5 shows the evaluation results of each method.

Table 5: Evaluation Results of Each Machine Learning Method for Each Dataset

dataset	method	accuracy	recall	AUC
BM+gene	random forest	0.951	0.949	0.951
	lightGBM	0.958	0.950	0.958
	SVC	0.948	0.939	0.949
	linear SVC	0.953	0.933	0.954
edgeR+gene	random forest	0.951	0.949	0.951
	lightGBM	0.957	0.951	0.958
	SVC	0.946	0.935	0.947
	linear SVC	0.952	0.930	0.953
BM+transcript	random forest	0.957	0.952	0.957
	lightGBM	0.961	0.955	0.962
	SVC	0.958	0.945	0.959
	linear SVC	0.959	0.945	0.959
edgeR+transcript	random forest	0.957	0.952	0.957
	lightGBM	0.962	0.956	0.962
	SVC	0.958	0.946	0.959
	linear SVC	0.956	0.938	0.957
sleuth+transcript	random forest	0.944	0.946	0.943
	lightGBM	0.952	0.944	0.953
	SVC	0.954	0.947	0.954
	linear SVC	0.958	0.936	0.959

LightGBM was the highest rated for all datasets except the sleuth + transcript dataset. Therefore, LightGBM was adopted in this paper.

4.4 Brunner-Munzel test + LightGBM results

For each value of hyperparameters, lambda_l1 was 0.0 (0.0158, respectively), lambda_l2 was 0.0 (1.176e-08), num_leaves was 5 (31), feature_fraction was 0.4 (0.4), bag-

ging_fraction was 0.604 (0.593), and bagging_freq was 5 (7) for the genes (transcripts).

The average accuracy, the average recall, and the average AUC of all 50 models were 0.958, 0.950, and 0.958 for the genes and 0.961, 0.955, and 0.962 for the transcripts.

Genes and transcripts with an average classification contribution of 0.10 or higher were aggregated from every 50 models of genes and transcripts. These are shown in Table 6 and Table 7, respectively.

Table 6: Genes with an Average Contribution Rate of 0.10 or Higher for Classification of Non-Tumor and Tumor Tissues Obtained by Using BM + LightGBM

id	gene	contribution			
		average	max	min	median
ENSG00000104938	CLEC4M	0.483	1	0	0.471
ENSG00000182566	CLEC4G	0.397	1	0	0.356
ENSG00000145824	CXCL14	0.369	1	0	0.298
ENSG00000160339	FCN2	0.339	1	0	0.188
ENSG00000008300	CELSR3	0.245	1	0	0.145
ENSG00000043355	ZIC2	0.232	1	0	0.151
ENSG00000263761	GDF2	0.197	1	0	0.011
ENSG00000136011	STAB2	0.170	1	0	0.004
ENSG00000089685	BIRC5	0.153	1	0	0.004
ENSG00000164362	TERT	0.151	0.897	0	0.073
ENSG00000165480	SKA3	0.132	1	0	0
ENSG00000158402	CDC25C	0.120	1	0	0.009
ENSG00000077152	UBE2T	0.114	1	0	0.000
ENSG00000135451	TROAP	0.113	1	0	0.003
ENSG00000156509	FBXO43	0.109	1	0	0
ENSG00000147003	CLTRN	0.100	0.614	0	0.026

Table 7: Transcripts with an Average Contribution Rate of 0.10 or Higher for Classification of Non-Tumor and Tumor Tissues Obtained by Using BM + LightGBM

id	transcript	contribution			
		average	max	min	median
ENST00000328853.9	CLEC4G-201	0.416	1	0	0.254
ENST00000310955.11	CDC20-201	0.353	1	0	0.322
ENST00000478222.1	CFP-205	0.349	1	0	0.207
ENST00000512158.5	CXCL14-202	0.231	1	0	0.138
ENST00000581492.3	GDF2-201	0.215	1	0	0.080
ENST00000366999.9	NEK2-202	0.178	1	0	0.000
ENST00000291744.11	FCN2-201	0.145	1	0	0.026
ENST00000376335.8	ZIC2-201	0.134	0.627	0	0.074
ENST00000425389.2	REPIN1-202	0.109	1	0	0.000

4.5 TMM-normalized edgeR + LightGBM results

For each value of hyperparameters, lambda_l1 was 1.18e-08 (3.15e-07, respectively), lambda_l2 was 4.33e-08 (1.05e-08), num_leaves was 31 (11), feature_fraction was 0.7 (0.4), bagging_fraction was 0.801 (0.554), and bagging_freq was 7 (6) for the genes (transcripts).

The average accuracy, the average recall, and the average AUC of all 50 models were 0.957, 0.951, and 0.958 for the genes and 0.962, 0.956, and 0.962 for the transcripts.

Genes and transcripts with an average classification contribution of 0.10 or higher were aggregated from every 50 models of genes and transcripts. These are shown in Table 8 and Table 9, respectively.

4.6 Sleuth + LightGBM results

For each value of hyperparameters, lambda_l1 was 5.16e-07, lambda_l2 was 3.56e-07, num_leaves was 8, feature_fraction was 0.716, bagging_fraction was 0.562, and bagging_freq was 3.

The average accuracy, the average recall, and the average AUC of all 50 models were 0.952, 0.944, and 0.953.

Transcripts with an average classification contribution of 0.10 or higher were aggregated from every 50 models of transcripts. These are shown in Table 10.

Table 8: Genes with an Average Contribution Rate of 0.10 or Higher for Classification of Non-Tumor and Tumor Tissues Obtained by Using EdgeR + LightGBM

id	gene	contribution			
		average	max	min	median
ENSG00000104938	CLEC4M	0.543	1	0	0.587
ENSG00000182566	CLEC4G	0.333	1	0.000	0.168
ENSG00000145824	CXCL14	0.284	1	0	0.248
ENSG00000008300	CELSR3	0.222	1	0	0.125
ENSG00000160339	FCN2	0.195	1	0	0.023
ENSG00000043355	ZIC2	0.193	1	0	0.093
ENSG00000165480	SKA3	0.173	1	0	0.000
ENSG00000164362	TERT	0.142	1	0	0.065
ENSG00000089685	BIRC5	0.122	1	0	0.001
ENSG00000135451	TROAP	0.121	1	0	0.002
ENSG00000263761	GDF2	0.114	0.980	0	0.009
ENSG00000158402	CDC25C	0.111	0.896	0	0.001

Table 9: Transcripts with an Average Contribution Rate of 0.10 or Higher for Classification of Non-Tumor and Tumor Tissues Obtained by Using EdgeR + LightGBM

id	transcript	contribution			
		average	max	min	median
ENST00000310955.11	CDC20-201	0.328	1	0	0.141
ENST00000512158.5	CXCL14-202	0.304	1	0	0.209
ENST00000581492.3	GDF2-201	0.264	1	0	0.158
ENST00000328853.9	CLEC4G-201	0.236	1	0	0.001
ENST00000291744.11	FCN2-201	0.235	1	0	0.071
ENST00000478222.1	CFP-205	0.200	1	0	0.062
ENST00000376335.8	ZIC2-201	0.175	1	0	0.007
ENST00000366999.9	NEK2-202	0.133	0.929	0	0
ENST00000461362.5	CELSR3-202	0.132	1	0	0.006
ENST00000423158.4	NTF3-202	0.121	1	0	0.001
ENST00000646651.1	UBE2T-205	0.120	1	0	0
ENST00000481771.5	MTX1-204	0.112	1	0	0

Table 10: Transcripts with an Average Contribution Rate of 0.10 or Higher for Classification of Non-Tumor and Tumor Tissues Obtained by Using Sleuth + LightGBM

id	transcript	contribution			
		average	max	min	median
ENST00000366860.9	CNIH4-204	0.623	1	0.002	0.653
ENST00000395577.2	CDKN3-203	0.486	1	0	0.454
ENST00000273352.8	ADGRG7-201	0.382	1	0.007	0.284
ENST00000478472.1	GBA-210	0.365	1	0	0.235
ENST00000472642.5	RHEB-204	0.254	1	0	0.170
ENST00000274562.13	LARS1-201	0.253	1	0	0.109
ENST00000310450.8	NAA20-201	0.237	0.751	0	0.198
ENST00000597316.1	TRPM4-208	0.147	0.691	0.010	0.114
ENST00000404674.7	PSMG3-203	0.147	1	0	0.017
ENST00000409299.8	PXMP4-203	0.146	0.666	0	0.092
ENST00000434715.7	DCTN2-201	0.137	1	0	0.001
ENST00000300057.4	MESP1-201	0.108	0.593	0	0.071

4.7 Genes and transcripts associated with liver cancer

Table 11 shows genes and transcripts with an average contribution of 0.10 or higher. Table 12 shows the results of the duplication of genes and transcripts with an average contribution of 0.10 or higher. Transcripts are indicated by their gene name.

Table 11: List of Genes and Transcripts with an Average Contribution Rate of 0.10 or Higher for Classification of Non-Tumor and Tumor Tissues Obtained by Using BM + LightGBM, EdgeR + LightGBM, and Sleuth + LightGBM

BM+gene	edgeR+gene	BM+transcript	edgeR+transcript	sleuth+transcript
CLEC4M	CLEC4M	CLEC4G-201	CDC20-201	CNIH4-204
CLEC4G	CLEC4G	CDC20-201	CXCL14-202	CDKN3-203
CXCL14	CXCL14	CFP-205	GDF2-201	ADGRG7-201
FCN2	CELSR3	CXCL14-202	CLEC4G-201	GBA-210
CELSR3	FCN2	GDF2-201	FCN2-201	RHEB-204
ZIC2	ZIC2	NEK2-202	CFP-205	LARS1-201
GDF2	SKA3	FCN2-201	ZIC2-201	NAA20-201
STAB2	TERT	ZIC2-201	NEK2-202	TRPM4-208
BIRC5	BIRC5	REPIN1-202	CELSR3-202	PSMG3-203
TERT	TROAP	-	NTF3-202	PXMP4-203
SKA3	GDF2	-	UBE2T-205	DCTN2-201
CDC25C	CDC25C	-	MTX1-204	MESP1-201
UBE2T	-	-	-	-
TROAP	-	-	-	-
FBXO43	-	-	-	-
CLTRN	-	-	-	-

Table 12: Duplicate Results of Genes with Transcripts for Genes with an Average Contribution Rate of 0.10 or Higher for Classification of Non-Tumor and Tumor Tissues Obtained by Using BM + LightGBM, EdgeR + LightGBM, and Sleuth + LightGBM

-	BM + gene	edgeR + gene	BM + transcript	edgeR + transcript	sleuth + transcript
BM + gene	-	CLEC4M CLEC4G CXCL14 FCN2 CELSR3 ZIC2 GDF2 BIRC5 TERT SKA3 CDC25C TROAP	CLEC4G CXCL14 FCN2 ZIC2 GDF2	CLEC4G CXCL14 FCN2 CELSR3 ZIC2 GDF2 UBE2T	0
edgeR + gene	CLEC4M CLEC4G CXCL14 FCN2 CELSR3 ZIC2 GDF2 BIRC5 TERT SKA3 CDC25C TROAP	-	CLEC4G CXCL14 FCN2 ZIC2 GDF2	CLEC4G CXCL14 CELSR3 FCN2 ZIC2 GDF2	0
BM + transcript	CLEC4G CXCL14 FCN2 ZIC2 GDF2	CLEC4G CXCL14 FCN2 ZIC2 GDF2	-	CLEC4G CDC20 CFP CXCL14 GDF2 NEK2 FCN2 ZIC2	0
edgeR + transcript	CLEC4G CXCL14 FCN2 CELSR3 ZIC2 GDF2 UBE2T	CLEC4G CXCL14 CELSR3 FCN2 ZIC2 GDF2	CLEC4G CDC20 CFP CXCL14 GDF2 NEK2 FCN2 ZIC2	-	0
sleuth + transcript	0	0	0	0	-

5 Discussion

Firstly, since the models got high scores by evaluating only the test data, which accounts for 30% of the total data, it is considered that overfitting did not occur.

Secondly, CLEC4M [14], BIRC5 [15], TERT [16], SKA3 [17], CDC25C [18], TROAP [19], and FBXO43 [20], which affected the classification only by genes, are reported related to liver cancer. CDC20 [21], NEK2 [22], CDKN3 [23], RHEB [24], and NAA20 [25], which affected the classification only by the transcripts, are reported related to liver cancer. CLEC4G [26], CXCL14 [27], CELSR3 [28], FCN2 [29], ZIC2 [30], GDF2 [31], and UBE2T [32] which affected the classification of both genes and transcripts, are reported related to liver cancer. If a gene is not distinguished from a transcript, 19 out of 34 genes (12 out of 25 transcripts and 14 out of 16 genes) have been reported to be related to liver cancer.

Therefore, it is considered that features that can accurately classify non-tumor tissues and tumor tissues in the liver could be extracted.

The two genes, STAB2 and CLTRN and the thirteen transcripts CFP-205, REPIN1-202, NTF3-202, MTX1-204, CNIH4-204, ADGRG7-201, GBA-210, LARS1-201, TRPM4-208, PSMG3-203, PXMP4-203, DCTN2-201, and MESP1-201, which affected the classification

of non-tumor tissues and tumor tissues in our results, were not found in papers that directly report they are related to liver cancer. In particular, the thirteen transcripts have potential for new discoveries because they have been difficult to analyze in terms of gene analysis technology. In addition, all other genes have been reported to be related to liver cancer. These results suggest that this group of genes may also be associated with liver cancer.

Figure 2 shows the average expression levels of the two genes in non-tumor tissues and tumor tissues. Figure 3 shows the distribution of the expression levels of the two genes in non-tumor tissues and tumor tissues. Similarly, Figure 4 and 5 show the thirteen transcripts. They have already been shown by BM, edgeR, and sleuth that there are significant differences in the expression levels of those genes between non-tumor tissues and tumor tissues. Therefore, it can be seen from these graphs that the expression levels of STAB2, CLTRN, CFP-205, NTF3-202, and ADGRG7-201 in tumor tissues are significantly lower than non-tumor tissues, and the expression levels of REPIN1-202, MTX1-204, CNIH4-204, GBA-210, LARS1-201, TRPM4-208, PSMG3-203, PXMP4-203, DCTN2-201, and MESP1-201 in tumor tissues are significantly higher than them. It is considered that the relationship with liver cancer can be investigated by manipulating the expression levels of those genes.

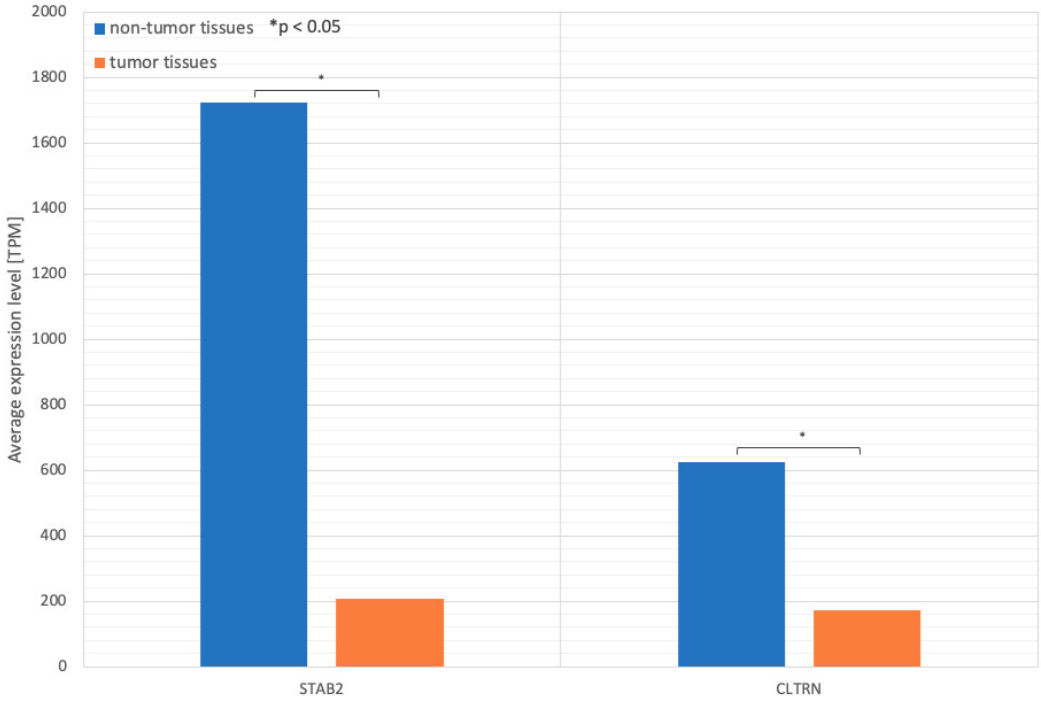
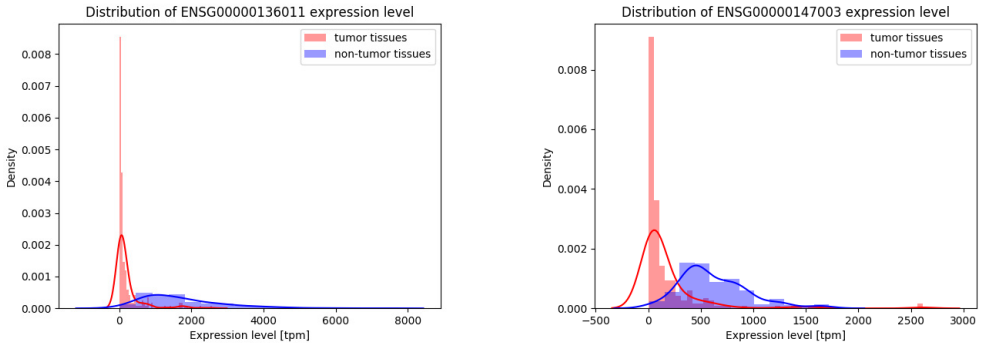


Figure 2: Average expression levels of the two genes which are considered to be associated with liver cancer in non-tumor and tumor tissues ($P < 0.05$, blue: non-tumor tissues, red: tumor tissues)



(a) STAB2 (b) CLTRN

Figure 3: Kernel density estimation results of expression levels of the two genes which are considered to be associated with liver cancer (vertical axis: density, horizontal axis: expression level, blue distribution: non-tumor tissues, red distribution: tumor tissues)

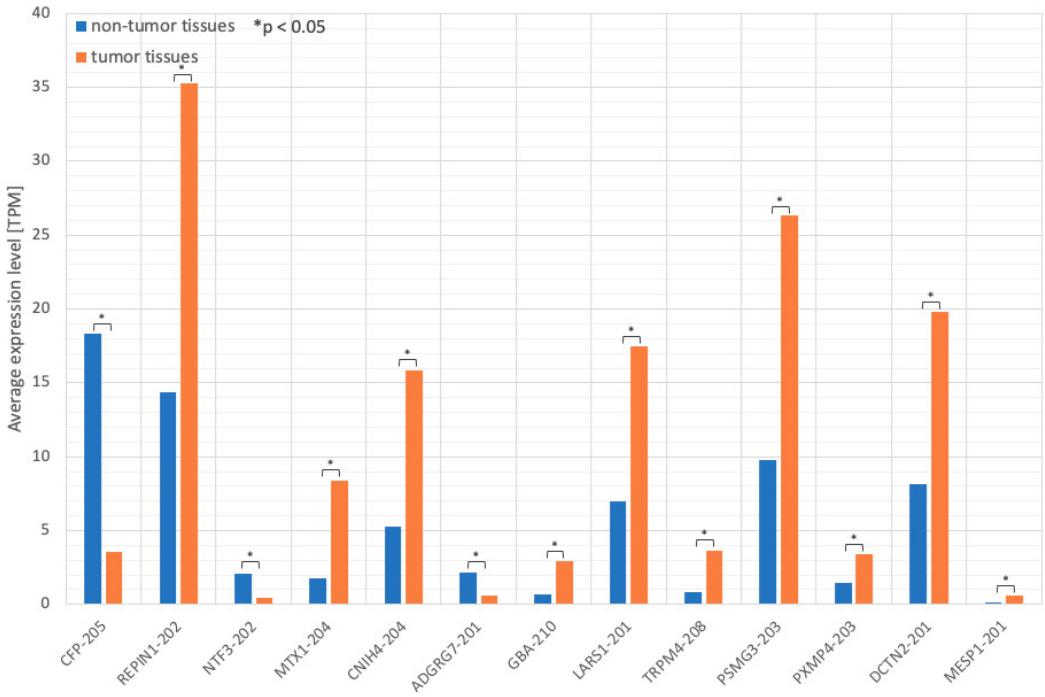
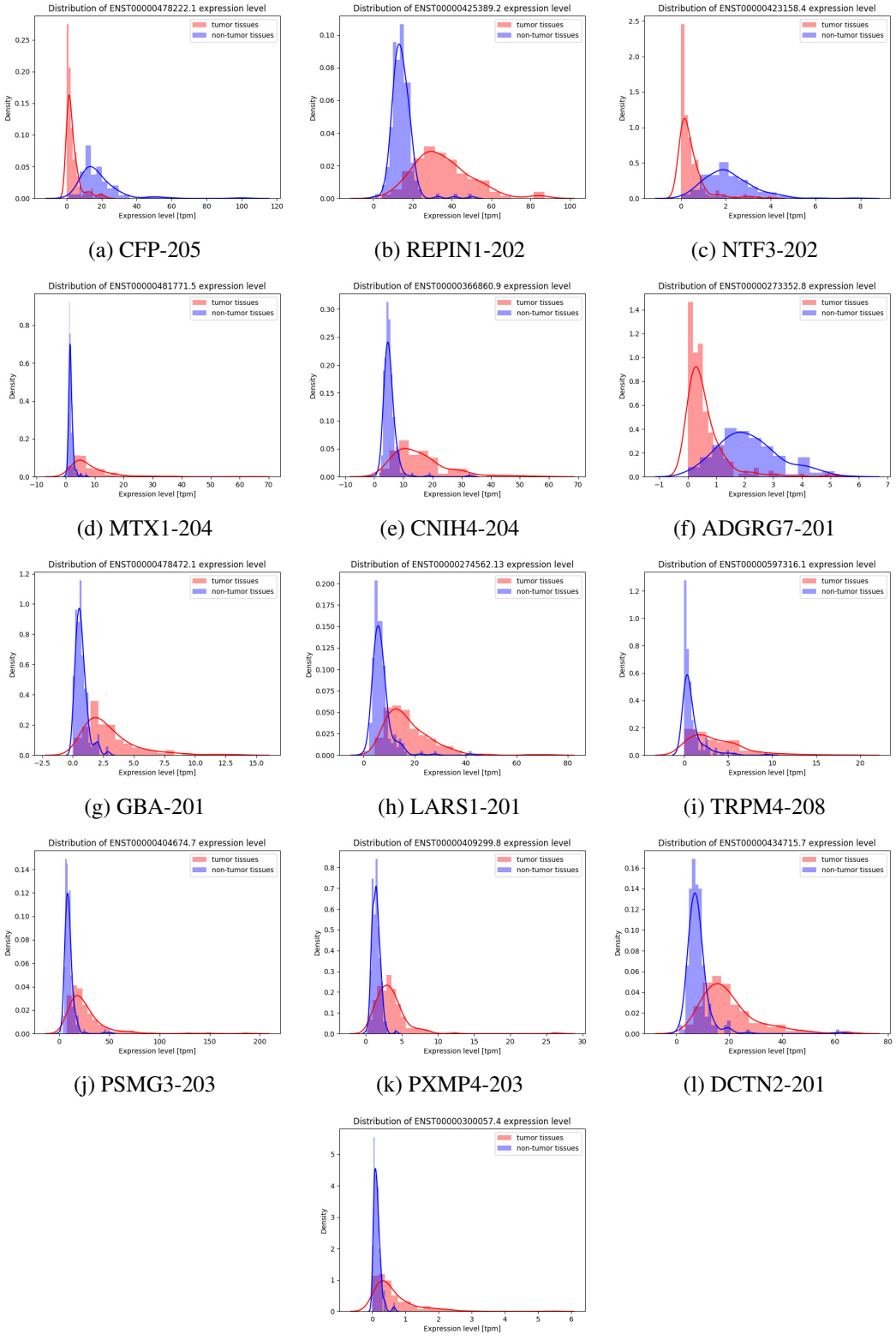


Figure 4: Average expression levels of the thirteen transcripts which are considered to be associated with liver cancer in non-tumor and tumor tissues ($P < 0.05$, blue: non-tumor tissues, red: tumor tissues)



(a) CFP-205 (b) REPIN1-202 (c) NTF3-202

(d) MTX1-204 (e) CNIH4-204 (f) ADGRG7-201

(g) GBA-201 (h) LARS1-201 (i) TRPM4-208

(j) PSMG3-203 (k) PXMP4-203 (l) DCTN2-201

(m) MESP1-201

Figure 5: Kernel density estimation results of expression levels of the thirteen transcripts which are considered to be associated with liver cancer (vertical axis: density, horizontal axis: expression level, blue distribution: non-tumor tissues, red distribution: tumor tissues)

6 Conclusion

With the recent rapid progress of large-scale analysis technology and the release of data related to gene expression, a comprehensive analysis of gene expression has been actively conducted. However, the relationship between gene expression and disease has not been fully elucidated due to the difficulty in finding appropriate data among the large amount of data. As a result, it is said that the discovery rate of gene mutations suitable for the selection of therapeutic methods is about 50%.

In this paper, we aimed to extract genes related to liver cancer. We extracted genes and transcripts with significant differences in expression levels between non-tumor tissues and tumor tissues of the liver dataset using the BM test, the edgeR test, and the sleuth. Then, they were used as input to the LightGBM, and the genes and transcripts with a high contribution of classification were obtained. If a gene is not distinguished from a transcript, 19 of 34 important genes are related to liver cancer in previous studies. The remaining fifteen genes, STAB2, CLTRN, CFP(CFP-205), REPIN1(REPIN1-202), NTF3(NTF3-202), MTX1(MTX1-204), CNIH4(CNIH4-204), ADGRG7(ADGRG7-201), GBA(GBA-210), LARS1(LARS1-201), TRPM4(TRPM4-208), PSMG3(PSMG3-203), PXMP4(PXMP4-203), DCTN2(DCTN2-201), and MESP1(MESP1-201), were judged to be of high importance in this paper, but these genes have not been reported related to liver cancer in previous studies. However, these genes are considered to be worthy of further investigation, as they include those that contribute more than other genes reported to be related to liver cancer.

Future work is to investigate the relationship between liver cancer and the fifteen genes obtained in this study in patients. And, analysis of ncRNA as well as mRNA is included in future work because ncRNA has various effects on the gene expression process, but its function is unclear.

Acknowledgments

We would like to thank RIKEN Center for Integrative Medical Sciences (IMS) (Representative Hidetoshi Nakagawa) for providing valuable datasets through the website (http://humandbs.biosciencedbc.jp/) of the National Bioscience Database Center (NBDC) of the Japan Science Technology Agency (JST).

References

[1] K. Sekine, T. Hochin, and H. Nomiya, "Extraction of Genes Associated with Liver Cancer Using Machine Learning," 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 7-12, 2020.

[2] H. Ide, K. Kanamori, and H. Ohwada, "Extraction of ncRNA associated with acute lung injury using machine learning," Proceedings of the Annual Conference of JSAI, vol. JSAI2016, pg. 4J44, 2016. (in Japanese).

[3] R. Díaz-Uriarte, Ramón and S. Alvarez de Andrés, "Gene selection and classi- fication of microarray data using random forest," BMC Bioinformatics, vol. 7, no. 1, pg. 3, 2006.

[4] O. Okun and P. Helen, "Random forest for gene expression based cancer classification: Overlooked issues," In Iberian Conference on Pattern Recognition and Image Analysis, pp. 483–490, 2007.

[5] N.L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic rna-seq quantification," Nature Biotechnology, vol. 34, no. 5, pp. 525–527, 2016.

[6] H. Pimentel, N.L. Bray, S. Punte, "Differential analysis of RNA-seq incorporating quantification uncertainty," Nature Methods, vol. 14, no. 7, pp. 687–690, 2017.

[7] Ensembl, Homosapiens.grch38.cdna.all.fa.gz, 2019. Accessed on March 20, 2021. [Online]. Available: http://ftp.ensembl.org/pub/release-99/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz

[8] C. Soneson, M. I. Love, and M. D. Robinson, "Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences," F1000Research, vol. 4, pg. 1521, 2016.

[9] A.D. Yates, P. Achuthan, W. Akanni, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai,K. Billis, S. Boddu, J. C. Marugán, C. Cummins, C. Davidson,K. Dodiya, R. Fatima, A. Gall, C. G. Giron, L. Gil, T. Grego,L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek,T. Juettemann, M. Kay, I. Lavidas, T. Le, D. Lemos, J. G. Martinez,T. Maurel, M. McDowall, A. McMahon, S. Mohanan, B. Moore,M. Nuhn, D. N. Oheh, A. Parker, A. Parton, M. Patricio, M. P. Sakthivel, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg,D. Sheppard, M. Sycheva, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, A. Vullo, B. Walts, A. Winterbottom, A. Zadissa, M. Chakiachvili, B. Flint, A. Frankish, S. E. Hunt, G. IIsley, M. Kostadima, N. Langridge, J. E. Loveland, F. J. Martin, J. Morales,J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, S. J. Trevanion,F. Cunningham, K. L. Howe, D. R. Zerbino, and P. Flicek, "Ensembl2020,"Nucleic Acids Research, vol. 48, no. D1, pp. D682–D688, 112019. [Online]. Available: https://doi.org/10.1093/nar/gkz966

[10] W. Su, J. Sun, K. Shimizu, and K. Kadota, "Tcc-gui: a shiny-based application for differential expression analysis of rna-seq count data," BMC Research Notes, vol. 12, no. 1, pg. 133, 2019.

[11] J. Sun, T. Nishiyama, K. Shimizu, and K. Kadota, "Tcc: an r package for comparing tag count data with robust normalization strategies," BMC Bioinformatics, vol. 14, no. 1, pg. 219, 2013.

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna:Anext- generation hyperparameter optimization framework," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631, 2019.

[13] H. Nakagawa, JGAD000229, NBDC Human Database, 2018. Accessed on: March 20, 2021. [Online]. Available: https://humandbs.biosciencedbc.jp/hum0158-v2

[14] L. Luo, L. Chen, K. Ke, B. Zhao, L. Wang, C. Zhang, F. Wang, N. Liao, X. Zheng, and Y. Wang, "High expression levels of clec4m indicate poor prognosis in patients with hepatocellular carcinoma," Oncology Letters, vol. 19, no. 3, pp. 1711–1720, 2020.

[15] A. C. Fields, G. Cotsonis, D. Sexton, R. Santoianni, and C. Cohen, "Survivin expression in hepatocellular carcinoma: correlation with proliferation, prognostic parameters, and outcome," Modern Pathology, vol. 17, no. 11, pp. 1378–1385, 2004.

[16] Y.-L. Chen, Y.-M. Jeng, C.-N. Chang, H.-J. Lee, H.-C. Hsu, P.-L. Lai, and R.-H. Yuan, "TERT promoter mutation in resectable hepatocellular carcinomas: A strong association with hepatitis C infection and absence of hepatitis B infection," International Journal of Surgery, vol. 12, no. 7, pp. 659–665, 2014.

[17] Y. Hou, Z. Wang, S. Huang, C. Sun, J. Zhao, J. Shi, Z. Li, Z. Wang, X. He, N. L. Tam, and L. Wu, "Ska3 promotes tumor growth by regulating cdk2/p53 phosphorylation in hepatocellular carcinoma," Cell death & disease, vol. 10, no. 12, pg. 929, 2019.

[18] X. Jin, H. Nagano, K. Sakon, H. Yamamoto, H. Eguchi, A. Kanmoto, K. Kondo, I. Arai, S. Morimoto, K. Dono, S. Nakamori, K. Umeshita, and M. Kadota, "Clinipathological study on cdc25 expression in hepa- tocellular carcinoma cases," in Liver, 2000, (in Japanese).

[19] B. Xu, W. LV, X. Li, L. Zhang, and J. Lin, "Prognostic genes of hepatocellular carcinoma based on gene coexpression network analysis," Journal of Cellular Biochemistry, vol. 120, 2019.

[20] H. Hu, L. Xu, Y. Chen, S.-J. Luo, Y.-z. Wu, S.-H. Xu, M.-T. Liu, F. Lin, Y. Mei, Q. Yang, Y.-y. Qiang, Y.-w. Lin, Y.-j. Deng, T. Lin, Y.-q. Sha, B.-J. Huang, and S.-J. Zhang, "The upregulation of trophinin-associated protein (troap) predicts a poor prognosis in hepatocellular carcinoma," J Cancer, vol. 10, pp. 957–967, 2019.

[21] J. Li, J.-Z. Gao, J.-L. Du, Z.-X. Huang, and L.-X. Wei, "Increased cdc20 expression is associated with development and progression of hepatocellular carcinoma," International journal of oncology, vol. 45, no. 4, pp. 1547–1555, 2014.

[22] Y. Zhang, W. Wang, Y. Wang, X. Huang, Z. Zhang, B. Chen, W. Xie, S. Li, S. Shen, and B. Peng, "NEK2 promotes hepatocellular carcinoma migration and invasion through modulation of the epithelial-mesenchymal transition," Oncology reports, vol.39, no. 3, pp. 1023–1033, 2018.

[23] W. Dai, H. Miao, S. Fang, T. Fang, N. Chen, and M. Li, "CDKN3 expression is negatively associated with pathological tumor stage and CDKN3 inhibition promotes cell survival in hepatocellular carcinoma," Molecular medicine reports, vol. 14, no. 2, pp. 1509–1514, 2016.

[24] F. Liu, Z. Pan, J. Zhang, J. Ni, C. Wang, Z. Wang, F. Gu, W. Dong, W. Zhou, and H. Liu, "Overexpression of RHEB is associated with metastasis and poor prognosis in hepatocellular carcinoma," Oncology letters, vol. 15, no. 3, pp. 3838–3845, 2018.

[25] T.-Y. Jung, J.-E. Ryu, M.-M. Jang, S.-Y. Lee, G.-R. Jin, C.-W. Kim, C.-Y. Lee, H. Kim, E. Kim, S. Park, S. Lee, C. Lee, W. Kim, T. Kim, S.-Y. Lee, B.-G. Ju, and H.-S. Kim, "Naa20, the catalytic subunit of NatB complex, contributes to hepatocellular carcinoma by regulating the LKB1–AMPK–mTOR axis," Experimental & Molecular Medicine, vol. 52, no. 11, pp. 1831–1844, 2020.

[26] D. W.-H. Ho, A. K.-L. Kai, and I. O.-L. Ng, "Tcga whole-transcriptome sequencing data reveals significantly dysregulated genes and signaling pathways in hepatocellular carcinoma," Frontiers of Medicine, vol. 9, no. 3, pp. 322–330, 2015.

[27] Y. Lin, B. Chen, X. Yu, H. Yi, J. Niu, and S. Li,"Suppressed expression of cxcl14 in hepatocellular carcinoma tissues and its reduction in the advanced stage of chronic hbv infection," Cancer Manag Res., vol. 11, pp. 10435–10443, 2019.

[28] X. Gu, H. Li, L. Sha, Y. Mao, C. Shi, and W. Zhao, "CELSR3 mRNA expression is increased in hepatocellular carcinoma and indicates poor prognosis," PeerJ, vol.7, pg.e7816, 2019.

[29] G. Yang, Y. Liang, T. Zheng, R. Song, J. Wang, H. Shi, B. Sun, C. Xie, Y. Li, J. Han, S. Pan, Y. Lan, X. Liu, M. Zhu, Y. Wang, and L. Liu, "Fcn2 inhibits epithelial-mesenchymal transition-induced metastasis of hepatocellular carcinoma via tgf-β/smad signaling," Cancer Letters, vol. 378, no. 2, pp. 80–86, 2016.

[30] S.-X. Lu, C. Z. Zhang, S.-P. Chen, C.-H. Wang, L. Liu, J. Fu, L. Zhang, H. Wang, D. Xie, and J.-P. Yun, "Zic2 promotes tumor growth and metastasis via PAK4 in hepatocellular carcinoma," Cancer letters, vol. 402, pp. 71–80, 2017.

[31] B. Herrera, M. García-Álvaro, S. Cruz, P. Walsh, M. Fernández, C. Roncero, I. Fabregat, A. Sánchez, and G. J. Inman, "Bmp9 is a proliferative and survival factor for human hepatocellular carcinoma cells," PLOS ONE, vol. 8, no. 7, pp. 1–12, 2013.

[32] N. P. Y. Ho, C. O. N. Leung, T. L. Wong, E. Y. T. Lau, M. M. L. Lei, E. H. K. Mok, H. W. Leung, M. Tong, I. O. L. Ng, J. P. Yun, S. Ma, and T. K. W. Lee, "The interplay of UBE2T and Mule in regulating Wnt/β-catenin activation to promote hepatocellular carcinoma progression," Cell Death & Disease, vol. 2021, no. 2, pg. 148, 2021.