Classification of early and late stage Liver Hepatocellular Carcinoma patients from their genomics and epigenomics profiles.

Harpreet Kaur[1], Sherry Bhalla[2,3], GPS Raghava[2*]

1. Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39A, Chandigarh-160036, India
2. Center for Computational Biology, Indraprastha Institute of Information Technology, New Delhi-110020, India
3. Centre for Systems Biology and Bioinformatics, Panjab University, Sector 14, Chandigarh-160014, India

Emails of Authors:

Harpreet Kaur: harpreet5aug@imtech.res.in

Sherry Bhalla: s.sherry308@gmail.com

GPS Raghava: raghava@iiitd.ac.in

[*] **Corresponding author**

Professor, Center for Computational Biology, Indraprastha Institute of Information Technology,

Okhla Industrial Estate, Phase III, New Delhi 110020,

India. Tel.: +91 011 26907444

E-mail address: raghava@iiitd.ac.in

Abstract

Background: Liver Hepatocellular Carcinoma (LIHC) is the second major cancer worldwide, responsible for millions of premature deaths every year. Prediction of clinical staging is vital to implement optimal therapeutic strategy and prognostic prediction in cancer patients. However, to date, no method has been developed for predicting stage of LIHC from genomic profile of samples.

Results: In current study, in silico models have been developed for classifying LIHC patients in early and late stage using RNA expression and DNA methylation data. The Cancer Genome Atlas (TCGA) dataset contains 173 early and 177 late stage samples of LIHC, was extensively analysed to identify differentially expressed RNA transcripts and methylated CpG sites that can discriminate early and late stages of LIHC samples with high precision. Naive Bayes model developed using 51 features that combine 21 CpG methylation sites and 30 RNA transcripts achieved maximum MCC 0.58 with accuracy 78.87% on validation dataset. Further, we also analysed genomics and epigenomics profiles of normal and LIHC samples and developed model to classify LIHC samples with AUROC 0.99. In addition, multiclass models developed for classifying samples in normal, early and late stage of cancer and achieved accuracy of 76.54% and AUROC of 0.86.

Conclusion: Our study reveals stage prediction of LIHC samples with high accuracy based on genomics and epigenomics profiling is a challenging task in comparison to classification of LIHC and normal samples. Comprehensive analysis, differentially expressed RNA transcripts, methylated CpG sites in LIHC samples and prediction models are available from CancerLSP (http://webs.iiitd.edu.in/raghava/cancerlsp/).

Keywords: Hepatocellular Carcinoma, Early stage, Late stage, CpG site, Methylation, RNA expression, Prediction, Machine learning, Biomarkers

Background

Hepatocellular Carcinoma (HCC) is the fifth most common cancer and considered as the second major cause of cancer-related mortality with nearly 7,88,000 deaths occurring worldwide due to liver cancer in the year 2015. Further in United States, approximately 30,200 deaths and 42,220 new cases are estimated in 2018. It is nearly four times more frequent in males than females [1]. Also, the higher number of HCC cases is reported in Africa and Asia in comparison to Europe, these observations indicate that HCC is defined by various factors. The pathogenesis of HCC is contributed by some risk factors like viral hepatitis infection (hepatitis B or C) or cirrhosis, smoking, alcohol and lifestyle *etc* [2]. Despite improved screening and discoveries, HCC exhibits rapid clinical course with elevated mortality rate. Patients with HCC are usually identified at advanced stages due to lack of pathognomonic symptoms, which consequently limits the potential treatment options and early death [3]. Furthermore, there is high recurrence rate of 70% even after curative resection treatment [4]. Therefore, absolute cure of this disease is quite challenging, indicating an urgent need for identification of sensitive diagnostic and prognostic markers for HCC [5, 6].

Traditionally in most of the developing countries, alpha-fetoprotein (AFP) is extensively employed as HCC biomarker. Its level becomes detectable in HCC once tumor is in advanced stage [7, 8]. Besides, AFP-L3, a glycoform of AFP (AFP reacts with Lens culinaris agglutinin) has also been employed as HCC biomarker due to its higher sensitivity and specificity than alone AFP. The lack of reliability, insufficient sensitivity and specificity are the major associated limitations with these markers [9]. Also, Des-gamma-carboxyprothrombin (DCP) is another vital HCC biomarker. Its levels have shown to be upregulated in advanced stages [10-12]. In recent times, next-

generation sequencing technology and bioinformatics analysis emergence has facilitated the identification of tumor and progression biomarkers of HCC [13]. Anomalous expression of cancer-associated genes is one of the major causes of tumorigenesis and plays vital role in hepatocarcinogenesis [14]. Evidently various reports have shown the elevated expression of *USP22, CBX6, NRAGE, ACTL6* and *CHMP4B* genes correlates with larger tumor size, advanced tumor stages, poor prognosis and short survival time of patients in HCC [15-19]. Moreover, the downregulation of *BTG1, FOXF2* and *CYP3A5* genes, have been observed to link with poorly differentiated and aggressive tumors, shorter disease survival rates and shorter recurrence times in HCC [20-22]. Beside this, recently, long non-coding RNAs have also been found to be aberrantly expressed and implicated in HCC pathogenesis [23-25]. *ZEB1- AS1* and *ANRIL* expression observed to be upregulated and their association with higher histological grade and stage in HCC [26, 27]. Hence, the potential reversibility of epigenetic abnormalities and restoration of the expression of tumor suppressor genes or other genes by specific inhibitors offer a rational therapeutic approach for HCC. Though in recent past, numerous tumor drivers like *TERT, CTNNB1 etc* have been identified in HCC, however, most of them have not been translated into efficient modalities [28].

In addition, various studies have shown that the alterations in epigenetics and miRNA pattern lead to progression from precancerous lesions to HCC [29, 30]. The epigenetic alterations including DNA hypermethylation or hypomethylation, dysregulation of histone modification patterns, chromatin remodeling *etc* are associated with HCC [31]. Tumor suppressor genes such as *SOCS1, hMLH1, GSTP1, MGMT, CDH1* and *TIMP3 etc* are inactivated by the virtue of promoter hypermethylation [32].

Detection of cancer at an early stage is vital to reduce the mortality rate by providing

appropriate treatment based on cancer stage. In previous studies, researchers focused mainly on identification of differentially expressed RNA transcripts or genes in LIHC. Best of our knowledge, no method has been developed for predicting stage of LIHC from genomic profile of samples. In this study, a systematic attempt has been made to develop model for discriminating early and late stage of LIHC samples. Firstly, we identified CpG sites that are differentially methylated in early and late stage of LIHC. Secondly, we identified aberrantly expressed RNA transcripts that can differentiate early stage from late stage of LIHC. Thirdly, models were developed based on different machine learning techniques for predicting stage of LIHC samples using the above derived genomic and epigenomic features. Using diverse feature spaces, we were able to establish models discriminating early and late stage of LIHC cancer. In addition, we made an attempt to develop models for discriminating normal and LIHC tissue samples. Our models successfully predict LIHC samples with high accuracy. It indicates that it is easy to predict LIHC samples but it is challenging to classify them in early and late stage. We also attempted to develop multiclass prediction models to classify samples in three classes; i) normal or control samples, ii) LIHC early stage and iii) LIHC late stage tissue samples.

Data Description

Main Dataset of Early & Late Stage Samples

We extracted the expression and methylation profiles of Liver Hepatocellular Carcinoma (LIHC) samples from GDC Data Portal (https://portal.gdc.cancer.gov/). In addition, manifest, metadata, clinical data, biospecimen files were downloaded to obtain clinical information using Biospecimen Core Resource (BCR) IDs of patients. Finally, we obtained 173 stage-I, 87 stage-II, 85 stage-III and 5 stage- IV stage

samples. Clinical characteristics of these patients displayed in figure S1 (Additional file 3, Figure S1). As the number of stage-IV samples in the dataset is small, we have considered stage-II, stage-III and stage-IV samples as late stage samples, while stage-I samples as early stage samples as stage I samples are of localized cancer which shows no sign of metastasis. We also downloaded the methylation profiles acquired using the Illumina Human- Methylation450K DNA Analysis BeadChip assay, based on genotyping of bisulfite-converted genomic DNA at individual CpG-sites. This provides Beta values, a quantitative measure of DNA methylation [33]. In addition, for each subject, RNA expression in terms of FPKM values for 60,483 RNA transcripts were reported. In this study, we have used FPKM values of RNA transcripts as quantification values.

Analyses

<u>Models for classification of early stage and late stage of LIHC samples</u>

Our primary objective is to identify potential markers *i.e.* CpG sites and RNA transcripts that can classify early stage tissue samples from late stage tissue samples. Subsequently in silico predictive models were developed based on these signature markers using various machine learning algorithms. Potential markers and prediction models based on them are explained in the following sections. The overall work flow represented in Figure 1.

Single feature based stage classification model using CpG sites

Here, we used single feature based classification technique to classify early and late stage samples. In this study, we first identified 447 CpG sites (from 3,74,292 sites) that are significantly differentially methylated in early and late stage of sample using t-test (FDR<0.05). These differentially methylated sites are further segregated in two

classes; i) 199 hypermethylated CpG sites (high level of methylation in early stage) and ii) 248 hypomethylated CpG site (low level methylation in early stage). Further, we used each differentially methylated CpG sites to develop a simple threshold-based model for classifying early and late stage samples. In the threshold-based model, a sample is classified in the early stage if the methylation level of CpG site (in case site is hypermethylated) is higher than a threshold value otherwise in late stage. In these models, the threshold is varied incrementally from minimum to maximum beta-value. In the final step, that threshold is selected which leads to maximum AUROC between early and late samples. Subsequently, all the 447 CpG sites are ranked according to AUROC to assess the ability of a CpG site to categorize early and late stage tissue samples (Additional file 1, Table_S1). As shown in Table_S1 (Additional file 1), there are 170 differentially methylated CpG sites (LS-CPG-AUROC) that can differentiate two types of samples with high precision (ROC $\geq$ 0.6). LS-CPG-AUROC further includes 105 hypermethylated and 65 hypomethylated CpG sites. The differential methylation pattern can be observed in heatmap (Additional file 4, Figure_S2). Hypermethylated CpG sites in early stage such as cg12595697, cg11232136, cg01055099 can distinguish early and late stage samples with AUROC 0.66, 0.65 and 0.65 at threshold 0.94, 0.90 and 0.84 respectively. It indicates that if the beta value of these CpG sites is greater than the mentioned threshold, then sample belongs to early stage, otherwise to the late stage. The hypomethylated sites in the early stage such as cg07132710 and cg22491320 can classify early stage samples from late stage with AUROC at 0.65 at threshold 0.63 and 0.43. It signifies that if the beta value of these CpG sites is lower than the mentioned threshold, then the sample belongs to early stage, otherwise late stage. Methylation pattern of the top 20 CpG

from 447 CpG sites is shown **Figure 2A** and their chromosome location and associated genes are represented in **Figure 2B**.

LS-CPG-AUROC signature is significantly enriched (adjusted *p*-value <0.05) in various immune system associated pathways and cancer associated cell signaling pathways present in BioCarta. For instance, 4 out of 55 genes of T Cell Receptor Signaling Pathway and 5 out of 33 genes of Integrin Signaling Pathway_Homosapiens, are enriched in LS-CPG-AUROC. Furthermore, *ITGB1*, *SHC1* and *PTK2* are associated with PTEN dependent cell cycle arrest and apoptosis, VEGF, Hypoxia, and AngiogenesisPathway. This signature is also enriched in actin filament binding (GO:0051015).

Multiple features based stage classification model using CpG sites

One of the challenges in classification models is to improve the performance of models. Though, we got a single CpG site like cg12595697, which can classify samples with AUROC 0.66. To utilize information from multiple features, we developed models using machine learning techniques. Firstly, models developed based on various machine learning techniques using above LS-CPG-AUROC.As shown in Table_S2 (Additional file 1), most of the models attain reasonable high performance on training dataset, maximum AUROC 0.79 for SVM, but poor AUROC *i.e.* 0.66 on external validation dataset.

Therefore, we selected 21 features using WEKA from 447 CpG differentially methylated sites with FDR (False discovery Rate) < 0.05 for early stage v/s late stage. These 21 selected features (CpG sites) named as LS-CPG-WEKA were used to develop models for classification using different machine learning techniques (Table 1). The SVM based model achieved the highest accuracy of 75.99% with AUROC 0.81andaccuracy of 73.24% with AUROC 0.78 on the training dataset and the

validation dataset respectively as shown in Table 1. Models based on Random forest, SMO and Naïve Bayes also achieved performance in the similar range. Among 21 CpG sites (selected features), 9 and 12 were hypomethylated and hypermethylated CpG sites respectively. Heatmap displaying the methylation pattern (Figure 3A) and Circos plot representing the chromosome location, genes associated with them and feature type shown in Figure 3B.

Table 1: The performance of different machine learning techniques based stage classification models developed using 21 methylation CpG sites selected by WEKA (LS-CPG-WEKA).

Machine Learning Techniques	**Dataset**	**Performance Measures**				
		Sensitivity	**Specificity**	**Accuracy (%)**	**MCC**	**AUROC**
SVM	Training	80.43	71.63	75.99	0.52	0.81
	Validation	71.43	75	73.24	0.46	0.78
Random Forest	Training	77.54	73.05	75.27	0.51	0.8
	Validation	71.43	83.33	77.46	0.55	0.79
Naïve Bayes	Training	87.68	62.41	74.91	0.52	0.82
	Validation	80	63.89	71.83	0.44	0.82
SMO	Training	82.61	69.5	75.99	0.53	0.76
	Validation	74.29	75	74.65	0.49	0.75
J48	Training	62.32	75.18	68.82	0.38	0.67
	Validation	45.71	61.11	53.52	0.07	0.57

Interestingly 9 hypermethylated CpG sites of LS-CPG-WEKA signature in early stage are associated with 12 genes (*EEF2, NOLC1, PPP1R15A, DNAJC14, DNAJB6, XYLT1* etc.) that are significantly enriched (adjusted p-value <0.05) in various GO biological processes such as two genes are involved in positive regulation of translation (GO:0045727) and rescue of stalled ribosome (GO:0072344). While, hypomethylated CpG sites are associated with 14 genes (*SMAD7, PCNX, RGL2, POLR1D, ZNF554, ZNF200 etc.*) and these genes are significantly enriched in different BioCarta pathways. For instance, *SMAD7*, a component of catenin complex (GO:0016342), is involved in TGF beta signaling pathway_Homosapiens_h_tgfbPathway and has been shown to inhibit TGF-beta (Transforming growth factor) and active in signaling by associating with their receptors. *ZNF554, PLOR1D* and *PSPC1* are involved in regulation of transcriptional activity.

Stage classification models using the expression of RNA transcripts

In above section, we studied methylation level of CpG sites in early and late stage samples as well as the classification models developed using methylation of CpG sites. Similarly, in this section, we analyzed expression of RNA transcripts in both types of samples as well as developed stage classification models. Firstly, RNA transcripts number was reduced to 103 (LS-RNA-AUROC) after applying stringent criteria (FDR adjusted p-value < 0.01) from 60,483. In LS-RNA-AUROC signature, 39 transcripts were over-expressed and 64 transcripts were under expressed in early stage LIHC samples. Secondly, we developed single feature based models using each transcript, similar to single features based models for CpG sites. Finally, we ranked RNA transcripts based on AUROC to discriminate early stage from late stage tissue samples. The 99 RNA transcript having AUROC greater or equal to 0.6 is shown in

Table_S3 (Additional file 1, Table S3) with their expression patterns illustrated in Figure_S3 (Additional file 5). Furthermore, *NCAPH*, *CYP4A22*, *HSD17B6*, *GLYATL1* and *FTCD* are top 5 RNA transcripts which distinguish early stage and late stage LIHC samples with AUROC >= 0.65. Among them, *NCAPH* is top performer with AUROC 0.66 at threshold 1.45. It means that if the log2-FPKM value of *NCAPH* is lower than 1.45, then the sample belongs to early stage otherwise to late stage as it is under-expressed in early stage. The expression pattern of top 20 genes in early stage samples in comparison to late stage tissue samples displayed in Figure 4A.

Gene enrichment analysis of LS-RNA-AUROC which includes 61 downregulated and 39 upregulated transcripts in early stage of LIHC using Enrichr displayed in Figure 5 (A and B). The enrichment analysis of these transcripts indicates that downregulated transcripts are mostly enriched in metabolic processes related pathways; whereas, upregulated transcripts are associated with pathways that involved in normal functioning of liver. Further, downregulated transcripts enriched in cell cycle GO terms that contribute to cell growth while upregulated transcripts enriched in GO terms associated with common functioning of liver. In addition downregulated transcripts enriched in MSigDB oncogenic signatures i.e. RPS14_DN.V1_DN and CSR_LATE_UP.V1_UP, while upregulated transcripts enriched in PKCA_DN.V1_UP.

Multiple features based stage classification model using RNA transcripts

In addition to single feature based models, we have also developed models using expression of multiple RNA transcripts implementing different machine learning techniques. First, 30 RNA transcripts (named as LS-RNA-WEKA) selected from 1,293 RNA transcript using WEKA, subsequently prediction models developed for stage classification. As shown in Table 2, the Naive Bayes based model achieved

maximum accuracy of 74.55% with AUROC 0.79 on training dataset and accuracy of 76.06% with AUROC 0.79 on validation dataset respectively (Table 2). Among 30-signature RNA transcripts (LS-RNA-WEKA), 24 RNA transcripts have been found to be underexpressed, while 6 observed as overexpressed in early stage (Additional file 1, Table S4). Further among 30 transcripts,15 are protein coding genes which include *CSNK1D, DCK, ZNF576, RPP25, SLC22A10, FNTB, LCAT, MAT1A* and *CDCA5 etc*, 9 processed pseudogenesinclude GAPDHP63, RP11-829H16.2, AC018712.2 *etc*, 2 unprocessed pseudogenes, 1 each of lincRNA, miRNA,sense_intronic transcript. Figure 4B represents the expression pattern of these RNA transcripts among early and late stage tissue samples. In addition, we also developed prediction model using 100 features selected by F-ANOVA method. The performance (accuracy 70.42, AUROC 0.74 on independent validation dataset) of model has decreased with increasing the number of features as shown in Table_S5 (Additional file 1).

It has been observed that most of overexpressed genes such as *CDCA5, CSNK1D, DCK, ZNF576* and *RPP25 etc* in late stage are involved in cell growth promotion processes. While underexpressed genes such as *LCAT, CFHR3, RAMP3, MAT1A etc* associated with normal functioning of liver and immune functions as shown in Table_S6 (Additional file 1).

Table 2: The performance of stage classification models developed using 30 RNA transcripts selected using WEKA from 103 RNA transcripts (LS-RNA-WEKA).

Machine Learning Techniques	**Dataset**	**Performance Measures**				
		Sensitivity	**Specificity**	**Accuracy (%)**	**MCC**	**AUROC**
SVM	Training	80.43	73.05	76.7	0.54	0.8

	Validation	68.57	75	71.83	0.44	0.77
Random Forest	Training	81.88	74.47	78.14	0.56	0.84
	Validation	65.71	58.33	61.97	0.24	0.67
Naïve Bayes	Training	84.78	64.54	74.55	0.5	0.79
	Validation	80	72.22	76.06	0.52	0.79
SMO	Training	75.36	76.6	75.99	0.52	0.76
	Validation	62.86	77.78	70.42	0.41	0.7
J48	Training	68.12	72.34	70.25	0.4	0.7
	Validation	57.14	66.67	61.97	0.24	0.63

Stage classification models using hybrid features

In addition to single type of features (CpG methylation or RNA transcripts), we also developed model by combining both type of features. Prediction models were developed using 51 features (21 CpG sites and 30 RNA transcripts) selected by WEKA named as LS-CpG-RNA-hybrid to classify early and late stage tissue samples. The model developed using Naïve Bayes based model achieved the highest accuracy 78.14% with AUROC 0.81 on the training dataset and accuracy 78.87% and 0.82 on the independent dataset (Table 3). Also, we also developed models using selected 38 features (15CpG sites and 23RNA transcripts) obtained from 1740 features (1293 RNA transcripts and 447CpG sites), (Additional file 1, Table_S7). In summary, we got best performance of Naïve Bayes based model developed using 51 hybrid features (21 CpG sites and 30 RNA transcripts). In spite of the best efforts, we were capable to discriminate the early and late stage samples with maximum AUROC of 0.82.

Table 3: The performance of stage classification hybrid models developed using 51 features (LS-CpG-RNA-hybrid) that comprise 21 CpG sites and 30 RNA transcripts.

Machine Learning Techniques	**Dataset**	**Performance Measures**				
		Sensitivity	**Specificity**	**Accuracy (%)**	**MCC**	**AUROC**
SVM	Training	81.16	73.76	77.42	0.55	0.82
	Validation	80	75	77.46	0.55	0.8
Random Forest	Training	79.71	75.89	77.78	0.56	0.85
	Validation	71.43	72.22	71.83	0.44	0.79
Naïve Bayes	Training	89.13	67.38	78.14	0.58	0.81
	Validation	85.71	72.22	78.87	0.58	0.82
SMO	Training	83.33	73.05	78.14	0.57	0.78
	Validation	80	72.22	76.06	0.52	0.76
J48	Training	70.29	68.09	69.18	0.38	0.69
	Validation	65.71	75	70.42	0.41	0.75

<u>Models for classification of LIHC and normal samples</u>

Models for classification of LIHC and normal samples using CpG sites

In this study, we also developed models for discriminating LIHC and normal tissue samples using genomic and epigenetic profiles. These models were developed on 424 samples (50 normal and 374 LIHC), where training dataset consists of 339 samples (40 normal and 299 LIHC) and independent dataset consists of 85 samples (10 normal and 75 LIHC). We identified 1,386 CpG sites that are differentially methylated in LIHC and normal tissue samples (minimum mean difference of methylation beta values is 0.4) with Bonferroni *p*-value < 0.0001. These 1,386 sites contain 125 hypermethylated and 1,261 hypomethylated in cancer samples (Additional file 1,

Table_S8). Further, we developed single feature based models using these CpG sites and ranked them based on performance of models. Top 496 CpG sites named as LCN-CpG-AUROC (68 hypemethylated and 428 hypomethylatedCpG sites in LIHC) with AUROC equal or greater than 0.9 enlisted in (Additional file 1, Table_S9). The methylation pattern of top 496CpG sites in cancer and normal tissue samples is shown in Figure_S4 (Additional file 6). We have obtained 14 CpG sites that can discriminate LIHC and normal samples with precision (ROC >= 0.95). Single feature based models developed using CpG site cg07274716, cg24035245 and cg20172627 were able to classify normal and LIHC tissue with AUROC 0.97, 0.96 and 0.95 at threshold 0.35, 0.44, 0.32 and 0.12 respectively. Methylation pattern of top 10 CpG sites is shown by Heatmap (Additional file 7, Figure_S5A) and chromosome locations and associated genes of the CpG sites are represented in Circos plot (Additional file 8, Figure_S6).

In addition to the single feature based models, we also developed models using multiple features to classify LIHC and normal samples. Our SVM based model developed using 104 CpG sites (selected using WEKA from 496CpG sites) achieves maximum accuracy of 98-99 % and AUROC 0.99 on independent and training dataset. We further reduce features and developed models using 100, 50, 25, 20, 10 and 5 features or CpG sites (selected using F-ANOVA), (data not shown). Our models based on even small number of features (*i.e.* 5 CpG sites named as LCN-5-CpG) got reasonably high performance (ROC ~0.99). Models based on Random Forest, Naïve Bayes, SMO algorithms performed reasonably good with accuracy 95% - 97%, and AUROC of 0.94 – 0.97 on training dataset and accuracy 96 - 97% and AUROC of 0.94 – 99 on an independent validation dataset (Table 4). This observation shows that

LIHC samples can be discriminated with high precision even using small number of CpG methylation based biomarkers.

Table 4: The performance models developed for discriminating LIHC and normal samples, using 5 methylation CpG sites.

Machine Learning Techniques	**Dataset**	**Performance Measures**				
		Sensitivity	**Specificity**	**Accuracy (%)**	**MCC**	**AUROC**
SVM	Training	99.33	90	98.23	0.91	0.99
	Validation	98.67	90	97.65	0.89	0.99
Random Forest	Training	99	87.5	97.64	0.88	0.97
	Validation	98.67	90	97.65	0.89	0.99
Naïve Bayes	Training	96.66	90	95.87	0.82	0.94
	Validation	97.33	90	96.47	0.84	0.95
SMO	Training	99	90	97.94	0.9	0.94
	Validation	98.67	90	97.65	0.89	0.94
J48	Training	97.66	80	95.58	0.79	0.86
	Validation	98.67	80	96.47	0.82	0.89

To detect the biological relevance of the genes associated with these signature CpG sites (LCN-5-CpG), enrichment analysis performed using Enrichr. This analysis indicates *MYH9* is significantly involved in molecular functions such as actin-dependent ATPase activity (GO:0030898), microfilament motor activity (GO:0060002, GO:0060001). It is also involved in GO biological processes such as positive regulation of protein processing in phagocytic vesicle (GO:1903923), PMA-inducible membrane protein ectodomain proteolysis (GO:0051088), malonyl-CoA

metabolic process (GO:2001293). *ACSF2* is involved in acyl-CoA metabolic process (GO:0006637), malonyl-CoA metabolic process (GO:2001293), *etc*. Furthermore, the enrichment analysis signifies the involvement of *MYH9* in Nicotinic acetylcholine receptor signaling pathway, Cytoskeletal regulation by Rho GTPase, Inflammation mediated by chemokine and cytokine signaling pathway_Homo sapiens_P00031 PANTHER pathways. In addition, Enrichr suggest the enrichment of *ACSF2* in LEF1_UP.V1_DN, while *CHAD* involved in AKT_UP_MTOR_DN.V1_DN, and AKT_UP.V1_DN MSigDB Oncogenic Signatures set (*p*-value <0.05)

Models for classification of LIHC and normal samples using RNA transcripts

Moreover, we have also identified 147 RNA transcripts that are significantly differentially expressed in LIHC and normal samples (Additional file 1, Table_S10).The performance of models developed using top 53 RNA transcripts (ROC equal or greater than 0.9) is shown in Table_S11 (Additional file 1). The expression pattern of top 53RNA transcripts (named as LCN-RNA-AUROC) in cancer and normal tissue samples have been shown in heatmap (Additional file 9, Figure_S7). Among 53 transcripts, 50 are protein coding, one each of Linc-RNA, snoRNA and processed transcripts. Further, among them 18 and 35 observed to be overexpressed and underexpressed respectively in LIHC in comparison to normal tissue samples. Moreover, *CLEC4G*, *CLEC4M*, *FCN2*, *COLEC10*, *CLEC1B* and *PLVAP* can classify LIHC and normal tissue samples with AUROC >= 0.96. *CLEC4G* which is underexpressed gene in cancer is chief performer with AUROC ~0.99 at threshold 2.9. It means if the log2 of FPKM value of *CLEC4G* is less than 2.9 then sample belongs to cancer otherwise normal. *PLVAP* is overexpressed gene in cancer samples, it can distinguish cancer and normal samples with AUROC 0.97 at threshold 3.8. It suggested if the log2 of FPKM value of *PLVAP* above 3.8 then the

sample is cancer otherwise normal. Expression pattern of top 10 RNA transcripts displayed in heatmap (Additional file 7, Figure_S5B)

Gene enrichment analysis of LCN-RNA-AUROC is represented in Figure_S8A and Figure_S8B (Additional file 10). This functional enrichment of these genes suggested their involvement in various growth progression events or cancer development processes. As these genes upregulated in LIHC or cancerous condition; this might lay insight toward their contribution in the diseased state. While the underexpressed genes in LIHC are enriched in processes and pathways associated with normal activity or functioning of liver. Moreover they are also enriched in various MSigDB oncogenic signatures (Additional file 1, Table_S12).

Further, we also proceed to develop multiple genes based models that can outperform single gene based model. Thus we developed models based on multiple RNA transcripts/features selected by different techniques i.e., WEKA and F-ANOVA method. In this, an attempt has been made to develop prediction model using 35 WEKA selected features (from 147 RNA transcripts) to distinguish liver cancer samples from the normal samples. The SVM based model attains maximum accuracy nearly 99% with AUROC 0.99on both training and independent validation dataset respectively (result not shown). To develop the prediction model based on least number of genes/features, that can categorize LIHC samples from normal tissue samples with reasonably good accuracy; we firstly selected 25, 20, 15, 10 and 5 feature sets (results not shown) using F-ANOVA method. Then we developed prediction models on training dataset and independent validation dataset. The model's performance on independent validation dataset is similar to the performance obtained using CpG sites. Interestingly models based on these feature sets almost performed comparably to that of model based on 35 features selected by WEKA. 5-RNA feature

set (LCN-5-RNA) based SVM prediction model categorize samples with accuracy 98.53 and AUROC 0.97 of training dataset and with accuracy 97.65 and AUROC 0.93 of the independent validation dataset. In addition models based on the Random forest, SMO, Naïve Bayes and J48 algorithms also performed almost equally to that of SVM as indicated in Table 5.

Table 5: The performance of models developed for discriminating LIHC and normal samples using 5 RNA transcripts.

Machine Learning Techniques	**Dataset**	**Performance Measures**				
		Sensitivity	**Specificity**	**Accuracy (%)**	**MCC**	**AUROC**
SVM	Training	97.32	97.5	97.35	0.89	0.99
	Validation	97.33	90	96.47	0.84	0.97
Random Forest	Training	97.66	90	96.76	0.85	0.98
	Validation	98.67	80	96.47	0.82	0.89
Naïve Bayes	Training	96.66	97.5	96.76	0.86	0.97
	Validation	97.33	90	96.47	0.84	0.94
SMO	Training	97.32	97.5	97.35	0.89	0.97
	Validation	97.33	90	96.47	0.84	0.94
J48	Training	96.99	97.5	97.05	0.87	0.96
	Validation	98.67	90	97.65	0.89	0.94

Enrichment analysis of LCN-5-RNA using Enrichr implies the involvement of *CLEC4M* in Phagosome KEGG pathways, and *GDF2* in the TGF-beta signaling pathway_Homo sapiens_P00052 (p-value <0.05). *CLEC4M* involved in receptor-mediated virion attachment to host cell (GO:0046813), evasion or tolerance of host defenses by virus (GO:0019049) *etc*. This analysis also denotes their enrichment in

various MSigDB Oncogenic Signatures sets such as KRAS, p53, PTEN, mTOR and SNF5 [KRAS.AMP.LUNG_UP.V1_UP, P53_DN.V2_DN and PTEN_DN.V1_UP] etc. (p-value<0.05).

Multi-class classification

With aim to identify the features that can classify normal, early and late stage tissue samples, we have independently selected 33 CpG sites (multiclass-CpG) and 5 RNA transcripts (multiclass-RNA) using WEKA. The features are selected from 440 CpG sites and 236 RNA transcripts that are significantly differentially methylated and differentially expressed respectively (between 3 groups: cancer v/s normal and early stage v/s late stage). Here prediction models developed based on multiclass-CpG using various techniques of WEKA. Naïve Bayes model is top performer with accuracy 77.43 and 76.54 and weighted average AUROC 0.88 and 0.86 on training and independent validation dataset respectively (Table 6). The Boxplot represents the methylation pattern of multiclass-CpG features in Figure_S9A (Additional file 11). Also, we developed prediction models based on multiclass-RNA by implementing different techniques of WEKA. Naïve Bayes model achieve maximum performance with accuracy 72.73% and 72.84 and weighted average AUROC 0.81 and 0.80 on training and independent validation dataset respectively (Table 6). Boxplot represents the expression pattern of multiclass-RNA transcripts in Figure_S9B (Additional file 11). In addition, we have also developed prediction model based on 284 features selected by WEKA from all 60,483 RNA transcripts using different techniques. Model based on random forest (tree based algorithm) classify samples with accuracy 78.99% and 70.37%, weighted average AUROC 0.88 and 0.80 of training and independent dataset (Additional file 1, Table_S13). The performance of models exploiting other techniques lowers than of above mentioned models. Complete results

of prediction models based on 33 CpG sites 5 and 284RNA transcripts using various techniques implementing WEKA have been shown in (Additional file 1, Table_S14).

Table 6: The performance Naïve Bayes model in form of confusion matrix developed for classifying normal, early and late stage samples. Model was developed using33 CpG sites (multiclass-CpG) and 5 RNA transcripts (multiclass-RNA).

33CpG sites					
Training dataset					
			Actual	Accuracy	Weighted average ROC
Late Stage	Early Stage	Normal			
102	36	3	Late Stage		
29	107	2	Early Stage		
0	2	38	Normal	77.43	0.88
Validation dataset					
			Actual	Accuracy	Weighted average ROC
Late Stage	Early Stage	Normal			
25	10	1	Late Stage		
7	27	1	Early Stage		
0	0	10	Normal	76.54	0.86
5 RNA transcripts					
Training dataset					
			Actual	Accuracy	Weighted average ROC
Late Stage	Early Stage	Normal			
87	54	0	Late Stage	72.73	0.81

33	105	0	Early Stage		
0	0	40	Normal		
Validation dataset					
Predicted as			**Actual**	Accuracy	Weighted average ROC
Late Stage	Early Stage	Normal			
18	18	0	Late Stage		
4	31	0	Early Stage		
0	0	10	Normal	72.84	0.80

Enrichment analysis of multiclass-RNA is represented in (Additional file 12, Figure_S10 (A & B)). *MT1E, CNDP1, C7, GMNN* and *EEF1A2* are among the 5 RNA transcripts identified in multiclass classification. Enrichment analysis indicated upregulated genes such as *GMNN* and *EEF1A2* enriched in various growth promoting processes and pathways, while downregulated genes like *CNDP1* and *MT1E* enriched in negative regulators of growth and *C7* enriched in innate immunity pathways respectively.

Implementation of Webserver

To serve the scientific community, we developed CancerLSP (Liver cancer stage prediction), a web server that executes in silico prediction models developed in the present study for prediction of cancer status *i.e.* stage and analysis using methylation profiling and RNA expression data. Further, CancerLSP mainly has two modules; Prediction Module and Data analysis Module.

Prediction Module

This module permits the users to predict the disease status i.e. cancerous or normal and stage of cancer of sample using methylation beta values of CpG sites and FPKM quantification values of RNA transcripts. Here, the user required to submit FPKM values of RNA markers (RNA transcripts) and methylation beta values of marker CpG sites of every subject. The output result displays a list for patient samples and corresponding predicted status cancer or normal and early or late stage. Moreover, the user can select among the models developed from LS-CpG-WEKA, LS-RNA-WEKA LS-CpG-RNA-hybrid, LCN-5-CpG, LCN-5-CpG, multiclass-RNA and multiclass-RNA which have been identified as putative markers sets for stage progression of LIHC.

Data analysis Module

This module is beneficial in assessing the role of each RNA transcript and CpG site to distinguish early stage from the late stage. In addition, it provides *p*-value for every feature that indicates whether there is significant difference in methylation pattern of CpG site or RNA expression in early and late stage. Furthermore, it also provides threshold-based ROC of each feature along with mean methylation of CpG site and mean expression of that RNA transcript in the early and late stage of LIHC. This web server is available from URL http://webs.iiitd.edu.in/raghava/cancerlsp/ for scientific community.

Conclusion and Discussion

Worldwide HCC has become a major threat owing to its high morbidity and mortality rate. Furthermore, the appropriate treatment options are affected not only by the degree of liver dysfunction but also by tumor stage[34]. Therefore, accurate stage classification and prediction of disease at early stage is crucial for patient management. Hence, the primary goal of the present study is to identify the genomic and epigenomic features that distinguish early stage from the late stage of LIHC based on the TCGA datasets retrieved from GDC data portal. Accordingly, we have made an attempt to distinguish early stage from late stage tissue samples of LIHC employing methylation profiling and RNA expression profiling data. Firstly differentially methylated CpG sites and differentially expressed transcripts identified and are ranked based on AUROC to categorize early stage and late stage tissue samples using simple threshold-based classification method.

3 CpG sites i.e. cg07132710, cg11232136 and cg03108697 of LS-CPG-AUROC signatures, are hypomethylated in early stage of LIHC are associated with *GATA2*, *ZNF566* and *SWAP70* respectively. Earlier studies have shown that these genes are epigenetically repressed and associated with cancer progression in different malignancies [35-38].

Various genes such as *NCAPH*, *CYP4A*, *GLYATL1*, *HILPDA*, *CANT1*, *SLC2A2*, *ALDOA*, *CNFN*, *A1BG*, *CYP2B6*, *HN1* and *ITIH1 etc*. are among the LS-RNA-AUROC signatures. Previously, different reports uncover their involvement in a variety of malignancies [39-46]. Evidently, recent studies reveal the association of *HN1* and *SLC2A2* with survival of HCC patients and *ITIH1* with progression of HCC [47-49]. Beside the protein coding genes, LINC01146 a long non coding RNA was

earlier identified as differentially expressed in HCC [50, 51] can classify stage samples with AUROC 0.65.

Further using state of art machine learning techniques 21 CpG (LS-CPG-WEKA) methylation sites achieved AUROC of 0.78 on validation data and 30 RNA transcripts (LS-RNA-WEKA) independently distinguish early stage from late stage tissue samples of validation dataset with 0.77 AUROC.

Among LS-CPG-WEKA signature CpG sites such as cg16657244, cg11232136, cg11023721 and cg27111890 are associated with *NOLC1, ZNF566, SMAD7* and *UBASH3A* respectively. Previously the aberrant methylation pattern of these genes has been observed in different types of cancers [38, 52, 53].

Furthermore, among the identified RNA transcripts in LS-RNA-WEKA to classify early stage and late stage of LIHC, out of 15 protein coding transcripts *DCK, CDCA5, RPP25, FUT11, NECAB1, FNTB, ZNF576, NETO2* and *ELOVL3* are observed as upregulated genes in late stage and are involved in cell division processes, transcriptional regulation, while genes upregulated in early stage such as *MAT1A, LCAT, CFHR3* are involved in normal functioning of liver involving lipid and lipoprotein metabolism [54] (Table_S6). Previously, various studies revealed the role of *MAT1A* in HCC [55-57]. Two of the genes in our biomarker panel *i.e. CDCA5* and *RAMP3* already identified as prognostic markers in HCC in literature [58, 59]. In recent past, role of pseudogenes has been revealed in pathogenesis of cancer [60-62] including HCC (25391452, 28823960). LS-RNA-WEKA also contains 12 pseudogenes (9 processed, 2 unprocessed and 1 transcribed processed pseudogenes) including RP3-375P9.2, GAPDHP63, PTMAP2, RPS3P7, FABP5P3, HNRNPA1P37, AC018712.2 *etc*. They are significantly upregulated in late stage of HCC. One of the studies reveals the role of RP3-375P9.2 in breast cancer [63].

Models based on 21 methylation CpG sites and 30 RNA transcripts achieved reasonable performance independently for classifying early and late stage tissue samples. On combining these features, hybrid models are able to classify samples with higher performance with accuracy of nearly 78% and AUROC 0.82 on independent validation dataset.

In this study, we have also identified top 10 CpG sites and top 10 RNA transcripts as candidate signatures based on their individual performance in terms of AUROC >0.9 to classify LIHC samples and normal liver tissue samples using simple threshold based approach.*CLEC4G* and CpG site cg07274716 (associated with *PITX1*) distinguish cancer samples from normal with AUROC 0.98 and 0.97 respectively. Earlier studies have shown the association of downregulation of CLEC4G expression with the progression of HCC [64, 65]. Our study corroborates this signature marker in HCC and further extends the importance of this gene as prognostic marker. Previous reports indicate that the hypermethylation of the *PITX1* correlated with tumor progression HNSCC [66] and ESCC [67]. Similarly, present study reveals the association hypermethylation of *PITX1* with progression of HCC.

Furthermore, prediction models developed using multiple features filtered by different techniques to distinguish cancer from control samples. Interestingly, our study shown that the model based on 5 features (RNA transcripts or CpG sites) performed quite similar to that of models based on much large dimension of features as indicated by the performance to discriminate cancer and normal samples with 96.47% accuracy and AUROC 0.97 on independent validation dataset. In current study, *CLEC4G*, *GDF5*, *CLEC1B*, *CLEC4M* and *FCN2* are among *LCN-5-RNA* signature that can classify cancer and normal tissue sample with high precision. Earlier, *CLEC4M* and *GDF5* have been recognized as differentially expressed gene in various other cancers

[68-70], while *CLEC4G*, *CLEC1B* and *FCN2* were identified as genes that associated with cancer progression and prognosis in hepatocellular carcinoma [64, 71-73]. Overall this analysis emphasizes the role of these signature genes in oncogenesis.

Interestingly one of the genes, *LCAT* is able to classify cancer and normal samples with AUROC 0.92, is also one of important markers to distinguish late stage samples from early stage. Thus our report suggests the role of *LCAT* as important signature marker progression of LIHC.

In addition to binary class classification, we have also developed two multiclass prediction models that can distinguish normal, early stage and late stage tissue samples based to 5 RNA transcripts sites with 72.84% accuracy with weighted AUROC of 0.81 and 33 CpG sites accuracy of 77.54% and weighted AUROC 0.86. Their enrichment analysis shows that upregulated genes enriched in various growth enhancing processes and pathways such as positives regulation of various kinase signaling, GTPase; while downregulated genes enriched in negative regulators of growth processes and innate immunity pathways. This might lay insight toward their involvement in progression of cancerous condition.

We have achieved only reasonable performance to classify early stage tissue samples from late stage tissue samples despite we tried different approaches to classify early stage tissue samples from late stage using methylation and expression data. These results suggest that stage classification is quite difficult and challenging task in comparison to cancer versus normal classification on the basis of their expression and methylation profiling. Moreover, HCC is a complex disease with various pathological bases, thus the exploration of biomarker combinations might offer more accurate and precise diagnosis or prognosis of HCC.

Potential Implications

We anticipate this study would be beneficial to recognize these important epigenetic and genomic-based signature markers for early diagnosis of liver cancer. We have integrated all our models in the form of webserver named CancerLSP for the use of scientific community. It has mainly two modules; Prediction Module and Data analysis Module. The prediction module allows the users to predict the disease status of sample *i.e.* early or late stage of cancer and cancerous or normal using methylation beta values of CpG sites and FPKM quantification values of RNA transcripts. Further, Data analysis module is beneficial in assessing the role of each RNA transcript and CpG site to distinguish early stage from the late stage

Methods

Pre-processing of Data

Methylation Data

There are total 4,85,577 Methylation CpG site (Probe IDs associated with CpG sites) for each tissue sample. The methylation score for every CpG site was defined in terms of beta value. All those CpG sites approximately 23-25% excluded from the study for which beta value is missing among any of samples using in-house bash script. Hence CpG sites number reduced to 3,74,292 for staging analysis and 3,69,221 for Cancer v/s Normal analysis.

Normalization of RNA Expression

The expression values of RNA transcripts from GDC portal were obtained in term of FPKM (fragments per kilobase of transcript per million mapped reads). There is a wide range of variation in FPKM values, thus we transformed values using log2 after addition of 1.0 as a constant number to each of FPKM value. Further those features have been

removed that have low variance using *caret* package in R, followed by z-score normalization of data. Thus, for each mRNA log_2-transformed FPKM values were centered and scaled by employing *caret* package in R. Following equations were used for computing the transformation and normalization:

$$x = log_2\,(FPKM + 1) \qquad (1)$$

$$Z_{_score} = \frac{x - \bar{x}}{sd} \qquad (2)$$

Where $Z_{_score}$ is the normalized score, x is the log-transformed expression, $\bar{x}$ is the mean of expression and *sd* is the standard deviation of expression.

Identification of differentially methylated CpG sites and expressed RNA transcripts

In order to identify differentially CpG methylated sites, first we computed the mean methylation score for each CpG site in early and late stage samples. Secondly, we computed whether difference in the mean of methylation score in early and late stage samples is statistically significant or not using *t-test*. Here, we implemented Welch t-test using in-house Rscript and bash script. Similarly, we have identified differentially expressed RNA transcripts in early and late stage samples from total of 60,483 RNA transcripts.

Feature Selection Techniques

One of challenges in developing prediction model is to identify important features from the large dimension of features. In this study, we used number of techniques for feature selection. Firstly, we used Area under Receiver operating characteristic (AUROC) based feature selection technique in which we developed single feature based models for discriminating early and late stage samples. Single feature based models are also called threshold based models in which feature having a score above threshold is assigned to early stage if it is upregulated in early stage [74]. We compute performance of each model based on a given feature and identify features having

highest performance in term of AUROC.Secondly, we perform feature selection using two different algorithms like employing attribute evaluator named, 'SymmetricalUncertAttributeSetEval' with search method of 'FCBFSearch' of WEKA software package and sklearn.feature_selection. F-ANOVA method using Scikit package.We also filtered CpG sites and RNA transcripts as features that can discriminate early stage from late stage cancer samples using attribute evaluator named, 'SymmetricalUncertAttributeSetEval' with search method of 'FCBFSearch' of WEKA software package. The FCBF (Fast Correlation-Based Feature) algorithm employed correlation to identify relevant features in high-dimensional datasets in small feature space [75].

Implementation of machine learning techniques

Primarily we have developed the Support Vector Machine (SVM) based prediction models using the package SVM$^{light\ [76]}$ and WEKA [75].In present study RBF (radial basis function) kernel employed to optimize various parameters to get best performance on training dataset.Furthermore, some of commonly used classifiers were also used for developing prediction models. These classifiers include Random forests, SMO, Naïve Bayes, J48 were implemented exploiting WEKA software.

Performance Evaluation of Models

In current study, we used both internal and external validation technique to evaluate the performance of models. Main dataset has been divided in two sets called training dataset and validation dataset in ratio of 80:20. The training set is used for developing model and for performing internal validation; whereas, validation dataset is used to perform external or independent validation. In this article, we used ten-fold cross validation technique where training dataset is randomly split into ten sets; of which nine out of ten sets were used as training sets and the remaining tenth set as testing dataset. This

process is repeated ten times in such a way that each set is exploited once for testing. The final performance of model is the mean performance of all the ten sets. In order to avoid optimization of parameters in case of ten-fold cross-validation, we also implement external validation. In case of external validation, we evaluate our model on an independent or external dataset not used for training. In this study**,** we used 20% of the main dataset for validation or independent testing and remaining 80% dataset for training.

In order to measure performance of models, we used standard parameters, commonly used to measure performance of classification models. Both threshold-dependent and threshold-independent parameters were employed to measure performance. In case of threshold-dependent parameters, we measure sensitivity, specificity, accuracy and Matthew's correlation coefficient (MCC) using following equations.

$$Sensitivity\ (Sn) = \frac{TP}{TP+FN} * 100 \qquad (1)$$

$$Specificity\ (Sp) = \frac{TN}{TN+FP} * 100 \qquad (2)$$

$$Accuracy\ (Ac) = \frac{TP+TN}{TP+FP+TN+FN} * 100 \qquad (3)$$

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (4)$$

where, FP, FN, TP and TN are false positive, false negative true positive and true negative predictions respectively.

While, for threshold-independent measures, we used standard parameter Area Under the Receiver Operating Characteristic curve (AUROC) or commonly known as Area Under (AUC) or Receiver Operating Characteristic curve (ROC).The AUROC curve is generated by plotting sensitivity or true positive rate against the false positive rate (1-

specificity) at various thresholds. Finally, the area under AUROC curve calculated to compute a single parameter called AUROC.

Functional annotation or enrichment analysis of genes

In order to discern the biological relevance of the signature genes or the genes associated with signature CpG sites, enrichment analysis performed using Enrichr [77, 78]. Enrichr applies Fisher exact test to identify enrichment score. It also provides Z-score which is derived by applying correction on a Fisher Exact test.

Contributions

H.K. and S.B. collected the data and created the datasets, developed classification programs, implemented algorithms, created the back-end server and front-end user interface. H.K., S.B. and G.P.S.R. analyzed the results. H.K. and S.B. wrote the manuscript. G.P.S.R. conceived and coordinated the project, helped in the interpretation and analysis of data, refined the drafted manuscript and gave complete supervision to the project. All of the authors read and approved the final manuscript.

Competing interests

The authors declare no competing financial interests.

Acknowledgement

The authors acknowledge funding agencies J. C. Bose National Fellowship (DST). HK and SB are thankful to CSIR and ICMR for providing fellowships.

Abbreviations

MCC: Matthews correlation coefficient; AUROC: Area under Receiver Operating Characteristics; FDR: False Discovery Rate

References

1. Siegel RL, Miller KD and Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68 1:7-30. doi:10.3322/caac.21442.
2. Herceg Z and Paliwal A. Epigenetic mechanisms in hepatocellular carcinoma: how environmental factors influence the epigenome. Mutat Res. 2011;727 3:55-61. doi:10.1016/j.mrrev.2011.04.001.
3. El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. Gastroenterology. 2012;142 6:1264-73 e1. doi:10.1053/j.gastro.2011.12.061.
4. Poon D, Anderson BO, Chen LT, Tanaka K, Lau WY, Van Cutsem E, et al. Management of hepatocellular carcinoma in Asia: consensus statement from the Asian Oncology Summit 2009. Lancet Oncol. 2009;10 11:1111-8. doi:10.1016/S1470-2045(09)70241-4.
5. Sukowati CH, El-Khobar KE, Ie SI, Anfuso B, Muljono DH and Tiribelli C. Significance of hepatitis virus infection in the oncogenic initiation of hepatocellular carcinoma. World J Gastroenterol. 2016;22 4:1497-512. doi:10.3748/wjg.v22.i4.1497.
6. Kimhofer T, Fye H, Taylor-Robinson S, Thursz M and Holmes E. Proteomic and metabonomic biomarkers for hepatocellular carcinoma: a comprehensive review. Br J Cancer. 2015;112 7:1141-56. doi:10.1038/bjc.2015.38.
7. Debruyne EN and Delanghe JR. Diagnosing and monitoring hepatocellular carcinoma with alpha-fetoprotein: new aspects and applications. Clin Chim Acta. 2008;395 1-2:19-26. doi:10.1016/j.cca.2008.05.010.
8. IuS T. [Detection of Embryo-Specific Alpha-Globulin in the Blood Serum of a Patient with Primary Liver Cancer]. Vopr Med Khim. 1964;10:90-1.
9. Bruix J, Sherman M and American Association for the Study of Liver D. Management of hepatocellular carcinoma: an update. Hepatology. 2011;53 3:1020-2. doi:10.1002/hep.24199.
10. Lou J, Zhang L, Lv S, Zhang C and Jiang S. Biomarkers for Hepatocellular Carcinoma. Biomark Cancer. 2017;9:1-9. doi:10.1177/1179299X16684640.
11. Li D and Satomura S. Biomarkers for Hepatocellular Carcinoma (HCC): An Update. Adv Exp Med Biol. 2015;867:179-93. doi:10.1007/978-94-017-7215-0_12.
12. Huang J and Zeng Y. Current clinical uses of the biomarkers for hepatocellular carcinoma. Drug Discov Ther. 2014;8 2:98-9.
13. Li L, Lei Q, Zhang S, Kong L and Qin B. Screening and identification of key biomarkers in hepatocellular carcinoma: Evidence from bioinformatic analysis. Oncol Rep. 2017;38 5:2607-18. doi:10.3892/or.2017.5946.
14. Umeda S, Kanda M and Kodera Y. Emerging evidence of molecular biomarkers in hepatocellular carcinoma. Histol Histopathol. 2017:11936. doi:10.14670/HH-11-936.
15. Zheng H, Jiang WH, Tian T, Tan HS, Chen Y, Qiao GL, et al. CBX6 overexpression contributes to tumor progression and is predictive of a poor prognosis in hepatocellular carcinoma. Oncotarget. 2017;8 12:18872-84. doi:10.18632/oncotarget.14770.
16. Shimizu D, Kanda M, Sugimoto H, Sueoka S, Takami H, Ezaka K, et al. NRAGE promotes the malignant phenotype of hepatocellular carcinoma. Oncol Lett. 2016;11 3:1847-54. doi:10.3892/ol.2016.4120.

749 17. Xiao S, Chang RM, Yang MY, Lei X, Liu X, Gao WB, et al. Actin-like 6A predicts
750 poor prognosis of hepatocellular carcinoma and promotes metastasis and
751 epithelial-mesenchymal transition. Hepatology. 2016;63 4:1256-71.
752 doi:10.1002/hep.28417.
753 18. Tang B, Tang F, Li B, Yuan S, Xu Q, Tomlinson S, et al. High USP22 expression
754 indicates poor prognosis in hepatocellular carcinoma. Oncotarget. 2015;6
755 14:12654-67. doi:10.18632/oncotarget.3705.
756 19. Hu B, Jiang D, Chen Y, Wei L, Zhang S, Zhao F, et al. High CHMP4B expression
757 is associated with accelerated cell proliferation and resistance to doxorubicin
758 in hepatocellular carcinoma. Tumour Biol. 2015;36 4:2569-81.
759 doi:10.1007/s13277-014-2873-1.
760 20. Shi Z, Liu J, Yu X, Huang J, Shen S, Zhang Y, et al. Loss of FOXF2 Expression
761 Predicts Poor Prognosis in Hepatocellular Carcinoma Patients. Ann Surg
762 Oncol. 2016;23 1:211-7. doi:10.1245/s10434-015-4515-2.
763 21. Jiang F, Chen L, Yang YC, Wang XM, Wang RY, Li L, et al. CYP3A5 Functions as
764 a Tumor Suppressor in Hepatocellular Carcinoma by Regulating mTORC2/Akt
765 Signaling. Cancer Res. 2015;75 7:1470-81. doi:10.1158/0008-5472.CAN-14-
766 1589.
767 22. Kanda M, Sugimoto H, Nomoto S, Oya H, Hibino S, Shimizu D, et al. Bcell
768 translocation gene 1 serves as a novel prognostic indicator of hepatocellular
769 carcinoma. Int J Oncol. 2015;46 2:641-8. doi:10.3892/ijo.2014.2762.
770 23. Mehra M and Chauhan R. Long Noncoding RNAs as a Key Player in
771 Hepatocellular Carcinoma. Biomark Cancer. 2017;9:1179299X17737301.
772 doi:10.1177/1179299X17737301.
773 24. Niu ZS, Niu XJ and Wang WH. Long non-coding RNAs in hepatocellular
774 carcinoma: Potential roles and clinical implications. World J Gastroenterol.
775 2017;23 32:5860-74. doi:10.3748/wjg.v23.i32.5860.
776 25. Fang TT, Sun XJ, Chen J, Zhao Y, Sun RX, Ren N, et al. Long non-coding RNAs
777 are differentially expressed in hepatocellular carcinoma cell lines with
778 differing metastatic potential. Asian Pac J Cancer Prev. 2014;15 23:10513-24.
779 26. Li T, Xie J, Shen C, Cheng D, Shi Y, Wu Z, et al. Upregulation of long noncoding
780 RNA ZEB1-AS1 promotes tumor metastasis and predicts poor prognosis in
781 hepatocellular carcinoma. Oncogene. 2016;35 12:1575-84.
782 doi:10.1038/onc.2015.223.
783 27. Hua L, Wang CY, Yao KH, Chen JT, Zhang JJ and Ma WL. High expression of
784 long non-coding RNA ANRIL is associated with poor prognosis in
785 hepatocellular carcinoma. Int J Clin Exp Pathol. 2015;8 3:3076-82.
786 28. Pezzuto F, Izzo F, Buonaguro L, Annunziata C, Tatangelo F, Botti G, et al.
787 Tumor specific mutations in TERT promoter and CTNNB1 gene in hepatitis B
788 and hepatitis C related hepatocellular carcinoma. Oncotarget. 2016;7
789 34:54253-62. doi:10.18632/oncotarget.9801.
790 29. Niu ZS, Niu XJ and Wang WH. Genetic alterations in hepatocellular
791 carcinoma: An update. World J Gastroenterol. 2016;22 41:9069-95.
792 doi:10.3748/wjg.v22.i41.9069.
793 30. Saito Y, Hibino S and Saito H. Alterations of epigenetics and microRNA in
794 hepatocellular carcinoma. Hepatol Res. 2014;44 1:31-42.
795 doi:10.1111/hepr.12147.

796 31. Zhang Y. Detection of epigenetic aberrations in the development of
797 hepatocellular carcinoma. Methods Mol Biol. 2015;1238:709-31.
798 doi:10.1007/978-1-4939-1804-1_37.
799 32. Esteller M, Corn PG, Baylin SB and Herman JG. A gene hypermethylation
800 profile of human cancer. Cancer Res. 2001;61 8:3225-9.
801 33. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA
802 methylation array with single CpG site resolution. Genomics. 2011;98 4:288-
803 95. doi:10.1016/j.ygeno.2011.07.007.
804 34. European Association For The Study Of The L, European Organisation For R
805 and Treatment Of C. EASL-EORTC clinical practice guidelines: management of
806 hepatocellular carcinoma. J Hepatol. 2012;56 4:908-43.
807 doi:10.1016/j.jhep.2011.12.001.
808 35. Tessema M, Yingling CM, Snider AM, Do K, Juri DE, Picchi MA, et al. GATA2 is
809 epigenetically repressed in human and mouse lung tumors and is not
810 requisite for survival of KRAS mutant lung cancer. J Thorac Oncol. 2014;9
811 6:784-93. doi:10.1097/JTO.0000000000000165.
812 36. Kandimalla R, van Tilborg AA, Kompier LC, Stumpel DJ, Stam RW, Bangma CH,
813 et al. Genome-wide analysis of CpG island methylation in bladder cancer
814 identified TBX2, TBX3, GATA2, and ZIC4 as pTa-specific prognostic markers.
815 Eur Urol. 2012;61 6:1245-56. doi:10.1016/j.eururo.2012.01.011.
816 37. Chiyomaru T, Tatarano S, Kawakami K, Enokida H, Yoshino H, Nohata N, et al.
817 SWAP70, actin-binding protein, function as an oncogene targeting tumor-
818 suppressive miR-145 in prostate cancer. Prostate. 2011;71 14:1559-67.
819 doi:10.1002/pros.21372.
820 38. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al.
821 Chromosome-wide and promoter-specific analyses identify sites of
822 differential DNA methylation in normal and transformed human cells. Nat
823 Genet. 2005;37 8:853-62. doi:10.1038/ng1598.
824 39. Chen XW, Yu TJ, Zhang J, Li Y, Chen HL, Yang GF, et al. CYP4A in tumor-
825 associated macrophages promotes pre-metastatic niche formation and
826 metastasis. Oncogene. 2017;36 35:5045-57. doi:10.1038/onc.2017.118.
827 40. Yin L, Jiang LP, Shen QS, Xiong QX, Zhuo X, Zhang LL, et al. NCAPH plays
828 important roles in human colon cancer. Cell Death Dis. 2017;8 3:e2680.
829 doi:10.1038/cddis.2017.88.
830 41. Zhang F, Lin JD, Zuo XY, Zhuang YX, Hong CQ, Zhang GJ, et al. Elevated
831 transcriptional levels of aldolase A (ALDOA) associates with cell cycle-related
832 genes in patients with NSCLC and several solid tumors. BioData Min.
833 2017;10:6. doi:10.1186/s13040-016-0122-4.
834 42. Kawai K, Uemura M, Munakata K, Takahashi H, Haraguchi N, Nishimura J, et
835 al. Fructose-bisphosphate aldolase A is a key regulator of hypoxic adaptation
836 in colorectal cancer cells and involved in treatment resistance and poor
837 prognosis. Int J Oncol. 2017;50 2:525-34. doi:10.3892/ijo.2016.3814.
838 43. Royse KE, Zhi D, Conner MG, Clodfelder-Miller B, Srinivasasainagendra V,
839 Vaughan LK, et al. Differential Gene Expression Landscape of Co-Existing
840 Cervical Pre-Cancer Lesions Using RNA-seq. Front Oncol. 2014;4:339.
841 doi:10.3389/fonc.2014.00339.

842 44. Canales NA, Marina VM, Castro JS, Jimenez AA, Mendoza-Hernandez G, Mc
843 CE, et al. A1BG and C3 are overexpressed in patients with cervical
844 intraepithelial neoplasia III. Oncol Lett. 2014;8 2:939-47.
845 doi:10.3892/ol.2014.2195.
846 45. Gerhardt J, Steinbrech C, Buchi O, Behnke S, Bohnert A, Fritzsche F, et al. The
847 androgen-regulated Calcium-Activated Nucleotidase 1 (CANT1) is commonly
848 overexpressed in prostate cancer and is tumor-biologically relevant in vitro.
849 Am J Pathol. 2011;178 4:1847-60. doi:10.1016/j.ajpath.2010.12.046.
850 46. Chen Q, Watson JT, Marengo SR, Decker KS, Coleman I, Nelson PS, et al. Gene
851 expression in the LNCaP human prostate cancer progression model:
852 progression associated expression in vitro corresponds to expression changes
853 associated with prostate cancer progression in vivo. Cancer Lett. 2006;244
854 2:274-88. doi:10.1016/j.canlet.2005.12.027.
855 47. Kim YH, Jeong DC, Pak K, Han ME, Kim JY, Liangwen L, et al. SLC2A2 (GLUT2)
856 as a novel prognostic factor for hepatocellular carcinoma. Oncotarget. 2017;8
857 40:68381-92. doi:10.18632/oncotarget.20266.
858 48. Nault JC, De Reynies A, Villanueva A, Calderaro J, Rebouissou S, Couchy G, et
859 al. A hepatocellular carcinoma 5-gene score associated with survival of
860 patients after liver resection. Gastroenterology. 2013;145 1:176-87.
861 doi:10.1053/j.gastro.2013.03.051.
862 49. Hamm A, Veeck J, Bektas N, Wild PJ, Hartmann A, Heindrichs U, et al.
863 Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITIH)
864 genes in multiple human solid tumors: a systematic expression analysis. BMC
865 Cancer. 2008;8:25. doi:10.1186/1471-2407-8-25.
866 50. Esposti DD, Hernandez-Vargas H, Voegele C, Fernandez-Jimenez N, Forey N,
867 Bancel B, et al. Identification of novel long non-coding RNAs deregulated in
868 hepatocellular carcinoma using RNA-sequencing. Oncotarget. 2016;7
869 22:31862-77. doi:10.18632/oncotarget.7364.
870 51. Miller KL. Nurse executive leadership: a corporate perspective. Nurs Adm Q.
871 1989;13 2:12-8.
872 52. Duan X, Zhang J, Liu S, Zhang M, Wang Q and Cheng J. Methylation of
873 nucleolar and coiled-body phosphoprotein 1 is associated with the
874 mechanism of tumorigenesis in hepatocellular carcinoma. Oncol Rep.
875 2013;30 5:2220-8. doi:10.3892/or.2013.2676.
876 53. Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance
877 M, et al. DNA methylation profiling reveals a predominant immune
878 component in breast cancers. EMBO Mol Med. 2011;3 12:726-41.
879 doi:10.1002/emmm.201100801.
880 54. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et
881 al. STRING v10: protein-protein interaction networks, integrated over the
882 tree of life. Nucleic Acids Res. 2015;43 Database issue:D447-52.
883 doi:10.1093/nar/gku1003.
884 55. Ramani K and Lu SC. Methionine adenosyltransferases in liver health and
885 diseases. Liver Res. 2017;1 2:103-11. doi:10.1016/j.livres.2017.07.002.
886 56. Lu SC and Mato JM. S-adenosylmethionine in liver health, injury, and cancer.
887 Physiol Rev. 2012;92 4:1515-42. doi:10.1152/physrev.00047.2011.

888 57. Lu SC and Mato JM. Role of methionine adenosyltransferase and S-
889 adenosylmethionine in alcohol-associated liver cancer. Alcohol. 2005;35
890 3:227-34. doi:10.1016/j.alcohol.2005.03.011.
891 58. Fang A, Zhou S, Su X, Liu C, Chen X, Wan Y, et al. RAMP3 is a prognostic
892 indicator of liver cancer and might reduce the adverse effect of TP53
893 mutation on survival. Future Oncol. 2018; doi:10.2217/fon-2018-0296.
894 59. Shen Z, Yu X, Zheng Y, Lai X, Li J, Hong Y, et al. CDCA5 regulates proliferation
895 in hepatocellular carcinoma and has potential as a negative prognostic
896 marker. Onco Targets Ther. 2018;11:891-901. doi:10.2147/OTT.S154754.
897 60. Shi X, Nie F, Wang Z and Sun M. Pseudogene-expressed RNAs: a new frontier
898 in cancers. Tumour Biol. 2016;37 2:1471-8. doi:10.1007/s13277-015-4482-z.
899 61. Xiao-Jie L, Ai-Mei G, Li-Juan J and Jiang X. Pseudogene in cancer: real
900 functions and promising signature. J Med Genet. 2015;52 1:17-24.
901 doi:10.1136/jmedgenet-2014-102785.
902 62. Cooke SL, Shlien A, Marshall J, Pipinikas CP, Martincorena I, Tubio JM, et al.
903 Processed pseudogenes acquired somatically during cancer development.
904 Nat Commun. 2014;5:3644. doi:10.1038/ncomms4644.
905 63. Ampuja M, Rantapero T, Rodriguez-Martinez A, Palmroth M, Alarmo EL,
906 Nykter M, et al. Integrated RNA-seq and DNase-seq analyses identify
907 phenotype-specific BMP4 signaling in breast cancer. BMC Genomics. 2017;18
908 1:68. doi:10.1186/s12864-016-3428-1.
909 64. Chen J, Qian Z, Li F, Li J and Lu Y. Integrative Analysis of Microarray Data to
910 Reveal Regulation Patterns in the Pathogenesis of Hepatocellular Carcinoma.
911 Gut Liver. 2017;11 1:112-20. doi:10.5009/gnl16063.
912 65. Hodo Y, Hashimoto S, Honda M, Yamashita T, Suzuki Y, Sugano S, et al.
913 Comprehensive gene expression analysis of 5'-end of mRNA identified novel
914 intronic transcripts associated with hepatocellular carcinoma. Genomics.
915 2010;95 4:217-23. doi:10.1016/j.ygeno.2010.01.004.
916 66. Sailer V, Charpentier A, Dietrich J, Vogt TJ, Franzen A, Bootz F, et al.
917 Intragenic DNA methylation of PITX1 and the adjacent long non-coding RNA
918 C5orf66-AS1 are prognostic biomarkers in patients with head and neck
919 squamous cell carcinomas. PLoS One. 2018;13 2:e0192742.
920 doi:10.1371/journal.pone.0192742.
921 67. Otsubo T, Yamada K, Hagiwara T, Oshima K, Iida K, Nishikata K, et al. DNA
922 hypermethyation and silencing of PITX1 correlated with advanced stage and
923 poor postoperative prognosis of esophageal squamous cell carcinoma.
924 Oncotarget. 2017;8 48:84434-48. doi:10.18632/oncotarget.21375.
925 68. Margheri F, Schiavone N, Papucci L, Magnelli L, Serrati S, Chilla A, et al. GDF5
926 regulates TGFss-dependent angiogenesis in breast carcinoma MCF-7 cells: in
927 vitro and in vivo control by anti-TGFss peptides. PLoS One. 2012;7 11:e50342.
928 doi:10.1371/journal.pone.0050342.
929 69. Jin K, Li G, Cui B, Zhang J, Lan H, Han N, et al. Assessment of a novel VEGF
930 targeted agent using patient-derived tumor tissue xenograft models of colon
931 carcinoma with lymphatic and hepatic metastases. PLoS One. 2011;6
932 12:e28384. doi:10.1371/journal.pone.0028384.
933 70. Potapenko IO, Haakensen VD, Luders T, Helland A, Bukholm I, Sorlie T, et al.
934 Glycan gene expression signatures in normal and malignant breast tissue;

possible role in diagnosis and progression. Mol Oncol. 2010;4 2:98-118. doi:10.1016/j.molonc.2009.12.001.

71. Yin F, Shu L, Liu X, Li T, Peng T, Nan Y, et al. Microarray-based identification of genes associated with cancer progression and prognosis in hepatocellular carcinoma. J Exp Clin Cancer Res. 2016;35 1:127. doi:10.1186/s13046-016-0403-2.

72. Yang G, Liang Y, Zheng T, Song R, Wang J, Shi H, et al. FCN2 inhibits epithelial-mesenchymal transition-induced metastasis of hepatocellular carcinoma via TGF-beta/Smad signaling. Cancer Lett. 2016;378 2:80-6. doi:10.1016/j.canlet.2016.05.007.

73. Xu G, Yang F, Ding CL, Zhao LJ, Ren H, Zhao P, et al. Small nucleolar RNA 113-1 suppresses tumorigenesis in hepatocellular carcinoma. Mol Cancer. 2014;13:216. doi:10.1186/1476-4598-13-216.

74. Bhalla S, Chaudhary K, Kumar R, Sehgal M, Kaur H, Sharma S, et al. Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. Sci Rep. 2017;7:44997. doi:10.1038/srep44997.

75. Frank E, Hall M, Trigg L, Holmes G and Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004;20 15:2479-81. doi:10.1093/bioinformatics/bth261.

76. Joachims T. Learning to Classify Text Using Support Vector Machines.: Springer; 2002.

77. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44 W1:W90-7. doi:10.1093/nar/gkw377.

78. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14:128. doi:10.1186/1471-2105-14-128.

Figures:

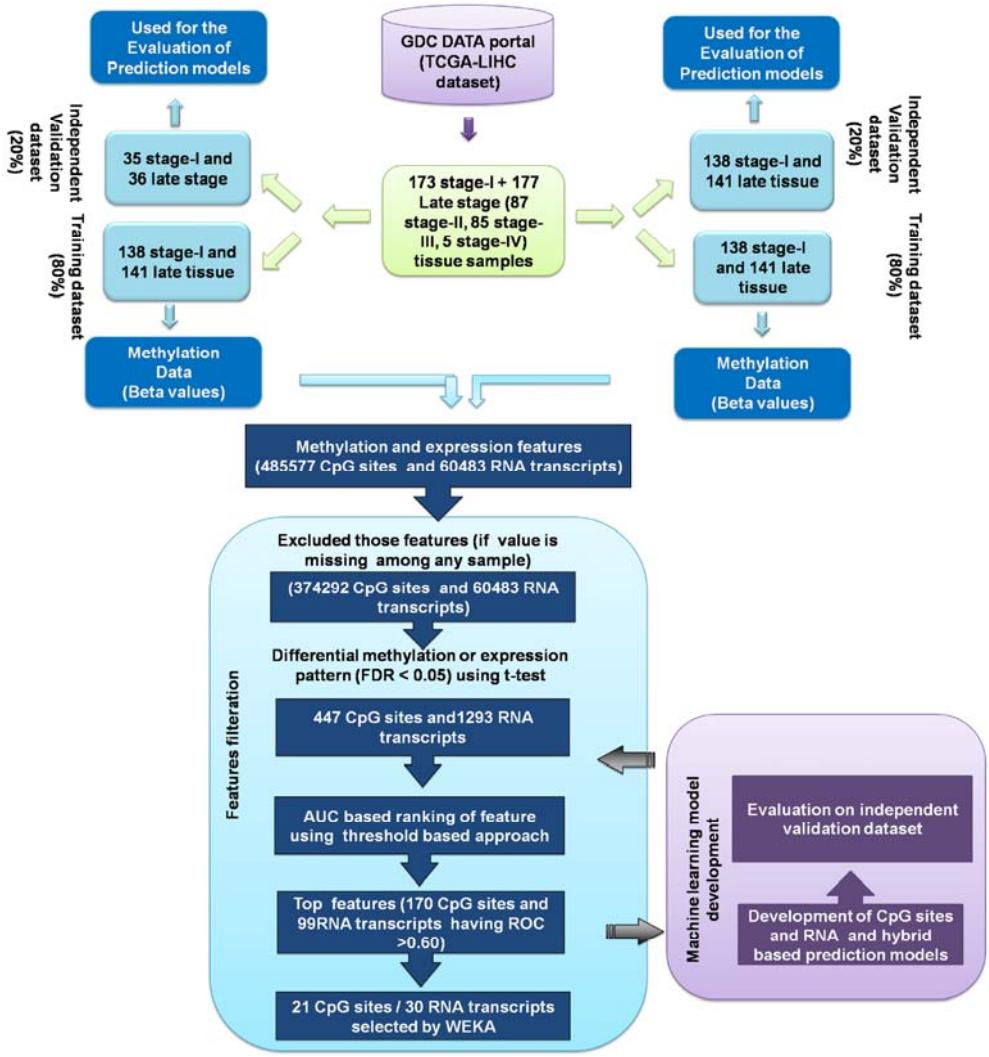


Figure 1: The workflow representing the analysis of methylation and expression profile of LIHC early and late stage samples

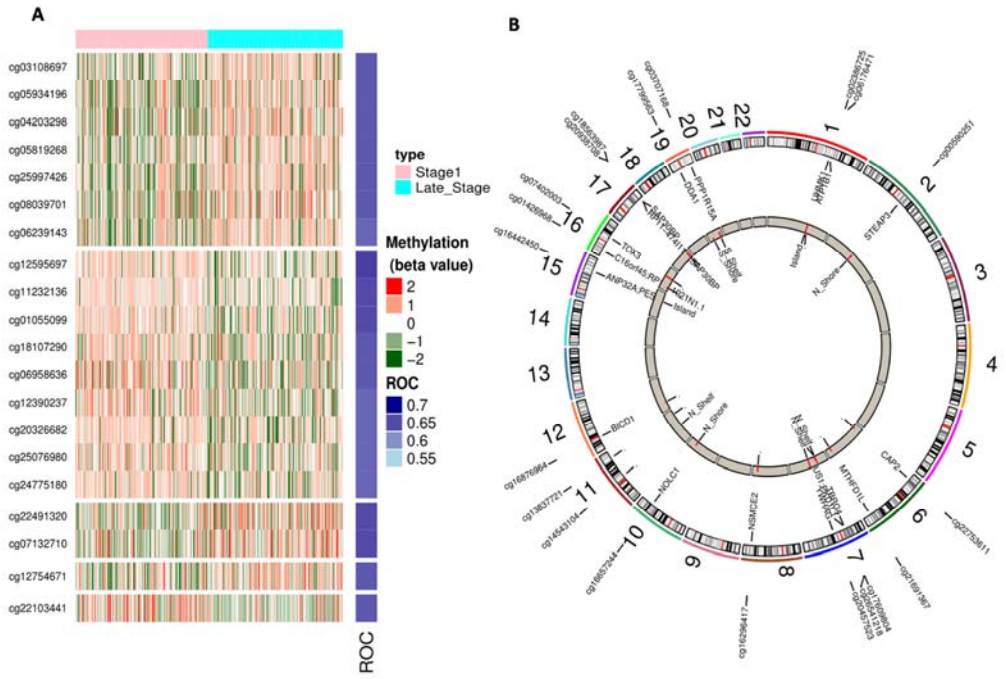


Figure 2: A) Heatmap displaying the differential methylation pattern (with FDR < 0.05) and B) Circos plot representing the chromosome location of top 20 CpG sites (AUROC based) in early versus late stage of LIHC.

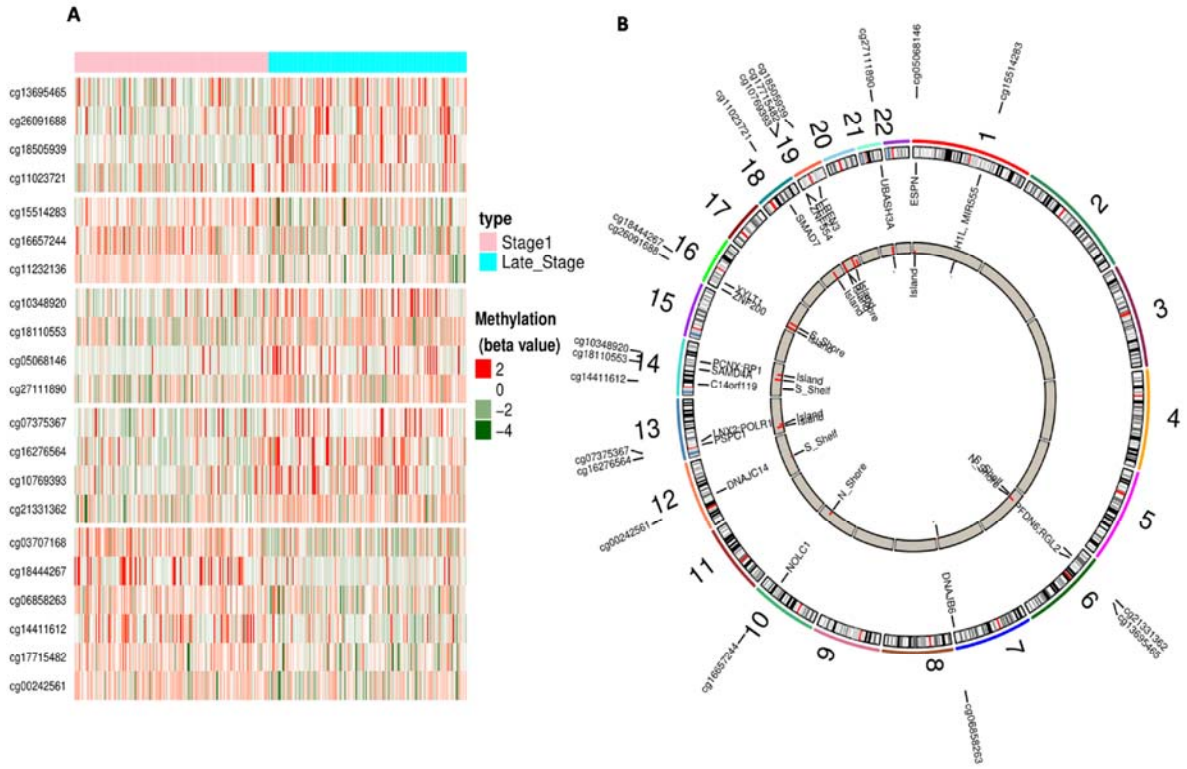


Figure 3: A) Heatmap displaying the differential methylation pattern (with FDR < 0.05) and B) Circos plot representing the chromosome location of top 21 CpG sites (LS-CPG-WEKA) in early versus late stage of LIHC.

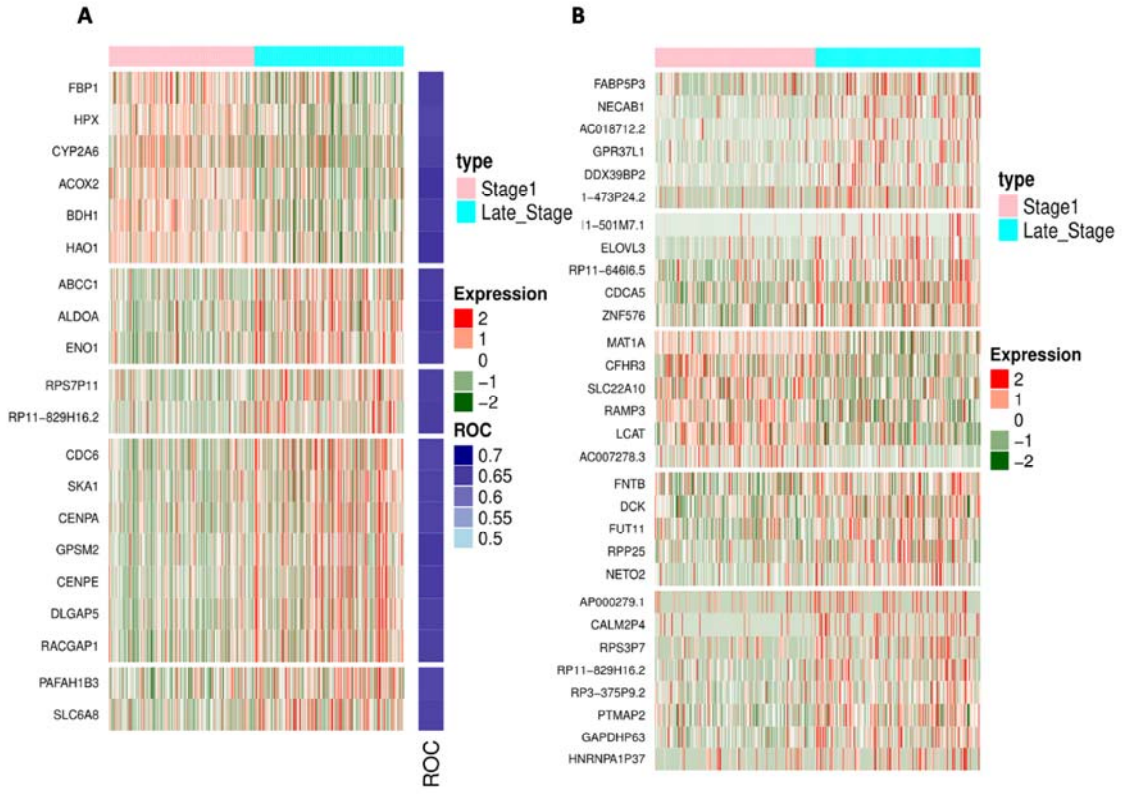


Figure 4: The differential expression pattern of A) top 20 RNA transcripts (AUROC based) and B) 30 RNA transcripts (LS-RNA-WEKA) in early stage versus late stage tissue samples (With FDR <0.01).

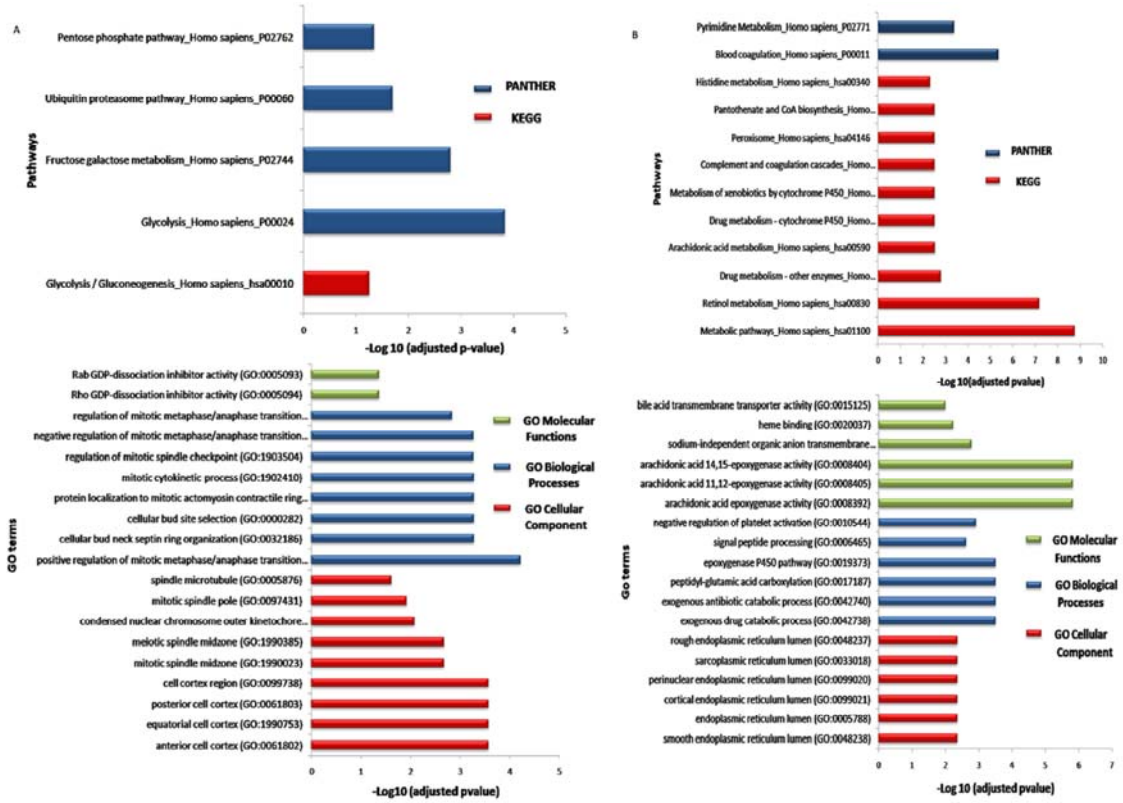
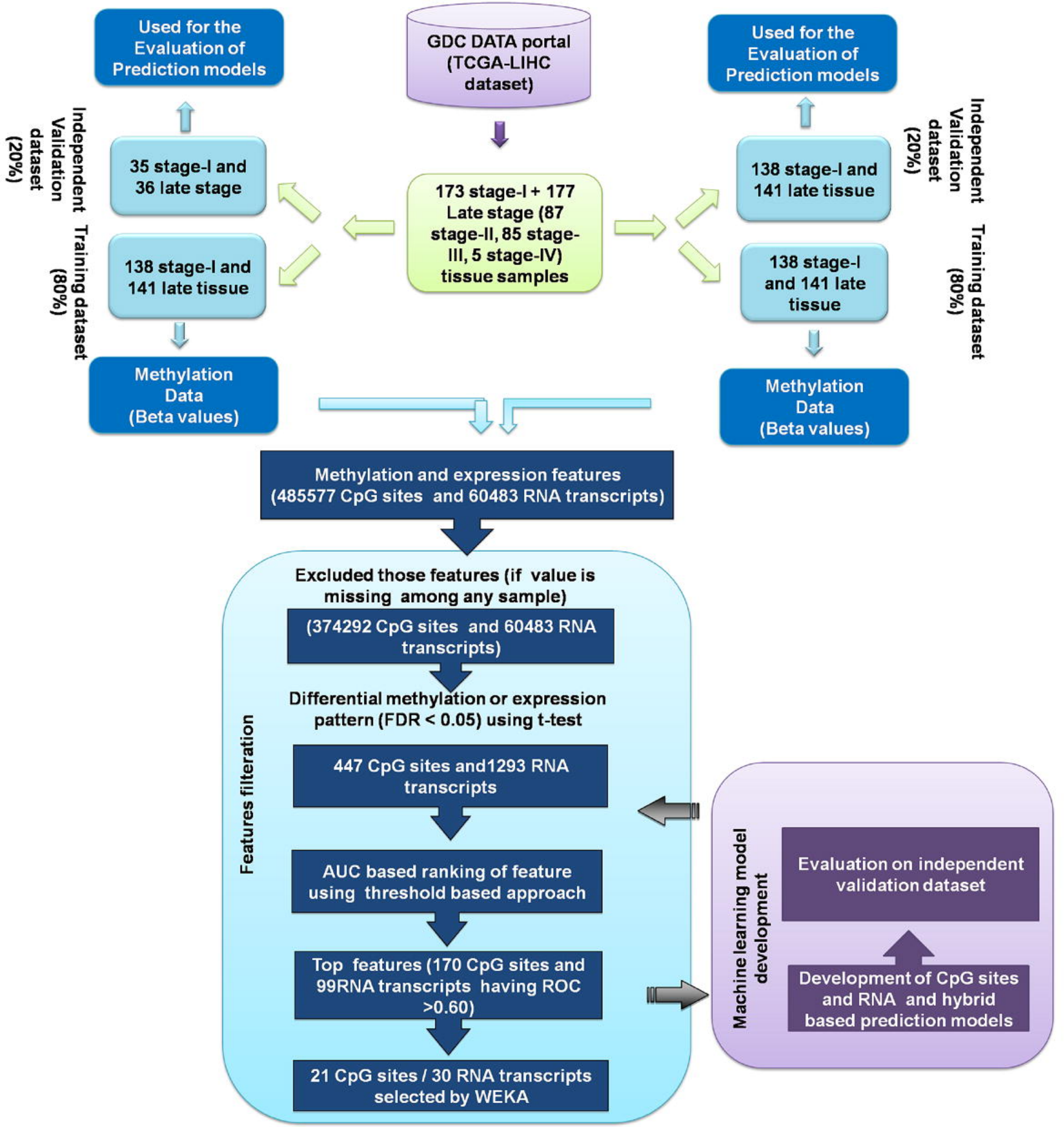
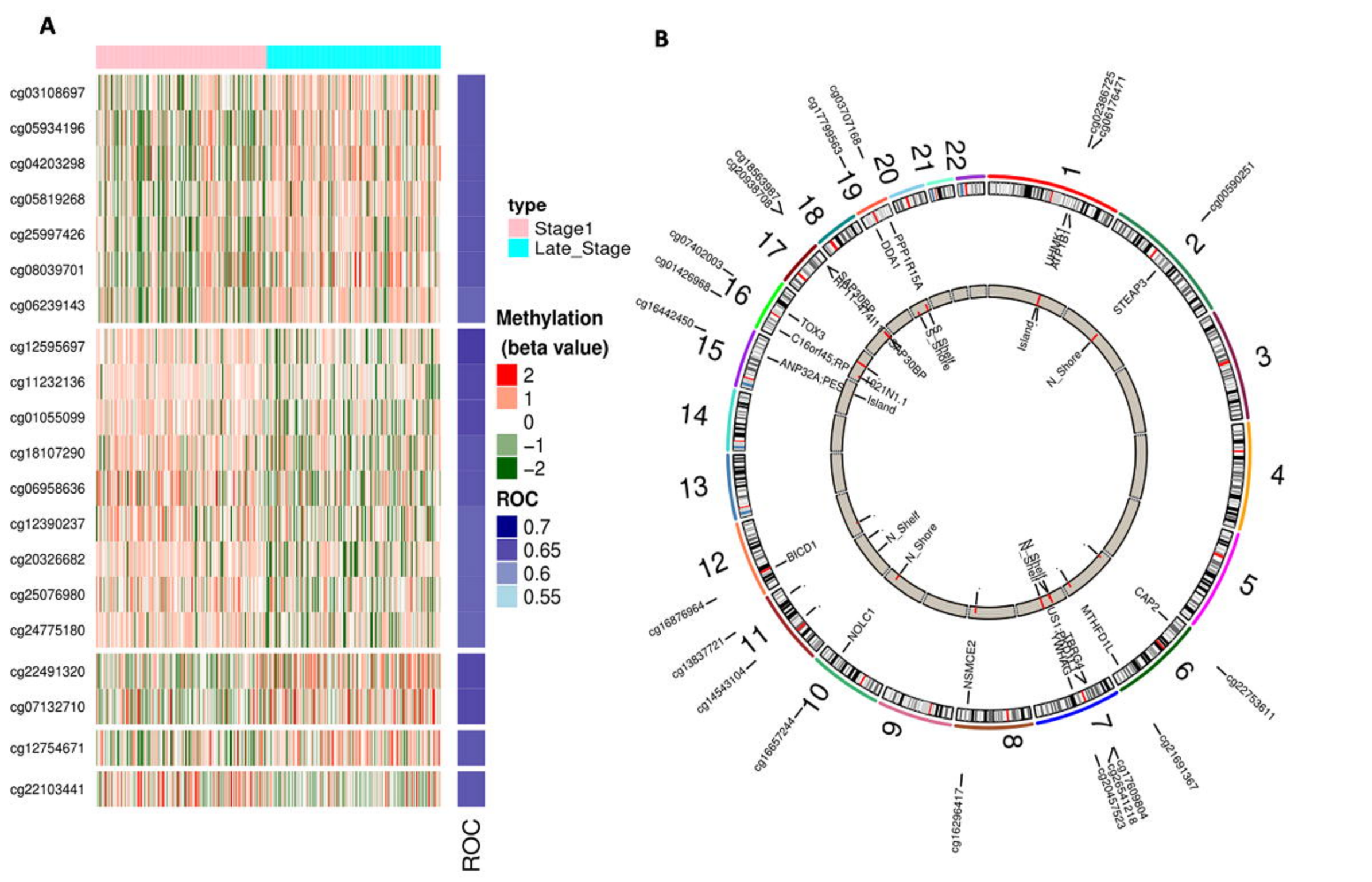
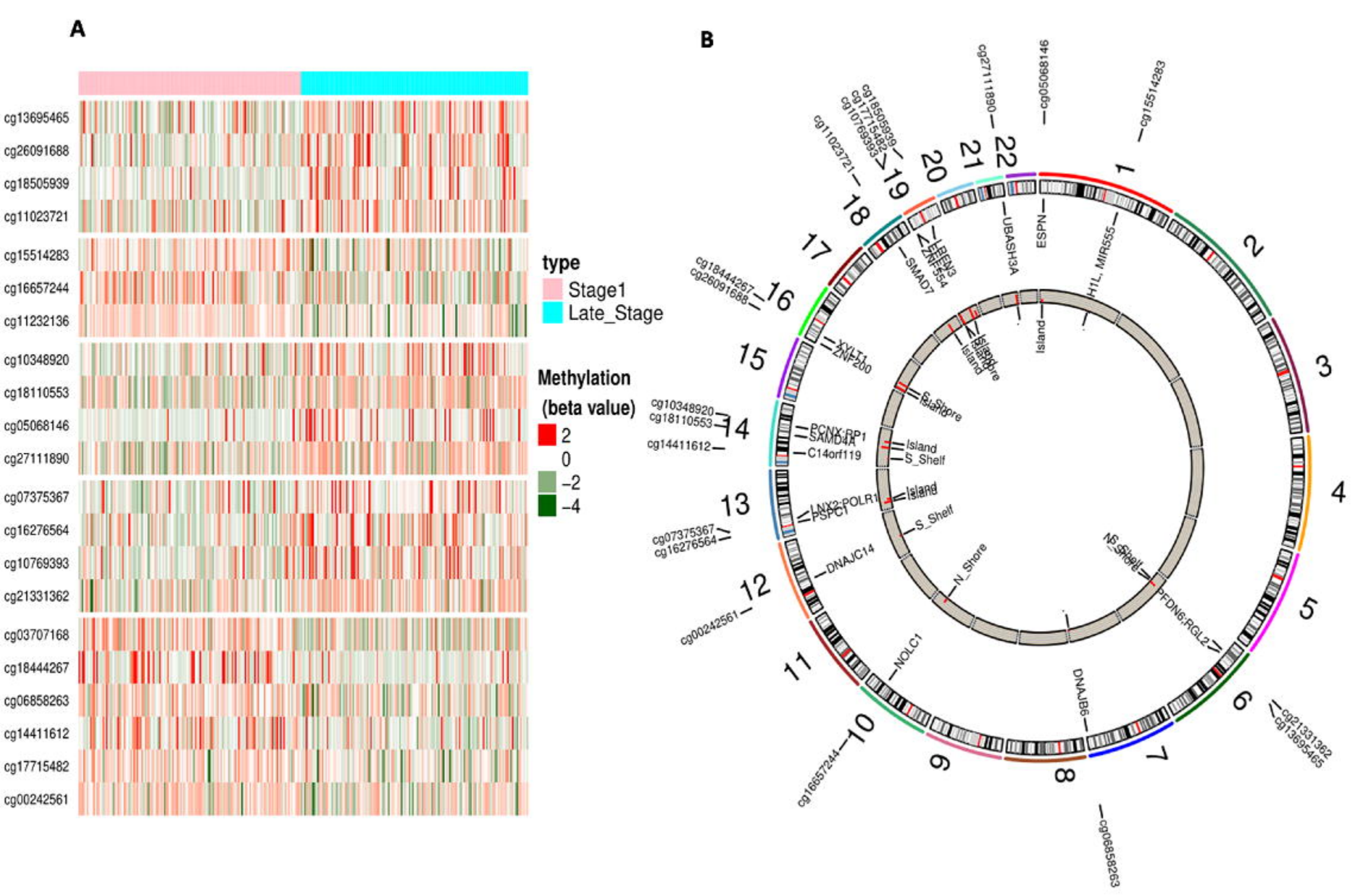
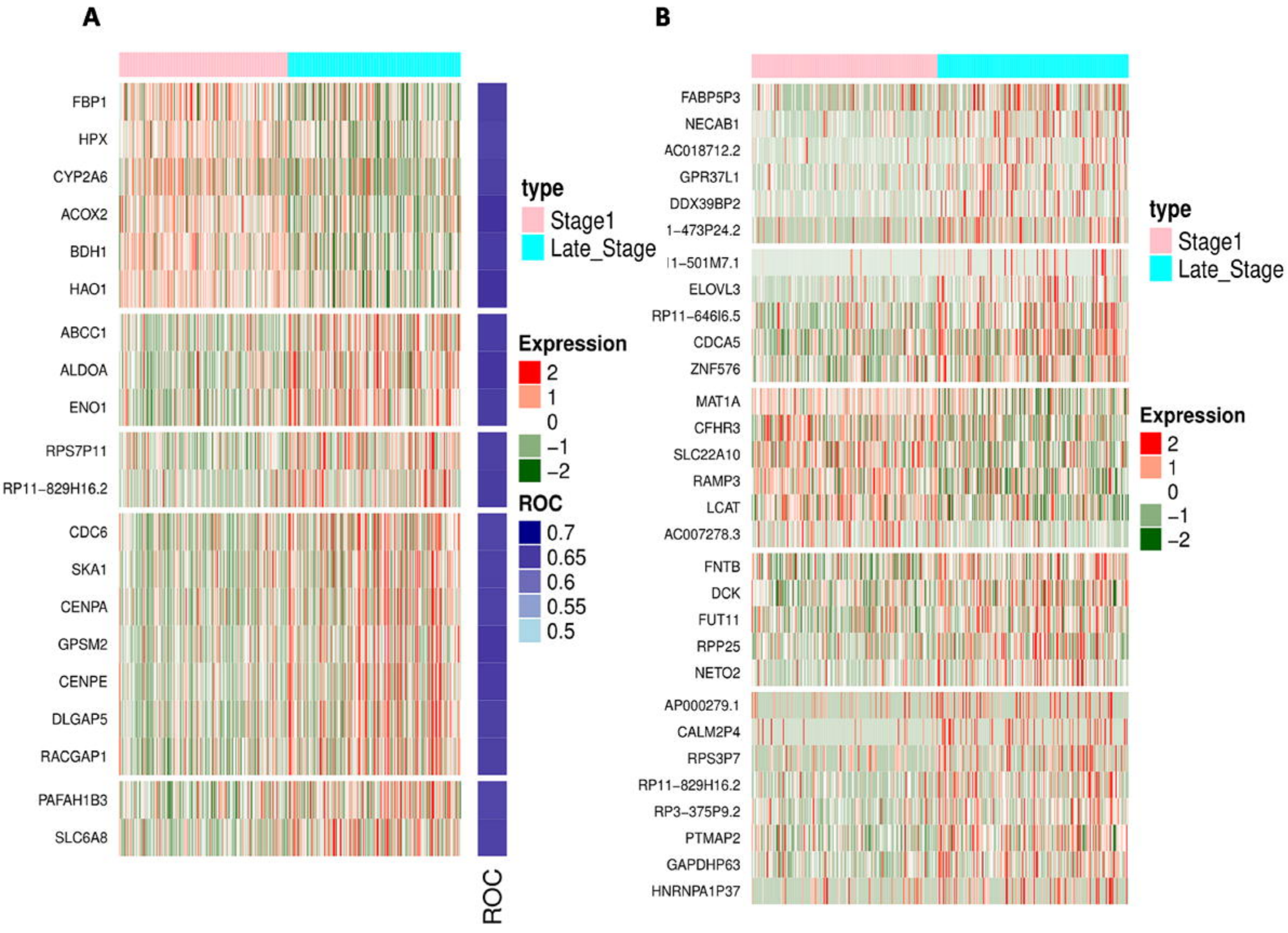


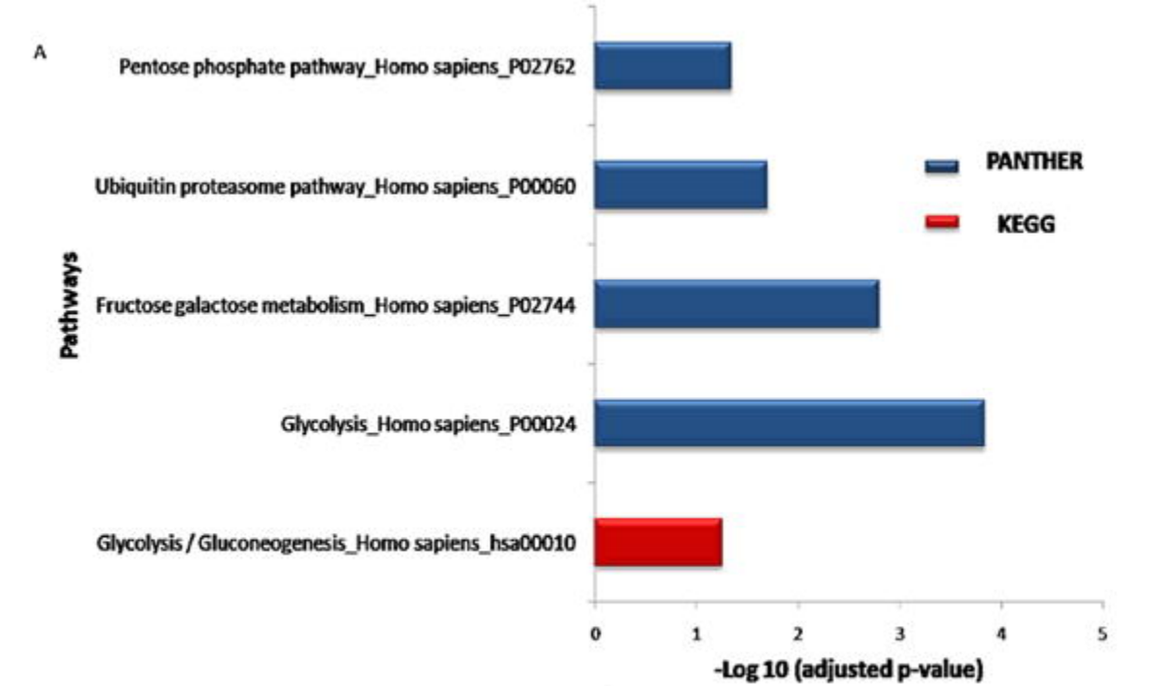
Figure 5: Gene Enrichment analysis of LS-RNA-AUROC, A) 61 downregulated and B) 39 upregulated RNA transcripts in early stage in comparison to late stage of LIHC.

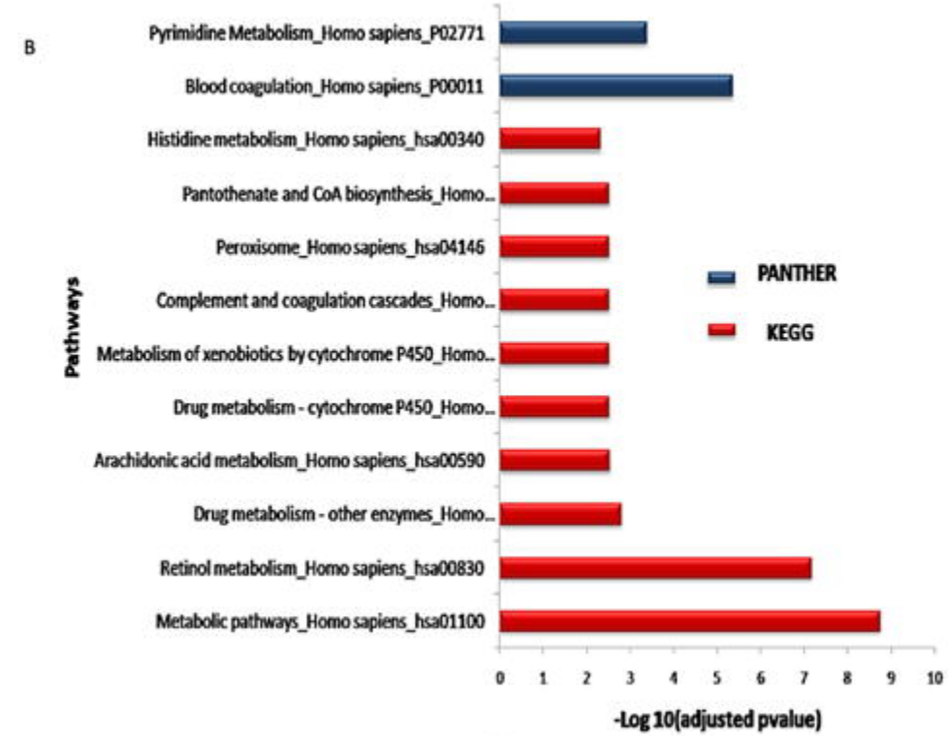

GDC DATA portal (TCGA-LIHC dataset)
173 stage-I + 177 Late stage (87 stage-II, 85 stage-III, 5 stage-IV) tissue samples
Independent Validation dataset (20%)
Training dataset (80%)
35 stage-I and 36 late stage
138 stage-I and 141 late tissue
Used for the Evaluation of Prediction models
Methylation Data (Beta values)
Independent Validation dataset (20%)
Training dataset (80%)
138 stage-I and 141 late tissue
138 stage-I and 141 late tissue
Used for the Evaluation of Prediction models
Methylation Data (Beta values)
Methylation and expression features (485577 CpG sites and 60483 RNA transcripts)
Features filteration
Excluded those features (if value is missing among any sample)
(374292 CpG sites and 60483 RNA transcripts)
Differential methylation or expression pattern (FDR < 0.05) using t-test
447 CpG sites and1293 RNA transcripts
AUC based ranking of feature using threshold based approach
Top features (170 CpG sites and 99RNA transcripts having ROC >0.60)
21 CpG sites / 30 RNA transcripts selected by WEKA
Machine learning model development
Evaluation on independent validation dataset
Development of CpG sites and RNA and hybrid based prediction models

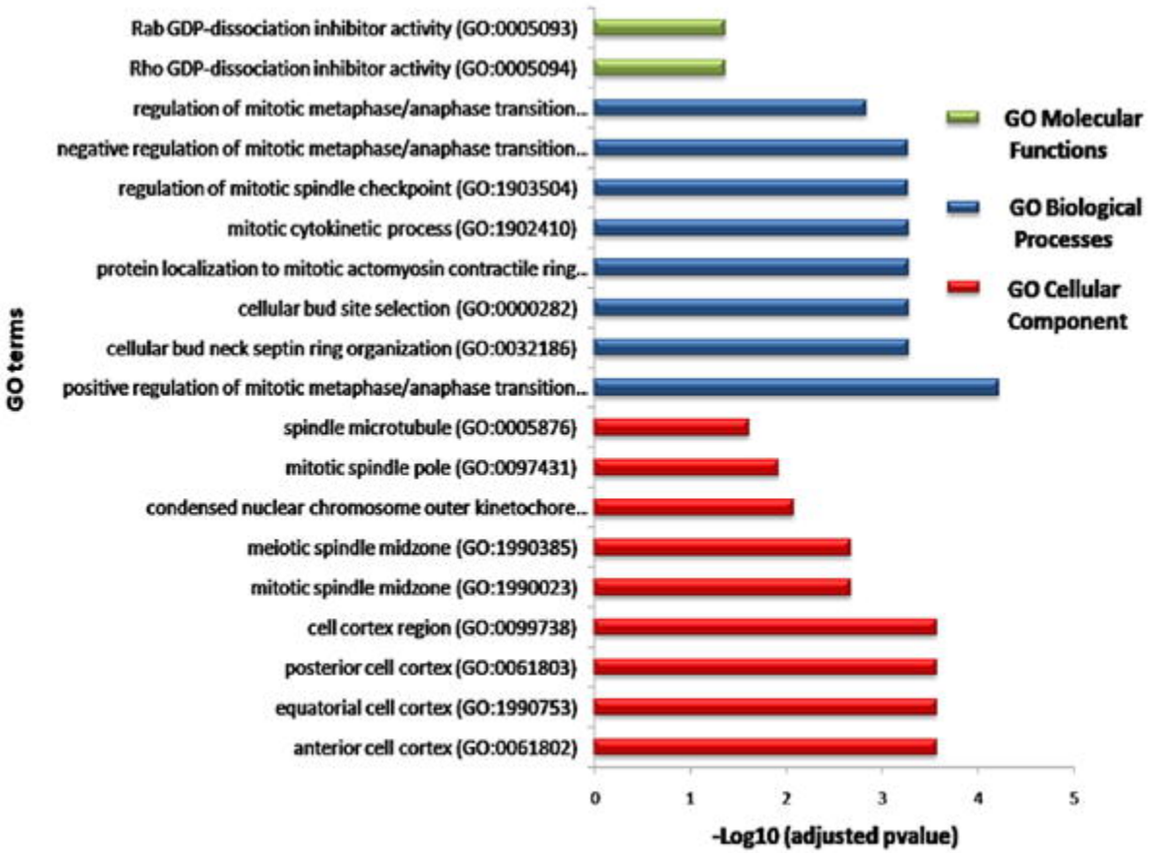

A
cg03108697
cg05934196
cg04203298
cg05819268
cg25997426
cg08039701
cg06239143
cg12595697
cg11232136
cg01055099
cg18107290
cg06958636
cg12390237
cg20326682
cg25076980
cg24775180
cg22491320
cg07132710
cg12754671
cg22103441
ROC
type
Stage1
Late_Stage
Methylation
(beta value)
2
1
0
−1
−2
ROC
0.7
0.65
0.6
0.55
B
cg02386725
cg06176471
cg00590251
STEAP3
CAP2
cg22753611
cg21691367
cg17609804
cg26541218
cg20457523
MTHFD1L
cg16296417
NSMCE2
cg16657244
NOLC1
cg14543104
cg13837721
cg16876964
BICD1
cg01426968
cg07402003
cg16442450
TOX3
C16orf45;RP
ANP32A;PES
cg18563987
cg20938708
cg17799563
cg03707168
DDA1
PPP1R15A
Island
N_Shore
S_Shelf
S_Shore
N_Shelf

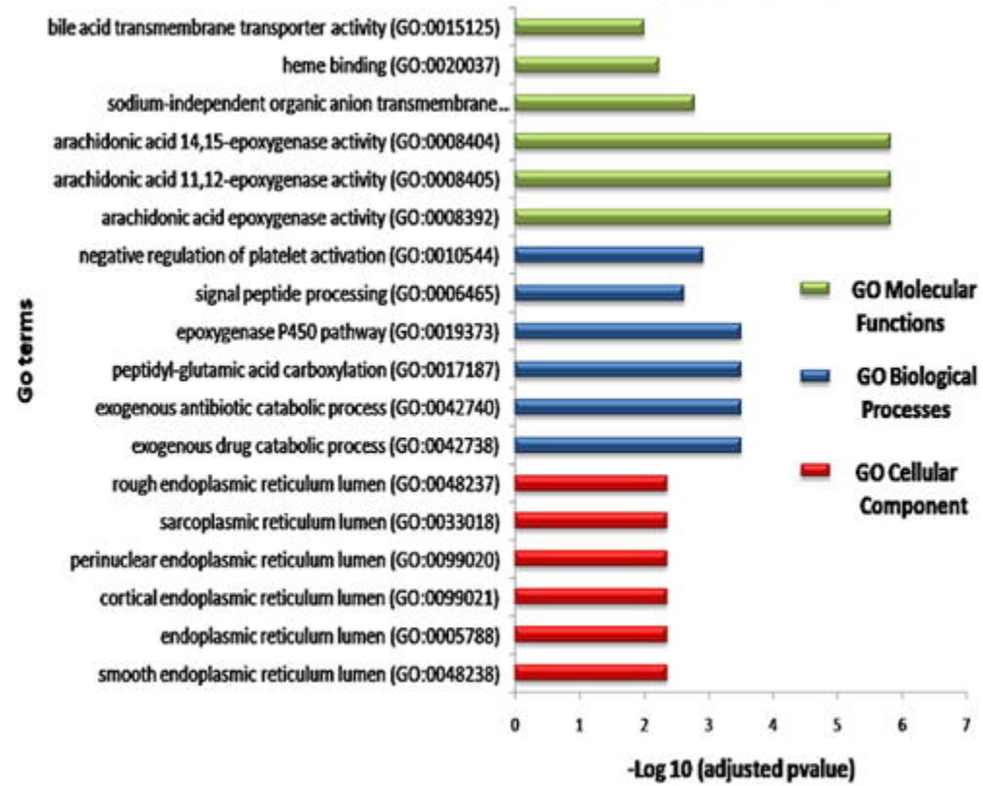

A
cg13695465
cg26091688
cg18505939
cg11023721
cg15514283
cg16657244
cg11232136
cg10348920
cg18110553
cg05068146
cg27111890
cg07375367
cg16276564
cg10769393
cg21331362
cg03707168
cg18444267
cg06858263
cg14411612
cg17715482
cg00242561
type
Stage1
Late_Stage
Methylation
(beta value)
2
0
−2
−4
B
cg27111890
cg05068146
cg15514283
cg18505939
cg17715482
cg10769393
cg11023721
UBASH3A
ESPN
H1L, MIR555
LBEN3
ZNF554
SMAD7
cg18444267
cg26091688
XYLT1
ZNF200
cg10348920
cg18110553
PCNX;RP1
SAMD4A
C14orf119
cg14411612
LNX2;POLR1
PSPC1
cg07375367
cg16276564
DNAJC14
cg00242561
NOLC1
cg16657244
DNAJB6
cg06858263
PFDN6;RGL2
cg21331362
cg13695465
Island
S_Shore
S_Shelf
N_Shore
N_Shelf


A
FBP1
HPX
CYP2A6
ACOX2
BDH1
HAO1
ABCC1
ALDOA
ENO1
RPS7P11
RP11-829H16.2
CDC6
SKA1
CENPA
GPSM2
CENPE
DLGAP5
RACGAP1
PAFAH1B3
SLC6A8
ROC
type
Stage1
Late_Stage
Expression
2
1
0
−1
−2
ROC
0.7
0.65
0.6
0.55
0.5
B
FABP5P3
NECAB1
AC018712.2
GPR37L1
DDX39BP2
1-473P24.2
1-501M7.1
ELOVL3
RP11-646I6.5
CDCA5
ZNF576
MAT1A
CFHR3
SLC22A10
RAMP3
LCAT
AC007278.3
FNTB
DCK
FUT11
RPP25
NETO2
AP000279.1
CALM2P4
RPS3P7
RP11-829H16.2
RP3-375P9.2
PTMAP2
GAPDHP63
HNRNPA1P37
type
Stage1
Late_Stage
Expression
2
1
0
−1
−2


A
Pentose phosphate pathway_Homo sapiens_P02762
Ubiquitin proteasome pathway_Homo sapiens_P00060
Fructose galactose metabolism_Homo sapiens_P02744
Glycolysis_Homo sapiens_P00024
Glycolysis / Gluconeogenesis_Homo sapiens_hsa00010
Pathways
PANTHER
KEGG
0
1
2
3
4
5
-Log 10 (adjusted p-value)


B
Pyrimidine Metabolism_Homo sapiens_P02771
Blood coagulation_Homo sapiens_P00011
Histidine metabolism_Homo sapiens_hsa00340
Pantothenate and CoA biosynthesis_Homo...
Peroxisome_Homo sapiens_hsa04146
Complement and coagulation cascades_Homo...
Metabolism of xenobiotics by cytochrome P450_Homo...
Drug metabolism - cytochrome P450_Homo...
Arachidonic acid metabolism_Homo sapiens_hsa00590
Drug metabolism - other enzymes_Homo...
Retinol metabolism_Homo sapiens_hsa00830
Metabolic pathways_Homo sapiens_hsa01100
Pathways
PANTHER
KEGG
0
1
2
3
4
5
6
7
8
9
10
-Log 10(adjusted pvalue)


Rab GDP-dissociation inhibitor activity (GO:0005093)
Rho GDP-dissociation inhibitor activity (GO:0005094)
regulation of mitotic metaphase/anaphase transition...
negative regulation of mitotic metaphase/anaphase transition...
regulation of mitotic spindle checkpoint (GO:1903504)
mitotic cytokinetic process (GO:1902410)
protein localization to mitotic actomyosin contractile ring...
cellular bud site selection (GO:0000282)
cellular bud neck septin ring organization (GO:0032186)
positive regulation of mitotic metaphase/anaphase transition...
spindle microtubule (GO:0005876)
mitotic spindle pole (GO:0097431)
condensed nuclear chromosome outer kinetochore...
meiotic spindle midzone (GO:1990385)
mitotic spindle midzone (GO:1990023)
cell cortex region (GO:0099738)
posterior cell cortex (GO:0061803)
equatorial cell cortex (GO:1990753)
anterior cell cortex (GO:0061802)
GO terms
GO Molecular Functions
GO Biological Processes
GO Cellular Component
0
1
2
3
4
5
-Log10 (adjusted pvalue)


bile acid transmembrane transporter activity (GO:0015125)
heme binding (GO:0020037)
sodium-independent organic anion transmembrane...
arachidonic acid 14,15-epoxygenase activity (GO:0008404)
arachidonic acid 11,12-epoxygenase activity (GO:0008405)
arachidonic acid epoxygenase activity (GO:0008392)
negative regulation of platelet activation (GO:0010544)
signal peptide processing (GO:0006465)
epoxygenase P450 pathway (GO:0019373)
peptidyl-glutamic acid carboxylation (GO:0017187)
exogenous antibiotic catabolic process (GO:0042740)
exogenous drug catabolic process (GO:0042738)
rough endoplasmic reticulum lumen (GO:0048237)
sarcoplasmic reticulum lumen (GO:0033018)
perinuclear endoplasmic reticulum lumen (GO:0099020)
cortical endoplasmic reticulum lumen (GO:0099021)
endoplasmic reticulum lumen (GO:0005788)
smooth endoplasmic reticulum lumen (GO:0048238)
Go terms
GO Molecular Functions
GO Biological Processes
GO Cellular Component
0
1
2
3
4
5
6
7
-Log 10 (adjusted pvalue)