Sarathi and Palaniappan *BMC Cancer* (2019) 19:663
https://doi.org/10.1186/s12885-019-5838-3

BMC Cancer

RESEARCH ARTICLE Open Access

Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma

Arjun Sarathi[1] and Ashok Palaniappan[2*]

Abstract

Background: Liver cancer is among top deadly cancers worldwide with a very poor prognosis, and the liver is a vulnerable site for metastases of other cancers. Early diagnosis is crucial for treatment of the predominant liver cancers, namely hepatocellular carcinoma (HCC). Here we developed a novel computational framework for the stage-specific analysis of HCC.

Methods: Using publicly available clinical and RNA-Seq data of cancer samples and controls and the AJCC staging system, we performed a linear modelling analysis of gene expression across all stages and found significant genome-wide changes in the log fold-change of gene expression in cancer samples relative to control. To identify genes that were stage-specific controlling for confounding differential expression in other stages, we developed a set of six pairwise contrasts between the stages and enforced a *p*-value threshold (< 0.05) for each such contrast. Genes were specific for a stage if they passed all the significance filters for that stage. The monotonicity of gene expression with cancer progression was analyzed with a linear model using the cancer stage as a numeric variable.

Results: Our analysis yielded two stage-I specific genes (CA9, WNT7B), two stage-II specific genes (APOBEC3B, FAM186A), ten stage-III specific genes including DLG5, PARI, NCAPG2, GNMT and XRCC2, and 35 stage-IV specific genes including GABRD, PGAM2, PECAM1 and CXCR2P1. Overexpression of DLG5 was found to be tumor-promoting contrary to the cancer literature on this gene. Further, GABRD was found to be signifincantly monotonically upregulated across stages. Our work has revealed 1977 genes with significant monotonic patterns of expression across cancer stages. NDUFA4L2, CRHBP and PIGU were top genes with monotonic changes of expression across cancer stages that could represent promising targets for therapy. Comparison with gene signatures from the BCLC staging system identified two genes, HSP90AB1 and ARHGAP42. Gene set enrichment analysis indicated overrepresented pathways specific to each stage, notably viral infection pathways in HCC initiation.

Conclusions: Our study identified novel significant stage-specific differentially expressed genes which could enhance our understanding of the molecular determinants of hepatocellular carcinoma progression. Our findings could serve as biomarkers that potentially underpin diagnosis as well as pinpoint therapeutic targets.

Keywords: LIHC transcriptomics, HCC stages, Stage-specific biomarkers, Differentially expressed genes, Pairwise contrasts, Significance analysis, Linear modelling, Tumorigenesis, Cancer progression, Metastasis, Monotonic expression

* Correspondence: apalania@scbt.sastra.edu
[2]Department of Bioinformatics, School of Chemical and BioTechnology, SASTRA deemed University, Thanjavur, Tamil Nadu 613401, India
Full list of author information is available at the end of the article

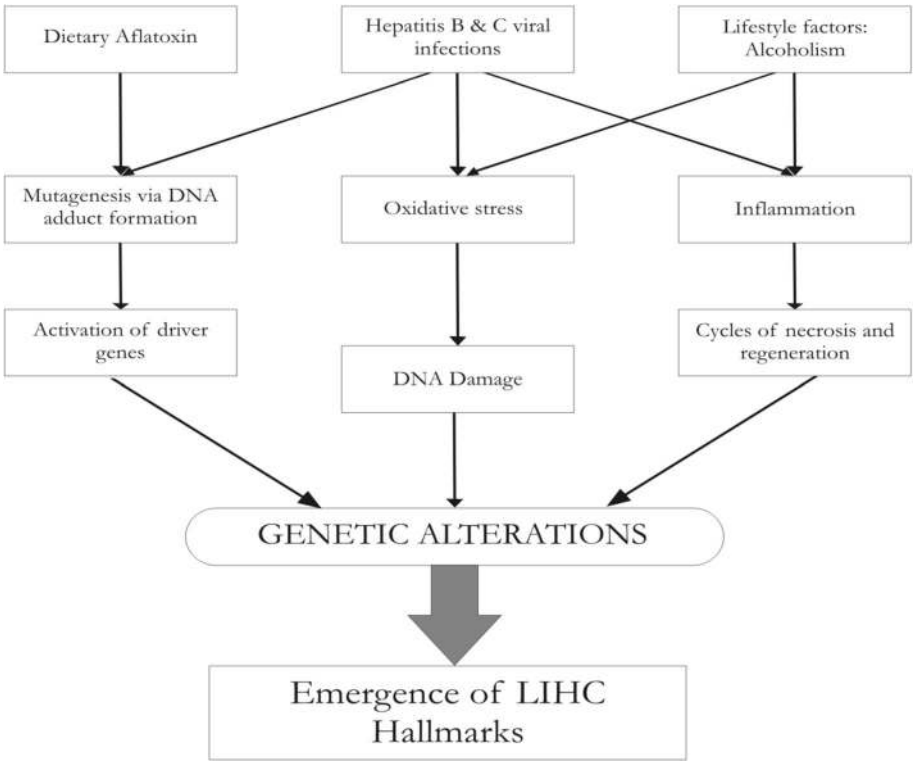

Fig. 1 Major causative pathways of hepatocarcinogenesis. All pathways converge to progressive genomic alterations, leading a normal cell to acquire the hallmarks of cancer

Background

Liver cancer is the second most deadly cancer in terms of mortality rate, with a very poor prognosis [60]. It accounted for 9.1% of all cancer deaths, and 83% of the annual new estimated 782,000 liver cancer cases worldwide occur in developing countries [13]. Liver cancer showed the greatest increase in mortality in the last decade for both males (53%) and females (59%) [8]. Liver hepatocellular carcinoma (LIHC) or simply hepatocellular carcinoma (HCC) is the most common type of liver cancer, accounting for nearly 85% of liver cancers. 78% of all reported cases of HCC were due to viral infections (53% Hepatitis B virus and 25% Hepatitis C virus) [38]. There are several non-viral causes of HCC as well, mainly aflatoxins and alcohol [10]. As shown in Fig. 1, all the factors converge to a common mechanism of genetic alterations that lead to the acquisition of cancer hallmarks [20] and the eventual emergence of a cancer cell [11]. Genetic alterations constitute the heart of the problem, and studying changes due to these genetic

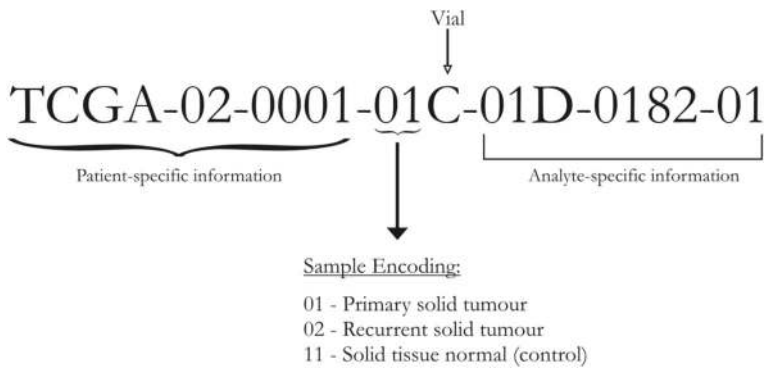

Fig. 2 TCGA 'Hybridization REF' Barcode. The first 10 characters constitute an anonymized unique patient identifier and the following two characters denote whether the sample is tumor or normal

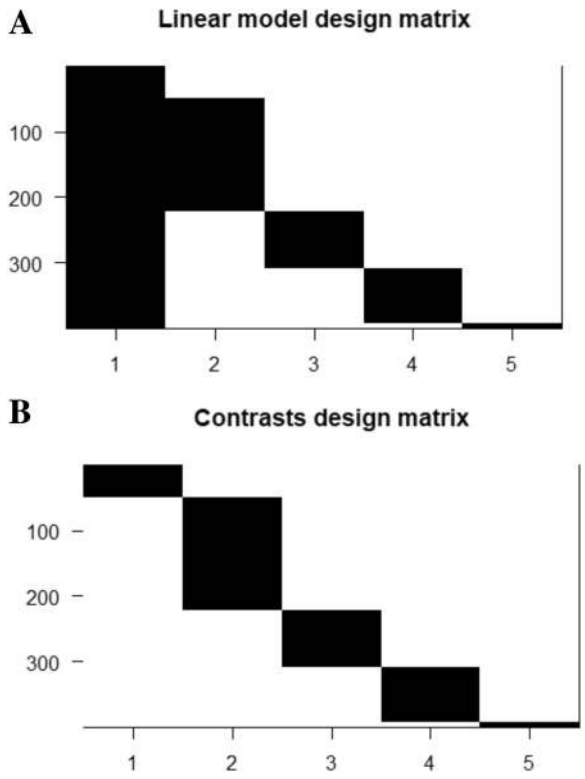

Fig. 3 Design matrices. **a** In the linear modeling, the control samples served as the baseline expression (intercept) of each gene against which the stage-specific expression was estimated. **b** the design matrix for the contrasts analysis

alterations is paramount to understand HCC. Earlier gene expression studies using EST data detected differential expression in cancer tissue compared to non-cancerous liver and proposed the existence of genetic aberrations and changes in transcriptional regulation in HCC [58]. The Cancer Genome Atlas (TCGA) research network [41] have subtyped and identified many potential targets for HCC based on a comprehensive multi-omics analysis. An independent analysis of TCGA RNA-Seq data encompassing 12 cancer tissues has uncovered liver cancer-specific genes [37]. Zhang et al. [63] have performed mutation analysis of HCC, and Yang et al. [59] combined TCGA expression data and natural language processing techniques to identify cancer-specific markers.

The burden of disease and mortality rate are both inversely correlated with the cancer stage. The response rate to therapy is also inversely correlated with stage. To the best of our knowledge, there are no reported research in the literature that have dissected the stage-specific features of HCC. The cancer staging system is based on gross features of cancer anatomical penetration, and one such standard is the American Joint Committee on Cancer (AJCC) Tumor-Node-Metastasis (TNM) staging [2]. It is reasonable to hypothesize that the stage-specific gross changes are associated with signature molecular events, and try to probe such molecular bases of stage-wise progression of cancer. We had earlier published on stage-specific "hub driver" genes in colorectal cancer [36]. A stage-focussed analysis of colorectal cancer transcriptome data yielded negative results vis-a-vis the AJCC staging system [25].

Methods

Data preprocessing

Normalized and $\log_2$-transformed Illumina HiSeq RNA-Seq gene expression data processed by the RSEM pipeline [29] were obtained from TCGA via the firebrowse.org portal [6]. The patient barcode (uuid) of each sample encoded in the variable called 'Hybridization REF' was parsed and used to annotate the controls and cancer samples (Fig. 2). To annotate the stage information of the cancer samples, we obtained the clinical information dataset for HCC from firebrowse.org (LIHC.Merge_

Table 1 Contrast matrix with control. Each stage (indicated by '1') is contrasted against the control (indicated by '-1') in turn

	Stage1-control	Stage2-control	Stage3-control	Stage4-control
Control	-1	-1	-1	-1
Stage1	1	0	0	0
Stage2	0	1	0	0
Stage3	0	0	1	0
Stage4	0	0	0	1

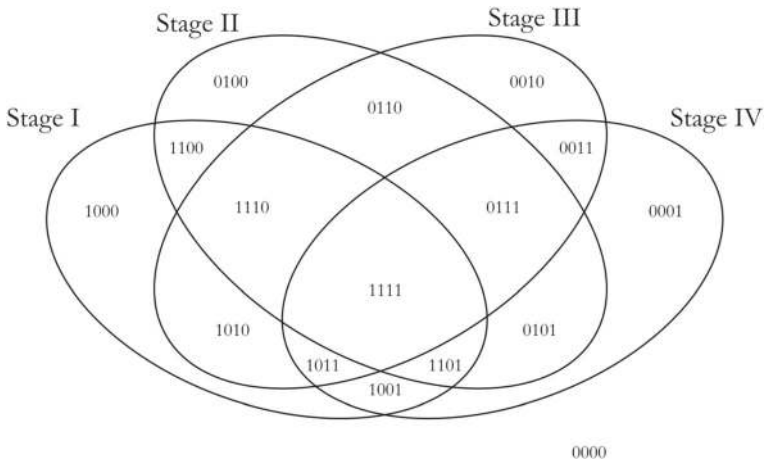

Fig. 4 A Venn representation of the pairwise stages contrasts. A gene could be differentially expressed in any combination of the four stages and this could be represented by a 4-bit string, one bit for each stage. For e.g., '1111' at the overlap of all four stages would be assigned to genes that are differentially expressed in all four stages

Clinical.Level_1.2016012800.0.0.tar.gz) and merged the clinical data with the expression data by matching the "Hybridization REF" in the expression data with the aliquot barcode identifier in the clinical data. The stage information of each patient was encoded in the clinical variable "pathologic stage". The pathologic stage is essentially the surgical stage, prior to any treatment received, determined with the tissue obtained at the time of surgery. This interpretation is reinforced in the TCGA HCC sample inclusion criteria as follows: "Surgical resection of biopsy biospecimens were collected from patients diagnosed with hepatocellular carcinoma (HCC), and had not received prior treatment for their disease (ablation, chemotherapy, or radiotherapy)" (The TCGA [41]). The availability of this unequivocal information enables the analysis of cancer stages. The substages (A,B,C) were collapsed into the parent stage, resulting in four stages of interest (I, II, III, IV). We retained a handful of other clinical variables pertaining to demographic features, namely age, sex, height, weight, and vital status. With this merged dataset, we filtered out genes that showed little change in expression across all samples (defined as $\sigma < 1$). Finally, we removed cancer samples from our analysis that were missing stage annotation (value 'NA' in the "pathologic stage"). The data preprocessing was done using R (www.r-project.org).

Table 2 Contrast matrix for inter-stage contrasts. There are six possible pairwise contrasts between the stages that are essential to identifying stage-specific genes

	Stage2-Stage1	Stage3-Stage1	Stage3-Stage2	Stage4-Stage1	Stage4-Stage2	Stage4-Stage3
Control	0	0	0	0	0	0
Stage1	-1	-1	0	-1	0	0
Stage2	1	0	-1	0	-1	0
Stage3	0	1	1	0	0	-1
Stage4	0	0	0	1	1	1

Table 3 AJCC Cancer staging. The correspondence between the AJCC staging and the TCGA staging for LIHC is noted, along with the number of LIHC cases in each stage in the TCGA dataset. Control indicates the number of normal tissue control samples, and NA denotes cases where the stage information is unavailable

TCGA Stage	TNM classification	Cases	
1A	T1a N0 M0	172	
1B	T1b N0 M0		
2	T2 N0 M0	87	
3A	T3 N0 M0	65	
3B	T4 N0 M0	8	
3C	-	9	85
3	-	3	
4A	T(any) N1 M0	1	
4B	T(any) N(any) M1	2	5
4	-	2	
CONTROL	-	50	
NA	-	24	

Linear modelling

Linear modelling of expression across cancer stages relative to the baseline expression (i.e, in normal tissue controls) was performed for each gene using the R *limma* package [42]. The following linear model was fit for each gene's expression based on the design matrix shown in Fig. 3a:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad (1)$$

where the independent variables are indicator variables of the sample's stage, the intercept α is the baseline expression estimated from the controls, and β_i are the estimated stage-wise log fold-change (lfc) coefficients relative to controls. The linear model was subjected to empirical Bayes adjustment to obtain moderated t-statistics [34]. To account for multiple hypothesis testing and the false discovery rate, the *p*-values of the F-statistic of the linear fit were adjusted using the method od Hochberg and Benjamini [22]. The linear trend across cancer stages for the top significant genes were visualized using boxplots to ascertain the regulation status of the gene relative to the control.

Monotonic mean expression

The linear model in eqn. (1) would not be sufficient to identify genes with an *ordered* monotonic trend of expression across cancer stages. Addressing this question would also help assess whether monotonic changes of gene expression were observed with disease progression. Towards this end, we designed a model of gene expression where the cancer stage was treated as a numeric variable:

$$y = aX + b \quad (2)$$

where X takes a value in [0,1,2,3,4] corresponding to the sample stage: [control, I, II, III, IV], respectively. It

Table 4 Summary of key demographic features of the dataset. For continuous variables (age,height, weight and BMI), the mean ± standard deviation is given. BMI is calculated only for patients with both height and weight data

Characteristic		Control	Stage1	Stage2	Stage3	Stage4	NA	Overall
Number of samples		50	172	87	85	5	24	423
Age (Years)		61.7±16.1	60.6±12.2	59.0±13.3	56.2±14.8	42.8±20.7	68.1±10.7	59.7±13.2
Height (cm)		170.6±9.5	166.5±12.3	167.9±8.3	169.0±8.9	162.3±4.9	166.2±11.1	167.7±10.6
Weight (Kg)		76.1±22.1	73.2±19.8	73.3±18.9	69.9±18.8	72.3±21.5	79.7±19.8	73.2±19.7
BMI		26.2±7.8	26.7±10.2	26.0±5.8	24.3±6.0	27.7±9.3	29.1±7.6	26.1±8.4
Sex	Male	28	122	60	55	1	14	280
	Female	22	50	27	30	4	10	143
Vital Status	Alive	20	134	71	65	2	12	304
	Dead	30	38	16	20	3	12	119

Table 5 Top 10 genes of the linear model. The log-fold change expression of the gene in each stage relative to the controls are given, followed by p-value adjusted for the false discovery rate, and the regulation status of the gene in the cancer stages with respect to the control

Genes	Stage I lfc (β_1)	Stage II lfc (β_2)	Stage III lfc (β_3)	Stage IV lfc (β_4)	Adj. p value	Regulation status
GABRD	5.08	5.11	5.24	6.55	5.529e-78	Up-regulated
PLVAP	3.51	3.24	3.24	3.79	7.498e-75	Up-regulated
CLEC4M	-8.32	-8.67	-8.48	-9.24	6.058e-74	Down-regulated
CXORF36	2.91	2.86	2.76	3.44	5.376e-73	Up-regulated
CLEC1B	-7.85	-8.46	-8.05	-9.44	6.292e-71	Down-regulated
BMP10	-4.66	-4.75	-4.67	-5.25	1.447e-66	Down-regulated
CLEC4G	-7.75	-8.23	-7.95	-8.75	2.437e-66	Down-regulated
CDH13	3.30	3.34	3.32	3.86	3.454e-66	Up-regulated
UBE2T	3.85	4.50	4.47	3.76	2.544e-65	Up-regulated
SLC26A6	3.10	3.39	3.34	3.07	7.438e-65	Up-regulated

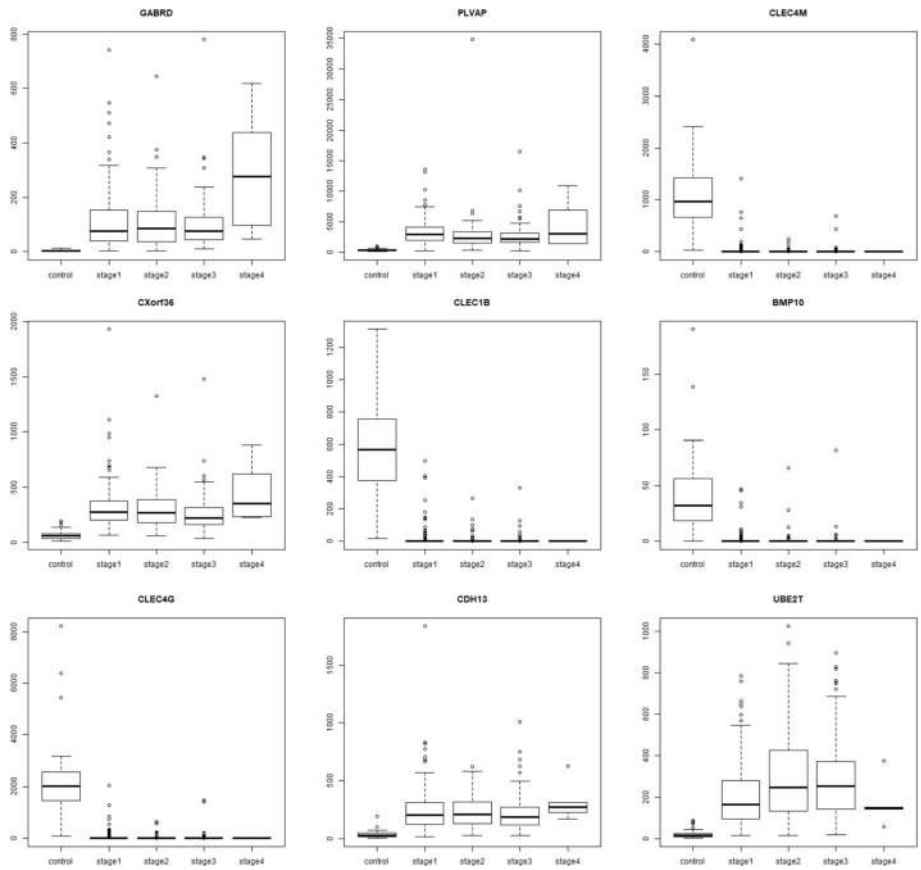

Fig. 5 Boxplots of top 9 linear model genes. For each gene, notice that the trend in expression could be either overexpression or downregulation relative to the control. For e.g., GABRD, PLVAP, CXorf36, CDH13 and UBE2T are overexpressed, while CLEC4M, CLEC1B, BMP10, and CLEC4G are downregulated. It could be seen that a linear trend does not imply maximal |lfc| in stage 4, as illustrated most clearly in the case of UBE2T

was noted the mean expression of a gene could show the following monotonic patterns across cancer stages:

(i). monotonic upregulation, where mean expression follows: control < I < II < III < IV.
(ii). monotonic downregulation, where mean expression follows: control > I > II > III > IV.

The sets of genes conforming to either (i) or (ii) were identified to yield monotonically upregulated and monotonically downregulated genes. These two sets were merged, and the final set of genes with monotonic changes of expression with cancer progression was obtained. This final set was ranked by the adj. *p*-values from the model estimated by eqn. (2).

Pairwise contrasts

To perform contrasts, a slightly modified design matrix shown in Fig. 3b was used, which would give rise to the following linear model of expression for each gene:

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad (3)$$

where the controls themselves are one of the indicator variables, and the β_i are all coefficients estimated only from the corresponding samples. Our first contrast of interest, between each stage and the control, was achieved using the contrast matrix shown in Table 1. Four contrasts were obtained, one for each stage vs control. A threshold of $|lfc| > 2$ was applied to each such contrast to identify differentially expressed genes (with respect to the control). We used the absolute value of the lfc, since driver genes could be either upregulated or downregulated. Genes could be differentially expressed in any combination of the stages or no stage at all. To analyze the pattern of differential expression (with respect to the control), we constructed a four-bit binary string for each gene, where each bit signified whether the gene was differentially expressed in the corresponding stage. For example, the string '1100' indicates that the gene was differentially expressed in the first and second stages. There are $2^4 = 16$ possible outcomes of the four-bit string for a given gene corresponding

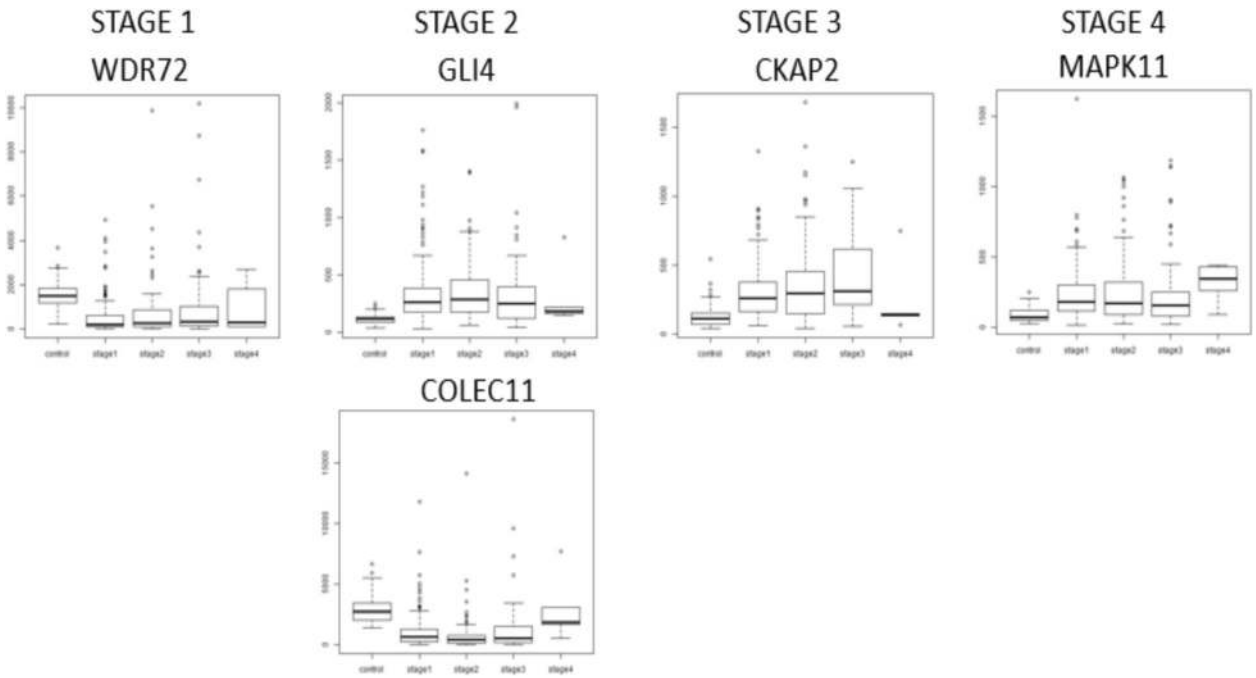

Fig. 6 Boxplots illustrating stage-specificity of differentially expressed genes. Extremal expression in a stage could be either maximal expression or minimal expression relative to the control and all other stages, and could be termed maximal *differential* expression. Here we show genes with maximal differential expression in stage-I (WDR72; minimum expression), stage-II (GLI4, maximum expression; COLEC11, minimum expression), stage-III (CKAP2; maximum expression), and stage-IV (MAPK11; maximum expression)

to the combination of stages in which it is differentially expressed. This is illustrated in set-theoretic terms in Fig. 4. In our first elimination, we removed genes whose $|lfc| < 2$ for all stages. For each remaining gene, we identified the stage that showed the highest |lfc| and assigned the gene as specific to that stage for the rest of our analysis.

Significance analysis

We applied a four-pronged criteria to establish the significance of the stage-specific differentially expressed genes.

(i). Adj. *p*-value of the contrast with respect to the control < 0.001. The expression profile of a driver gene in cancer samples would markedly depart from that for the controls, which motivates the use of a stringent threshold here.
(ii). (ii)-(iv) *P*-value of the contrast with respect to other stages < 0.05. The use of a more relaxed cutoff would improve the sensitivity of stage-specific detection.

To obtain the above *p*-values (ii) - (iv), we used the contrast matrix shown in Table 2, which was then used an an argument to the `contrastsFit` function in *limma*.

Further analyses

Principal component analysis (PCA) were performed using `prcomp` in R. To choose 100 random genes, we used the `rand` function. Gene set enrichment analysis were performed on KEGG (https://www.genome.jp/kegg/) and Gene Ontology [5] using `kegga` and `goana` in *limma*, respectively. In order to visualize outlier genes that are significant with a large effect size, volcano plots could be obtained by plotting the -log10 transformed *p*-value vs. the log fold-change of gene expression. Heat maps of significant stage-specific differentially expressed genes were visualized using `heatmap` and clustered using `hclust`. Novelty of the identified stage-specific genes was ascertained by screening against the Cancer Gene Census v84 [14].

Results

The TCGA expression data consisted of expression values of 20,532 genes in 423 samples. After the completion of data pre-processing, we obtained a final dataset of expression data for 18,590 genes across 399 samples annotated with the corresponding sample stage (available in Supplementary File S1). The stagewise distribution of TCGA samples along with the corresponding AJCC staging is shown in Table 3. A statistical summary of demographic details including age, sex, height, weight, and vital status is shown in Table 4. The body mass index (BMI) distribution was derived from patient clinical data that had both height and weight (i.e, neither was 'NA'). The average age of

onset of HCC was around 60 years, and the average BMI was about 26, indicating a possible link with ageing-associated pathology and obesity.

The dataset was processed through `voom` in *limma* to prepare for linear modelling [28]. At a *p*-value cutoff of 0.05, 14,843 genes were significant for the linear model given by eqn. (1). Even raising the bar to 1E-5, 9618 genes remained significant in the linear modelling, thus implying a strong linear trend in their expression across cancer stages relative to control. This was not entirely surprising since one of the hallmarks of cancer phenotype is genome-wide instability [20]. The linear modelling highlighted top ranked genes, some upregulated in HCC (GABRD, PLVAP, CDH13) and some downregulated (CLEC4M, CLEC1B, CLEC4G). The lfc for each stage with respect to control of top ten genes (ranked by adjusted *p*-value) are shown in Table 5, along with their inferred regulation status. Boxplots of the expression of the top 9 genes (Fig. 5) indicated elevated expression across cancer stages relative to control for up-regulated genes, while depressed expression across cancer stages relative to control was indicative of downregulated genes. (Boxplots of all other genes in the top 200 are provided in the Supplementary Fig. S1) It is worthwhile to note that a given gene might have maximal differential expression in any stage (not necessarily stage 4), and the linear trend does not suggest the order of expression across stages (Fig. 6).

A PCA of the top 100 genes from the linear model was visualized using the top two principal components (Fig. 7a). A clear separation of the controls and the cancer samples could be seen, suggesting the extent of differential expression of these genes in cancer samples. Hence linear modelling yields cancer-specific genes versus normal controls, and the results for the all the genes, including the top 100, are provided in order in Supplementary File S2. For comparison, a PCA plot of 100 randomly sampled genes (Fig. 7b) failed to show any separation of the cancer and control samples.

To ascertain an ordered trend of expression across cancer stages, the linear model given by eqn. (2) was fit. At a *p*-value of 0.05, 14,127 genes were significant, and raising the bar to 1E-5 still left 8032 genes significant. A goodness of fit with eqn. (2) does not equate with a monotonic trend of expression; i.e., a a gene with a significant linear fit is not required to follow a monotonic trend of mean expression with cancer stage. Using the definition of monotonicity given in the Methods section, we found 2109 genes showing strictly monotonic expression with the cancer stage and reaching maximum absolute mean expression in stage IV. Each such gene was annotated and ranked with the *p*-value from eqn. (2). This yielded 1977 genes with significant (i.e, $p\text{-val} < 0.05$) monotonic trends of mean expression across cancer stages, with 1602 upregulated and 375 downregulated. The top 20 such genes are presented in Table 6.

The results from the linear modelling were in contrast with those obtained by Huo et al. [25] and were most likely driven by an improved design and the inclusion of 51 controls in our study. These positive results provided the impetus to pursue stage-driven analysis. Given the conventional AJCC staging, gene expression differences would play a major role in driving the cancer progression. To identify the stage-specific differentially expressed genes, we applied the first contrast matrix (Table 2) and constructed the four-bit stage string of each gene. Based on the stage strings, we binned all the genes, and the string-specific gene lists corresponding to all the partitions in the Venn diagram (Fig. 4) is made available in Supplementary File S3. The size of each such partition is illustrated in Fig. 8. We eliminated the 16,135 genes corresponding to the stage string '0000' ($|\text{lfc}| < 2$ in all stages). To establish the significance of the remaining genes, we applied the second contrast (Table 3) and passed each gene through the four filter criteria. The gradual reduction in candidate stage-specific genes as each criterion was applied, is shown in Table 7. Only genes that passed all criteria were retained as significant stage-specific differentially expressed genes. We obtained 2 stage-I specific, 2 stage-II specific, 10 stage-III specific and 35 stage-IV specific genes (Table 8). Figure 9 shows the volcano plot of these 49 stage-specific genes.

In view of the limited sample size for stage-IV and consequent low power for rejecting false-positives, we stipulated that each stage-IV specific gene would display a smooth increasing or decreasing expression trend through cancer progression culminating in maximum differential expression in stage-IV. On this basis, we pruned the 35 stage-IV specific genes to just the top ten by significance in the linear modelling. This yielded a total of 24 stage-specifc genes of interest.

A heatmap of the lfc expression of these stage-specific genes across the stages was generated (Fig. 10a) and revealed a systematic gradient in expression relative to control, involving both downregulation and overexpression. The map was clustered on the basis of differential expression (i.e, |lfc|) both across stages and across features (i.e, genes) (Fig. 10b). It was seen that stage I genes clustered together, stage II genes co-clustered with NCAPG2 and DLG5 from stage-III, all the other stage-III genes clustered together, while the stage-IV genes formed two separate clusters. It was interesting to note that GABRD emerged as an outgroup to all the clusters, demonstrating its uniqueness.

To identify the biological processes specific to each stage, we used the genes with maximal |lfc| in each stage and performed a stagewise gene set enrichment analysis on two ontologies, the GO and KEGG pathways. Salient

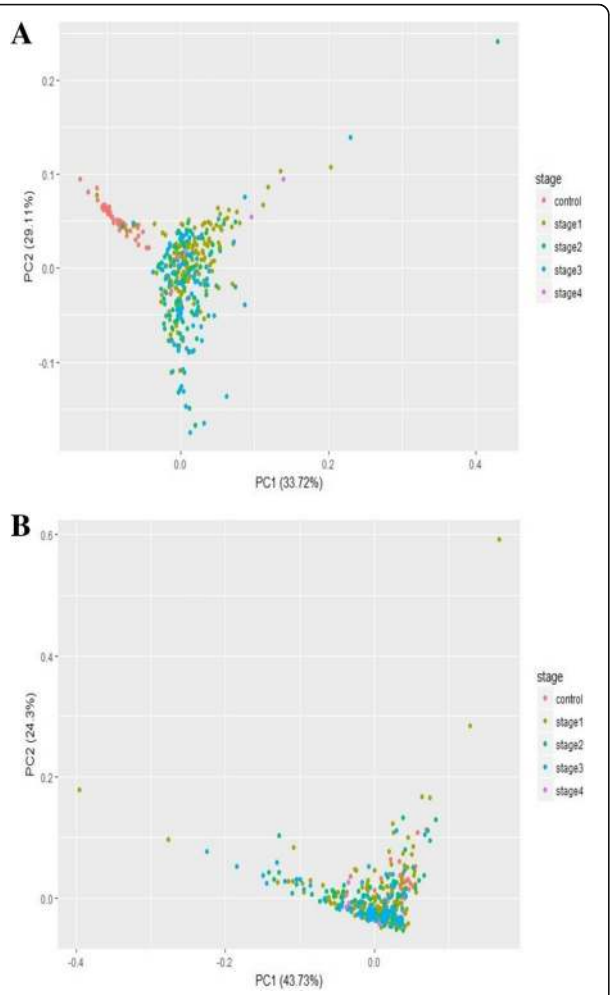

Fig. 7 Principal components analysis of cancer vs control. **a** The first two principal components of the top 100 genes from linear modeling are plotted. It could be seen that control samples (red) clustered independent of the cancer samples (colored by stage). **b** The same analysis repeated with 100 random genes failed to effect a clustering of the control samples relative to the cancer samples

results with respect to KEGG pathways are presented below (Table 9) and the complete KEGG and GO results are available in Supplementary Tables S1 and S2, respectively. In stage I, we found the significant enrichment of cell-cycle signaling pathways (Hippo, Wnt, HIF-1), and viral infection-related pathways (cytokine-cytokine receptor interaction, human papillomavirus infection, HTLV-I infection). In stage II, key signalling pathways (Ras, MAPK) were aberrant. Two liver-specific pathways, alcoholism and cytochrome P450 mediated metabolism of xenobiotics were enriched, as well as standard cancer pathways of bladder, brain, stomach, and skin that might involve generic genetic alterations necessary for cancer cell growth. In stage III, we noticed the significant enrichment of Metabolic pathways that summarize cellular metabolism. This might indicate the metabolic shift needed by the cancer to grow and invade neighboring tissues. Other salient significantly enriched pathways pertained to increased cell cycle progression, DNA replication, chemical carcinogenesis, p53 signaling pathway and cellular senescence, all hallmark processes critical to cancer progression. Stage IV gene set was significantly enriched for bile-related processes (bile secretion, primary bile acid biosynthesis), and ABC transporters (possibly conferring a drug-resistant advanced cancer phenotype). A signaling pathway related to diabetic complications was enriched as well, indicating the role of co-morbidities in driving liver cancer progression. The enrichment analysis of the top 100 genes of the linear model is included in the Supplementary Table S3.

Discussion

When differentially expressed genes are identified in a two-class cancer vs control manner, the information about stage-specificity of differential expression is lost. By applying our protocol, this information is recovered and available for dissection. The top linear model genes and all the stage-specific differentially expressed genes (Table 10) were analyzed with respect to the existing literature.

Top genes of linear models

Three C-type lectin domain proteins (CLEC4M, CLEC1B, CLEC4G) were detected in the top ten genes of linear model given by eqn. (1). Interestingly, this identical cluster of three genes was detected as the most significantly downregulated liver cancer-specific genes in a qPCR study of an independent cohort of 65 tumor-normal matched cases [21]. On screening the top 200 linear model (1) genes against cancer driver genes in the Cancer Gene Census, only four genes were found, namely BUB1B, CDKN2A, EZH2, and RECQL4. The top 200 genes of the linear model given by eqn. (2) overlapped with 111 genes of linear model (1) and yielded six genes from the Cancer Gene Census, namely BUB1B, EZH2, CDKN2C, CANT1, POLD1, and STIL. Both CDKN2A and CDKN2C are cyclin-dependent kinase inhibitors. CDKN2A was a member of the gene signatures for HCC prognosis independently proposed by Gillet et al. [16] and Yang et al. [59]. It was remarkable that GABRD stood out as the top gene in both the linear models, *and* with a monotonic order of expression with the cancer stage. GABRD is discussed further in the section on Stage-IV specific genes. A gene with a monotonicity of expression may be increasingly upregulated as the cancer initiates, progresses and metastasizes, signalling its oncogenic progression; or conversely, it may be increasingly downregulated with the cancer stages, signalling the loss of tumor suppressor activity. Screening the top 200 genes with monotonic expression against

Table 6 Top 20 genes with significant monotonic patterns of expression. Intercept, Coefficient and Adj. p-value are from the linear model given by eqn.(2). Status indicates monotonic upregulation (UP) or monotonic downregulation (DOWN). The genes are sorted by significance (adj.p-value)

Gene	Intercept	Coefficient	Adj.p-value	status
GABRD	-0.96	1.09	2.33E-023	UP
PIGU	3.80	0.43	2.45E-023	UP
NDUFA4L2	2.42	0.80	3.09E-023	UP
CRHBP	2.59	-1.52	3.14E-023	DOWN
C20orf20	3.26	0.39	4.84E-023	UP
PPIA	7.00	0.34	1.65E-022	UP
C12orf47	2.31	0.39	2.75E-022	UP
VIPR1	2.40	-1.12	4.60E-022	DOWN
FLVCR1	0.73	0.72	5.12E-022	UP
TTC13	2.73	0.50	1.06E-021	UP
NXPH4	-2.29	1.31	3.53E-021	UP
CACYBP	4.87	0.42	4.65E-021	UP
MFSD5	3.98	0.34	5.15E-021	UP
PIGC	4.08	0.40	5.90E-021	UP
ATP6AP1	6.40	0.33	6.52E-021	UP
DYNLL1	6.37	0.32	7.09E-021	UP
CCT4	6.28	0.30	1.02E-020	UP
SLC41A3	4.29	0.36	1.11E-020	UP
ZBTB9	2.31	0.38	1.17E-020	UP
NEU1	5.92	0.43	1.20E-020	UP

the Cancer Gene Census yielded a completely different set of six genes: HSP90AB1, ALDH2, ESR1, PPP2R1A, HIST1H4I, SEPT5. HSP90AB1, a heat shock protein and molecular chaperone, was a key result of Xu et al. [56] where it played a dual role, one in the set of 50 hub genes correlated with Barcelona Clinic Liver Cancer (BCLC) staging of HCC patients, and another, in the set of 13 hub genes correlated with overall survival of HCC patients. HSP90AB1 might have a significant role in the aetiology of HCC, given that its expression is known to be upregulated by hepatitis B virus encoded X protein [31]. The monotonic changes in HSP90AB1 might further facilitate its known roles in angiogenesis [19]. The top 200 genes with monotonic expression had 15 genes in common with the top 200 of linear model (1) and 16 genes in common with the top 200 of linear model (2). However, only six genes were common to the top 200 of all three (namely GABRD, PIGU, NDUFA4L2, CRHBP, FLVCR1, TTC13; Fig. 11). NDUFA4L2 has been identified as a target gene of HIF-1 (hypoxia-inducible transcription factor-1), and a key factor driving the metabolic reprogramming in hypoxic microenvironments [46]. Our findings established that not only was NDUFA4L2 significantly overexpressed in HCC (as noted in [27]), but its overexpression follows a significant monotonic pattern across cancer stages, a much stronger statement that would support the role of NDUFA4L2 in driving HCC progression. Similarly, the expression of CRHBP has been recently shown to be negatively associated with the tumor size in HCC [55].

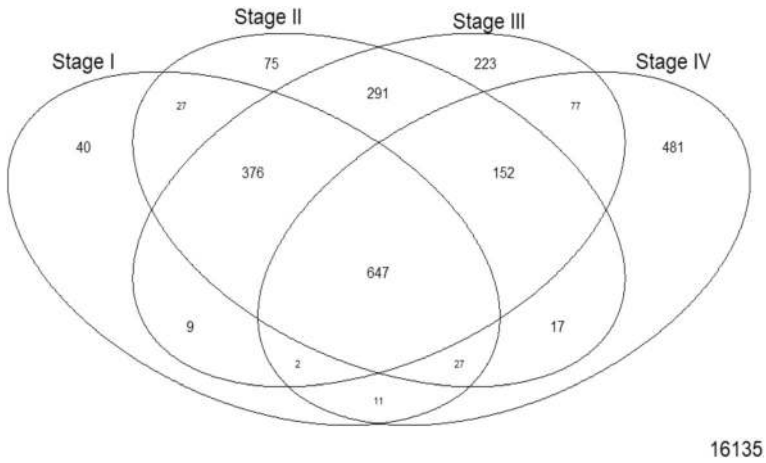

Fig. 8 Venn illustration of the size of each 4-bit string. The numbers of genes with each pattern of differential expression are shown

Our study provides a more quantitative account of the significant monotonic downregulation of CRHBP with the HCC stage. Two proteins of the glycosylphosphatidylinositol (GPI) anchoring system, PIGU and PIGC, were top genes with respect to significant monotonic expression (Table 6); of these, PIGU is a known bladder cancer oncogene [18].

Stage-I specific DEGs (Fig. 12)

CA9 is a member of carbonic anhydrases, which are a large family of zinc metalloenzymes that catalyse the reversible hydration of carbon dioxide. Its expression in clear cell Renal carcinoma, but not in functional kidney cells has gained attention for its use as a pre-operative biomarker [30]. The WNT7B protein is part of the Wnt family, a family of secreted signalling proteins. Elevated WNT7B in pancreatic adenocarcinoma has been found to mediate anchorage independent growth [4]. Surprisingly, both CA9 and WNT7B are downregulated in HCC, most so in stage-I, contrary to their role in other cancers. A concrete interpretation of the role of these genes in HCC awaits appropriately designed experimental studies.

It is pertinent to ask the following question here: which genes are essential for the initiation of HCC?

Table 7 Number of genes in each step of the significance analysis. Differential expression is defined with respect to a threshold |logFC| = 2. Significance analysis proceeds first by significance (i.e, p-value) with respect to control, followed by p-value in each possible pairwise contrast between the different stages. Exclusive DE genes refer to genes differentially expressed in only one of the four stages (corresponding to the bit strings '1000', '0100', '0010' and '0001')

Filtering criteria	STAGE 1	STAGE 2	STAGE 3	STAGE 4	Total
Exclusive DE genes	40	75	223	481	819
DE genes	122	407	844	1082	2455
Adj.p-value w.r.to control	120	406	839	293	1658
p-value 1 x 2	26	187	-	-	213
p-value 1 x 3	19	-	670	-	689
p-value 1 x 4	2	-	-	88	90
p-value 2 x 3	-	13	70	-	83
p-value 2 x 4	-	2	-	46	48
p-value 3 x 4	-	-	10	35	45
Final genes	2	2	10	35	45

Table 8 Final set of highlighted genes in each stage. The genes in each stage are ordered by increasing adjusted p-values of the linear modelling analysis. Stage-IV specific genes with monotonic changes of expression correlating with disease progression are highlighted

STAGE 1	STAGE 2	STAGE 3	STAGE 4
CA9	FAM186A	C12orf48	***GABRD***
WNT7B	APOBEC3B	C15orf42	PECAM1
		ORC6L	***LOC25845***
		ECT2	***CEND1***
		WDHD1	GBX2
		DLG5	***PGAM2***
		XRCC2	***NR1I2***
		NCAPG2	GDF5
		GNMT	CXCR2P1
		PRR11	GPR1
			MUSTN1
			EHD2
			LOC143188
			HIST3H2BB
			CA12
			CDX1
			MYO16
			CPE
			LPPR3
			ZMYND12
			KCNF1
			GPR126
			MCCD1
			GABRB2
			SNCB
			TRIM50
			MT3
			KCNQ2
			DUXA
			C14orf72
			ECEL
			FOXE
			MYH13
			ARHGAP42
			BMP7

Clearly these genes would be differentially expressed in stage I relative to control. All significantly differentially expressed genes with maximal |lfc| in stage-I would be the best candidates for genes involved in the initiation of HCC. These 122 genes are provided in the Supplementary File S3.

Stage-II specific DEGs (Fig. 13)

APOBEC3B, a DNA cytidine deaminase, is a known cancer driver gene in the Cancer Gene Census, but there are no literature reports of its stage-specificity in any cancer. It is known to account for half the mutational load in breast carcinoma, and its target sequence context

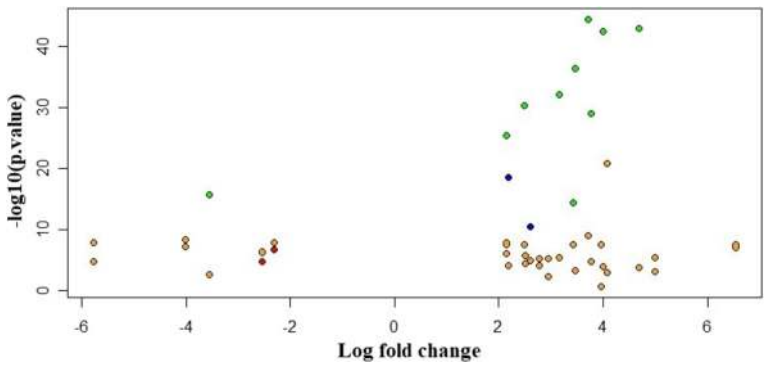

Fig. 9 Volcano plot of the 49 significant stage-specific differentially expressed genes. Stage 1 genes, red; Stage 2, blue; Stage 3, green; and Stage 4, orange. The genes are seen to orient away from the origin and the axes, indicating significance and effect size

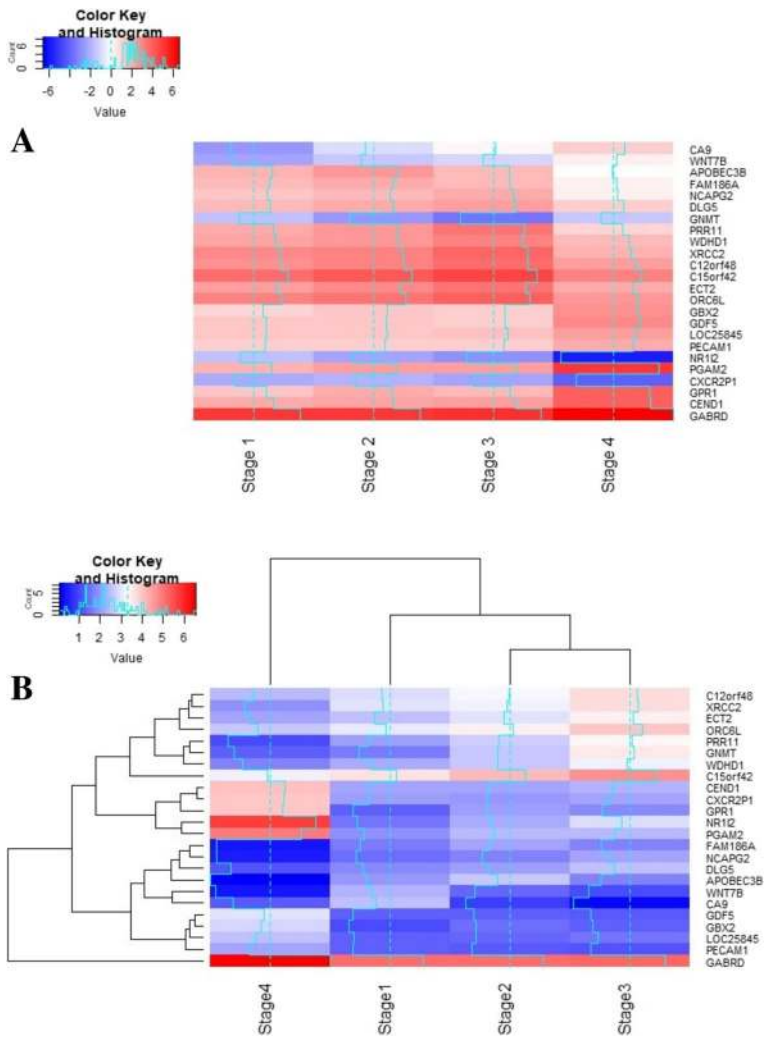

Fig. 10 Heatmap plots of the final 24 stage-specific genes. **a** heatmap generated from the lfc values of all the stage-specific genes (arranged stagewise). The color gradient spans the spectrum from downregulation (blue) to overexpression (red). Log fold changes upto sixfold are seen, indicating 64 times differential expression with respect to control. **b** Representation of the stagewise gene expression based on clustering of differential expression profiles

Table 9 Gene set enrichment analysis. Stage-specific gene sets (all the differentially expressed genes, corresponding to row 'DE genes' in Table 6) were analyzed for significant enrichment with respect to KEGG Pathways. Significance was based on p-value <0.05

Stage	Enriched pathways	p-value
Stage 1	Hippo signalling pathway	3.276e-03
	Cytokine-cytokine receptor interaction	1.218e-02
	Wnt signalling pathway	1.528e-02
	Human papillomavirus infection	1.763e-02
	HTLV-I infection	2.552e-02
	HIF-1 signalling pathway	2.787e-02
Stage 2	Bladder cancer	4.643e-03
	Ras signalling pathway	5.264e-03
	Pathways in cancer	6.211e-03
	Glioma	6.457e-03
	Alcoholism	1.027e-02
	Gastric cancer	1.210e-02
	MAPK signalling pathway	2.526e-02
	Melanoma	3.183e-02
	Metabolism of xenobiotics by cytochrome P450	3.472e-02
Stage 3	Cell cycle	2.881e-18
	DNA replication	6.526e-11
	Chemical carcinogenesis	1.233e-06
	Metabolic pathways	1.204e-03
	Cellular senescence	7.203e-03
	p53 signalling pathway	7.275e-03
Stage 4	Bile secretion	2.479e-06
	ABC transporters	7.146e-06
	Primary bile acid biosynthesis	2.357e-03
	AGE-RAGE signalling pathway in diabetic complications	3.024e-02

was found to be highly mutated in Bladder, lung, cervix, neck, and head cancers as well [7]. Further studies have attributed specific hypermutation signatures across all cancers to the APOBEC family, including APOBEC3B [1]. Here APOBEC3B is upregulated, increasing its capacity to inflict the hypermutator phenotype, and highlighting an intriguing stage-specificity in its action. FAM186A polymorphisms have been reported in GWAS and SNP studies on colorectal cancer patients and shown to have a significant odds ratio in risk heritability [48]. FAM163A was a component of the 8-gene signature used for the risk stratification of HCC patients [39].

Stage-III specific DEGs (Fig. 14)

C12orf48, also known as PARI, participates in the homologous recombination pathway of DNA repair, and its overexpression has been reported in pancreatic cancer[35]. Further PARI was recently identified as a transcriptional target of FOXM1 [62], which is a well-validated upregulated gene in HCC [21]. DLG5 is a cell

Table 10 Stagewise effect sizes and significance of stage specific genes. The stagewise log foldchanges of differential expression of each candidate stage-specific gene in tumor samples relative to normal control samples are shown, along with significance values, and its inferred regulation status. In stage-IV, only the top 10 genes are shown. The stage-specificity of the genes are emphasized

GENE	β_0	β_1	β_2	β_3	β_4	Adj.p.value (from contrasts against control)	Adj.p.value (from linear model)	Regulation status (Up/Down)
STAGE 1								
CA9	0.1	**-2.5**	-0.9	0.2	1.3	3.66e-05	3.44e-09	Down
WNT7B	-1.8	**-2.3**	-1.4	-1.1	0.3	5.45e-07	2.07e-06	Down
STAGE 2								
APOBEC3B	-0.9	2.0	**2.6**	1.7	-0.6	1.12e-10	1.17e-09	Up
FAM186A	-4.2	1.7	**2.2**	1.8	0.4	2.62e-18	1.50e-12	Up
STAGE 3								
NCAPG2	1.5	1.5	1.8	**2.1**	0.4	6.03e-25	1.76e-24	Up
DLG5	2.3	1.8	2.1	**2.5**	1.1	1.23e-29	2.90e-29	Up
GNMT	6.9	-1.6	-2.6	**-3.6**	-1.3	9.69e-16	2.40e-15	Down
PRR11	-3.8	2.1	2.7	**3.4**	1.0	1.74e-14	4.65e-13	Up
WDHD1	-1.4	2.3	2.7	**3.2**	1.8	2.25e-31	8.87e-31	Up
XRCC2	-3.8	2.9	3.2	**3.8**	1.9	2.29e-28	7.06e-28	Up
C12orf48	-1.9	2.8	3.2	**3.7**	2.4	5.88e-43	1.19e-43	Up
C15orf42	-3.5	3.7	4.2	**4.7**	3.2	1.52e-41	1.17e-42	Up
ECT2	0.2	2.5	3.0	**3.5**	2.2	2.62e-35	4.29e-35	Up
ORC6L	-2.4	3.1	3.5	**4.0**	2.6	4.19e-41	5.55e-42	Up
STAGE 4								
GABRD	-3.8	5.1	5.1	5.2	**6.5**	3.15e-17	5.53e-78	Up
PECAM1	3.8	1.3	1.2	1.2	**2.1**	4.69e-07	7.64e-24	Up
LOC25845	2.0	1.3	1.4	1.6	**2.5**	4.47e-06	1.29e-20	Up
CEND1	-5.0	2.2	2.2	2.4	**4.1**	4.42e-06	1.08e-17	Up
GBX2	-6.3	1.0	1.5	1.3	**2.8**	1.20e-05	2.59e-13	Up
PGAM2	-1.7	1.9	2.4	2.5	**5.0**	5.73e-05	1.72e-11	Up
NR1I2	5.6	-1.5	-2.3	-2.9	**-5.8**	8.41e-05	8.06e-11	Down
GDF5	-6.1	1.3	1.3	1.2	**2.9**	2.73e-05	8.22e-11	Up
CXCR2P1	2.0	-2	-2	-2.2	**-4.0**	4.52e-04	1.77e-10	Down
GPR1	-5.5	1.3	2.1	1.8	**3.9**	1.42e-04	4.33e-10	Up

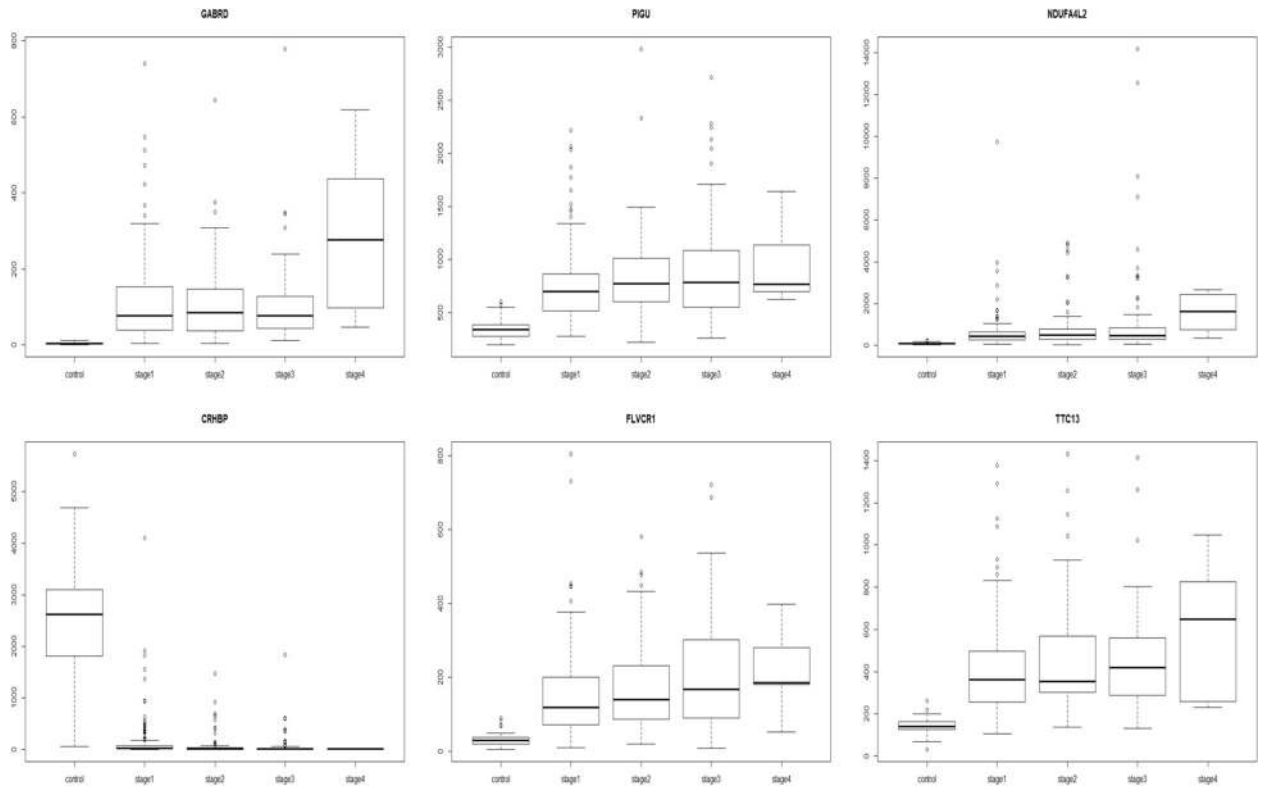

Fig. 11 Boxplot of top genes with monotonic expression. These six genes (GABRD, PIGU, NDUFA4L2, CRHBP, FLVCR1, TTC13) showed monotonic trends of expression across the cancer stages, and were topranked in both the linear models given by eqns. (1) and (2)

polarity gene and its downregulation has been implicated in the malignancy of breast [32], prostate [49] and bladder cancers [65]. It has been recently found that lower DLG5 expression is correlated with advanced stages of HCC and essential for invadopodium formation, an event critical to cancer metastasis [26]. It is surprising that our study has identified a stage-III specific upregulation in DLG5. Interestingly, evidence is emerging to lend support to our finding that DLG5 might be tumor-promoting. In a very recent review, Saito et al. [43] reinterpreted published results on cell polarity and cancer, and advanced an alternative perspective on the role of polarity regulators in cancer biology. They argued that both cellular and subcellular polarity would be regulated by DLG5 and related polarity proteins. Subcellular polarity might improve the

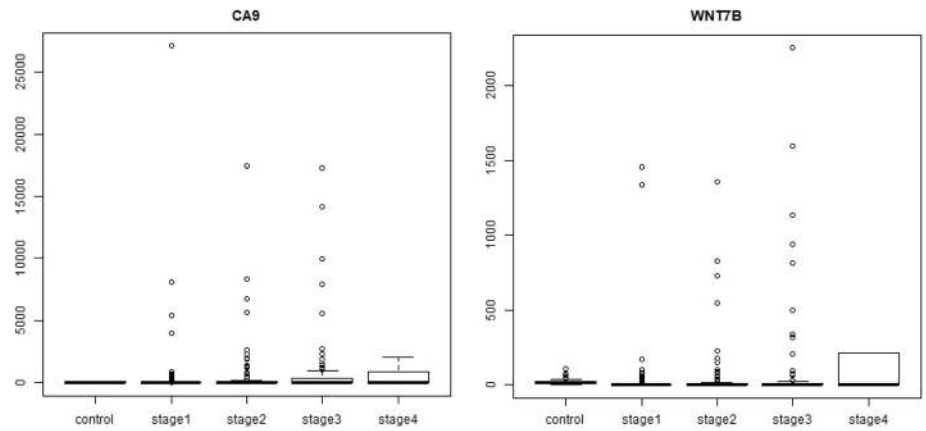

Fig. 12 Boxplot of stage-I specific genes. It is seen that CA9 and WNT7B are both maximally downregulated in stage-I

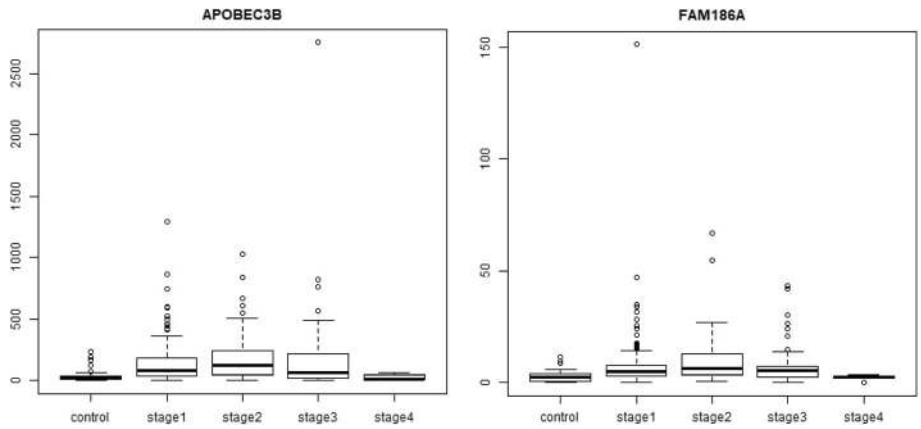

Fig. 13 Boxplot of stage-II specific genes. It is seen that both APOBEC3B and FAM186A are maximally overexpressed in stage-II, the trend following an inverted U-shape

cellular fitness for proliferation and stemness, thereby causing tumor promotion. Hence cell polarity regulation is anti-tumorigenic and subcellular polarity regulation is pro-tumorigenic, and our analysis has uncovered the pro-tumorigenic upregulated activity of DLG5. ECT2 encodes a guanine nucleotide exchange factor that remains elevated during the G2 and M phase in cellular mitosis. ECT2 is found to be upregulated in lung adenocarcinoma and lung squamous cell carcinoma [66], as well as in invasive breast cancer [52]. NCAPG2 is a component of the condensing II complex and involved in chromosome segregation during mitosis. NCAPG2 level were found to be increased in non-small cell lung cancer, and its overexpression was found to be correlated with lymph node metastasis, thus enabling the use of NCAPG2 as a poor prognostic biomarker in lung adenocarcinoma [61]. GNMT is a methyltransferase that catalyses conversion of S-adenosine methionine to s-adenosyl cysteine. In the absence of GNMT, S-adenosine methionine causes hypermethylation of DNA, which represses GNMT levels and is found in HCC samples [24]. This is an epigenetic mechanism for loss of function of tumor suppressors and our study here confirmed the downregulation of GNMT expression. PRR11 is found to be over-expressed in lungs, and its silencing using siRNA resulted in cell cycle arrest and apoptotic cell death, followed by decreased cell growth and viability [64]. A similar knock out experiment of PRR11 in hilar cholangiocarcinoma cell lines resulted in decreased cellular proliferation, migration, and tumor growth [9]. WDHD1 is a key post-transcriptional regulator of centromeric, and consequently genomic, integrity [23] and its overexpression has been identified as biomarker of acute myeloid leukemia [53], and lung and esophageal carcinomas [44]. C15orf42 has been implicated in nasopharyngeal carcinoma [3]. ORC6L overexpression has been identified as a prognostic biomarker of colorectal cancer possibly by enhancing chromosomal instability [54]. XRCC2 was found to increase locally advanced rectal cancer radioresistance by repairing DNA double-strand breaks and preventing cancer cell apoptosis [40]. XRCC2 was also highlighted in the gene signature for HCC prognosis advanced by Gillet et al. [16].

Stage-IV specific DEGs (Fig. 15)

GABRD, which was the top gene in the linear models as well, encodes for the delta subunit of the gamma-amino butyric acid receptor. The GABA receptor family was found to be frequently downregulated in cancers, except for GABRD, which was found to be up-regulated. Gross et al. [17] proposed that the GABA receptor gene family might play a role in the proliferation independent differentiation of cancer cells. GBX2 is part of the GBX gene family, which are homeobox containing DNA binding transcription factors. GBX2 is overexpressed in prostate cancer and studies show that expression of GBX2 is required for malignant growth of human prostate cancer [15]. PECAM1 overexpression has been linked to peritoneal recurrence of stage II/III gastric cancer patients [47]. CEND1 has been identified as a cell-cycle protein [50]. PGAM2 is a glycolytic enzyme whose upregulation is essential for tumor cell proliferation [57]. NR1I2 downregulation has been used in constructing a prognostic 9-genes expression signature of gastric cancer [51]. GDF5 has been shown to be a downstream target of the TGF-beta signaling pathway [33], stimulating angiogenesis required for the growth and spread of the cancer. GPR1 has been reported to be involved in promoting cutaneous squamous cell carcinoma migration [12]. Two other stage-IV specific genes, namely the downregulated CXCR2P1, which is a C-X-C motif chemokine receptor 2 pseudogene 1, and LOC25845, are minimally documented in the literature in the context of HCC, other cancers or any other

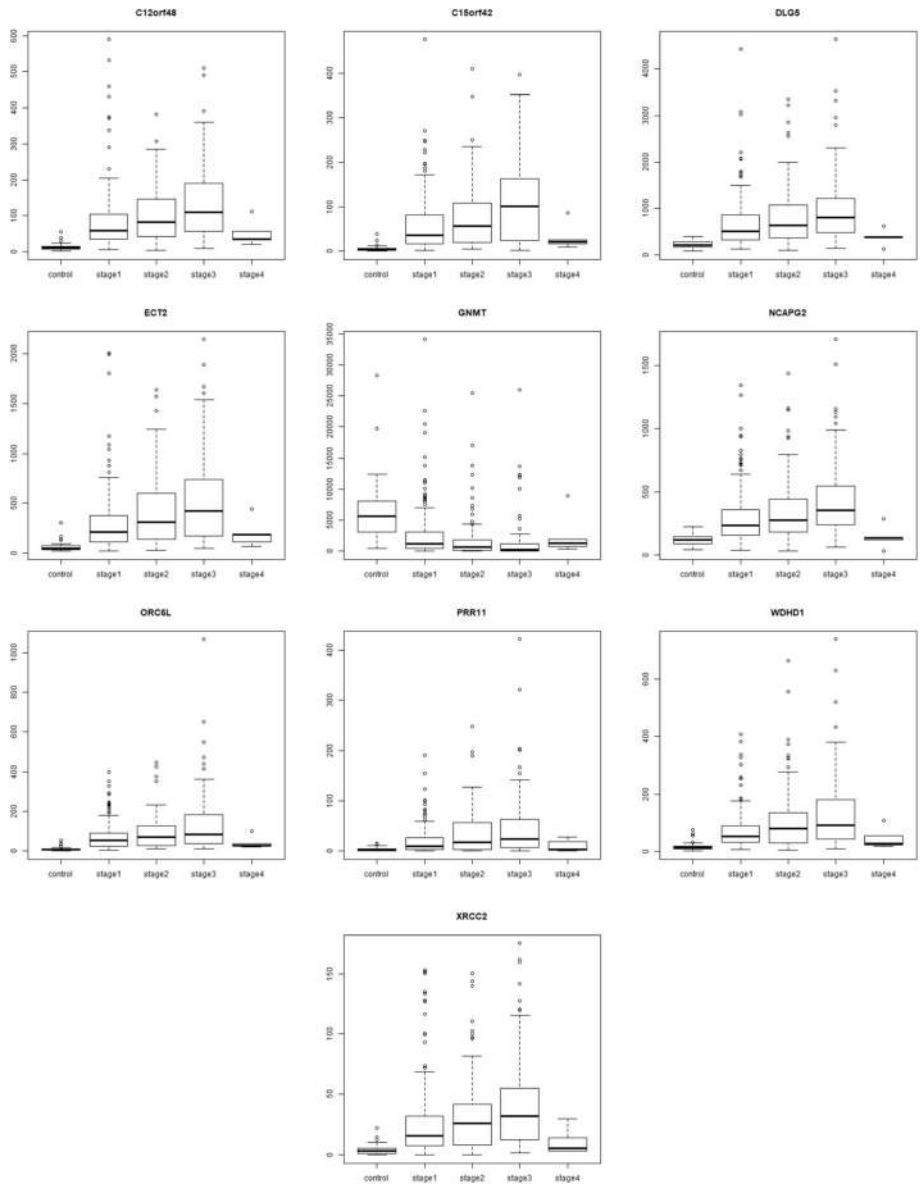

Fig. 14 Boxplot of stage-III specific genes. Except for GNMT, the expression of stage-III specific genes show a peak in stage-III, with the expression trend following an inverted U-shape across the stages. The expression trend is convex and reversed for the downregulated GNMT, with minimum expression in stage-III

condition. It is worth mentioning however that CXCR2, a member of the GPCR protein family binding the interleukin IL8, has been reported as an effective non-invasive blood based biomarker for HCC [45]. It is notable that ARHGAP42, a Rho GTPase activating protein, was another key result of Xu et al. [56], finding a place both in their set of 50 hub genes correlated with the BCLC staging of HCC patients, and in the set of 13 hub genes correlated with overall survival of HCC patients.Most of the stage-IV specific genes show contra-regulation (i.e, no clear trend) across cancer stages, and only 15 of the 35 genes revealed a monotonic pattern of expression (highlighted in Table 8). The other 20 genes could be unique to the hallmarks of stage-IV cancer, e.g., processes related to lymph node involvement and/or metastasis.

Conclusion

We have developed an original protocol for the stage-wise dissection of the HCC transcriptome. We were able to successfully fit a linear model across cancer stages and detected genes with a strong linear expression trend in the cancer phenotype. These genes were found to effectively separate the control and cancer samples. We

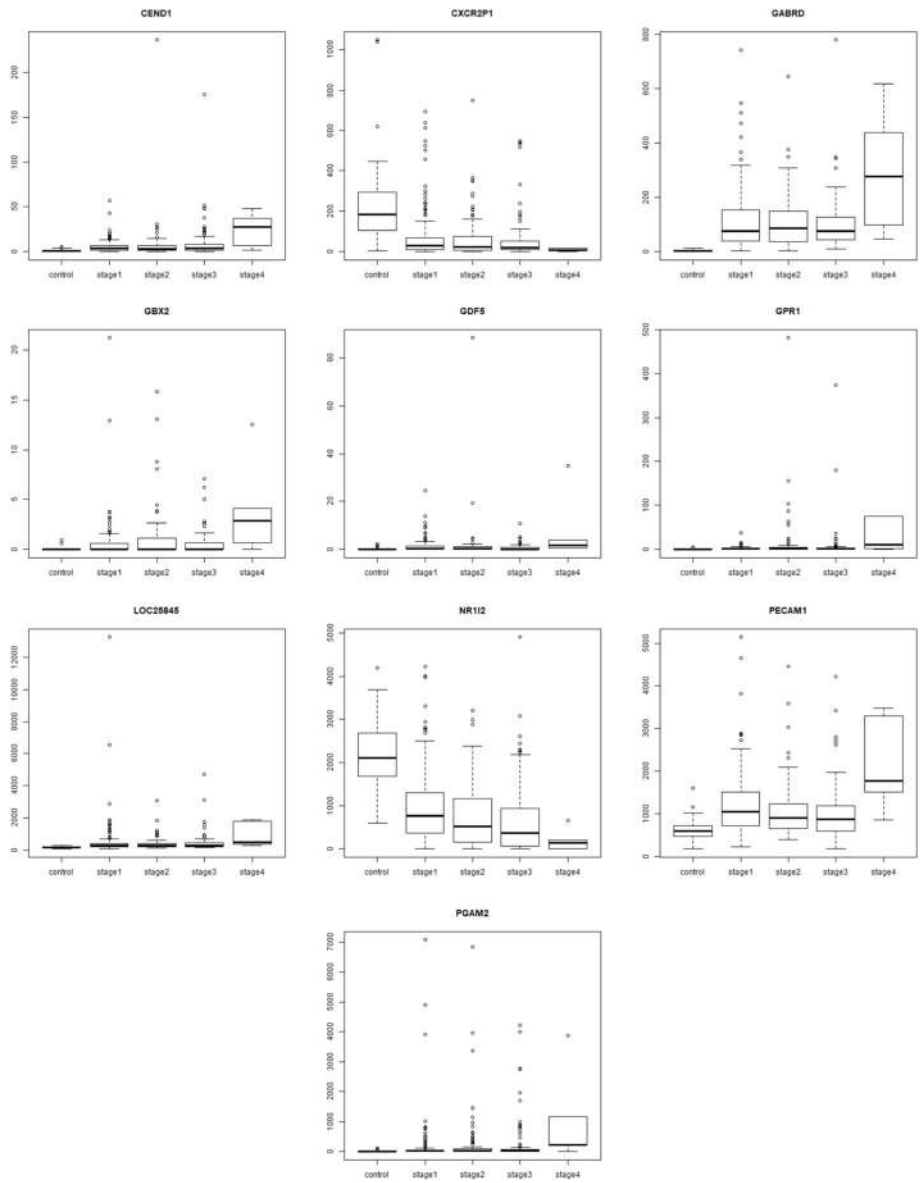

Fig. 15 Boxplot of top 10 stage-IV specific genes. All genes, except NR1I2 and CXCR2P1, show a smooth increasing expression trend reaching peak expression in stage-IV. In the case of NR1I2 and CXC2RP1, the trend is reversed, with the expression decreasing smoothly to touch the minimum in stage-IV

were able to assign 2455 differentially expressed genes into one of four stages and visualized their stage specific expression using boxplots. Using a multi-layered approach, we were able to assess the significance of each stage-specific DEG and narrowed down to a handful of candidate significant stage-specific DEG's. Our analysis yielded two stage-I specific genes (CA9, WNT7B), two stage-II specific genes (APOBEC3B, FAM186A), ten stage-III specific genes (including DLG5, NCAPG2, GNMT and XRCC2) and 35 stage-IV specific genes (including GABRD and CXCR2P1). Though most of these genes constituted novel findings in the context of HCC, a comprehensive literature search indicated connections with other cancer conditions. The analysis of monotonicity of expression has uncovered two genes with documented HCC connection, namely NDUFA4L2 and CRHBP. Correlation of our analysis with gene signatures based on the BCLC staging system revealed two common genes, namely HSP90AB1 and ARHGAP42. Our study might deepen our understanding of the mechanistic basis of HCC progression, and lay the foundation for the development of HCC diagnosis and treatment strategies. Translational research could transform our results into a panel of biomarkers for early clinical decision-making and

rational drug development. It is straightforward to extend our computational methodology to the stage-based analysis of other cancers to obtain a fuller view of disease initiation, progression, and metastasis.

Abbreviations

AJCC: American Joint Committee on Cancer; BCLC: Barcelona Clinic Liver Cancer; BMI: Body mass index; DEG: Differentially expressed genes; EST: Expressed sequence tag; GABA: Gamma Amino Butyric Acid; GO: Gene Ontology; HCC: Hepatocellular carcinoma; KEGG: Kyoto Encyclopaedia of Genes and Genomes; lfc: log fold-change; LIHC: Liver hepatocellular carcinoma; PCA: Principal Components Analysis; qPCR: quantitative Polymerase Chain Reaction; RSEM: RNA-Seq by Expectation Maximization; TCGA: The Cancer Genome Atlas; TNM: Tumor, Node, Metastasis

Acknowledgements

We would like to thank the peer reviewers for improving an earlier version of the manuscript. We would like to thank the departments of Bioinformatics and Bioengineering, SASTRA deemed University, for infrastructure and computing support. A.P. acknowledges support from DST-SERB grant no. EMR/2017/000470, Government of India.

Authors' contributions

AP conceived and designed the study. AS and AP performed the experiments and analyzed the data. AP and AsS wrote the manuscript. All authors approved the final manuscript.

Funding

This work was supported in part by DST-SERB grant no. EMR/2017/000470, Government of India.

Availability of data and materials

All data and material are available as supplementary information (https://doi.org/10.6084/m9.figshare.6455024).

Ethics approval and consent to participate

All data used in the study is available in the public domain, constituting anonymized patient data provided by the TCGA consortium.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Bioengineering, School of Chemical and BioTechnology, SASTRA deemed University, Thanjavur, Tamil Nadu 613401, India. [2]Department of Bioinformatics, School of Chemical and BioTechnology, SASTRA deemed University, Thanjavur, Tamil Nadu 613401, India.

Received: 18 August 2018 Accepted: 16 June 2019
Published online: 05 July 2019

References

1. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jäger N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdés-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415–21. https://doi.org/10.1038/nature12477.
2. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, Meyer L, Gress DM, Byrd DR, Winchester DP. The eighth edition AJCC Cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin. 2017;67(2):93–9.
3. An F, Zhang Z, Xia M. Functional analysis of the nasopharyngeal carcinoma primary tumor-associated gene interaction network. Mol Med Rep. 2015; 12(4):4975–80. https://doi.org/10.3892/mmr.2015.4090.
4. Arensman MD, Kovochich AN, Kulikauskas RM, Lay AR, Yang PT, Li X, Donahue T, Major MB, Moon RT, Chien AJ, Dawson DW. WNT7B mediates autocrine WNT/β-catenin signaling and anchorage-independent growth in pancreatic adenocarcinoma. Oncogene. 2014;33(7):899.
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25.
6. Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized TCGA data from broad GDAC firehose 2016_01_28 run. Broad institute of MIT and Harvard. Dataset; 2016. https://doi.org/10.7908/C11G0KM9.
7. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat Genet. 2013;45(9):977.
8. Cancer Research UK, http://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer, Accessed 05 May 2018.
9. Chen Y, Cha Z, Fang W, Qian B, Yu W, Li W, Yu G, Gao Y. The prognostic potential and oncogenic effects of PRR11 expression in hilarcholangiocarcinoma. Oncotarget. 2015;6(24):20419.
10. Chuang SC, La Vecchia C, Boffetta P. Liver cancer: descriptive epidemiology and risk factors other than HBV and HCV infection. Cancer Lett. 2009;286(1):9–14.
11. Farazi PA, DePinho RA. Hepatocellular carcinoma pathogenesis: from genes to environment. Nat Rev Cancer. 2006;6(9):674.
12. Farsam V, et al. Senescent fibroblast-derived Chemerin promotes squamous cell carcinoma migration. Oncotarget. 2016;7(50):83554–69. https://doi.org/10.18632/oncotarget.13446.
13. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:E359–86.
14. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.
15. Gao AC, Lou W, Isaacs JT. Down-regulation of homeobox gene GBX2 expression inhibits human prostate cancer clonogenic ability and tumorigenicity. Cancer Res. 1998;58(7):1391–4.
16. Gillet JP, Andersen JB, Madigan JP, Varma S, Bagni RK, Powell K, et al. A gene expression signature associated with overall survival in patients with hepatocellular carcinoma suggests a new treatment strategy. Mol Pharmacol. 2016;89(2):263–72.
17. Gross AM, Kreisberg JF, Ideker T. Analysis of matched tumor and normal profiles reveals common transcriptional and epigenetic signals shared across cancer types. PLoS One. 2015;10(11):e0142618.
18. Guo Z, Linn JF, Wu G, Anzick SL, Eisenberger CF, Halachmi S, Cohen Y, Fomenkov A, Hoque MO, Okami K, Steiner G, Engles JM, Osada M, Moon C, Ratovitski E, Trent JM, Meltzer PS, Westra WH, Kiemeney LA, Schoenberg MP, Sidransky D, Trink B. CDC91L1 (PIG-U) is a newly discovered oncogene in human bladder cancer. Nat Med. 2004;10(4):374–81.
19. Haase M, Fitze G. HSP90AB1: helping the good and the bad. Gene. 2016;575:171–86.
20. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
21. Ho DWH, Kai AKL, Ng IOL. TCGA whole-transcriptome sequencing data reveals significantly dysregulated genes and signaling pathways in hepatocellular carcinoma. Front Med. 2015;9(3):322–30.
22. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med. 1990;9:811–8.
23. Hsieh CL, et al. WDHD1 modulates the post-transcriptional step of the centromeric silencing pathway. Nucleic Acids Res. 2011;39(10):4048–62. https://doi.org/10.1093/nar/gkq1338.
24. Huidobro C, Toraño EG, Fernández AF, Urdinguio RG, Rodríguez RM, Ferrero C, Martínez-Camblor P, Boix L, Bruix J, García-Rodríguez JL, Varela-Rey M. A DNA methylation signature associated with the epigenetic repression of glycine N-methyltransferase in human hepatocellular carcinoma. J Mol Med. 2013;91(8):939–50.

25. Huo T, Canepa R, Sura A, Modave F, Gong Y. Colorectal cancer stages transcriptome analysis. PLoS One. 2017;12(11):e0188697.
26. Ke Y, et al. Discs large homolog 5 decreases formation and function of invadopodia in human hepatocellular carcinoma via Girdin and Tks5. Int J Cancer. 2017;141(2):364–76. https://doi.org/10.1002/ijc.30730.
27. Lai RK, Xu IM, Chiu DK, Tse AP, Wei LL, Law CT, Lee D, Wong CM, Wong MP, Ng IO, Wong CC. NDUFA4L2 fine-tunes oxidative stress in hepatocellular carcinoma. Clin Cancer Res. 2016;22(12):3105–17. https://doi.org/10.1158/1078-0432.CCR-15-1987.
28. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
29. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(1):323.
30. Li G, Feng G, Zhao A, Péoc'h M, Cottier M, Mottet N. CA9 as a biomarker in preoperative biopsy of small solid renal masses for diagnosis of clear cell renal cell carcinoma. Biomarkers. 2017;22(2):123–6.
31. Li WH, Miao XH, Qi ZT, Ni W, Zhu SY, Fang F. Proteomic analysis of differently expressed proteins in human hepatocellular carcinoma cell lines HepG2 with transfecting hepatitis B virus X gene. Chin Med J. 2009;5:15–23.
32. Liu J, et al. Loss of DLG5 promotes breast cancer malignancy by inhibiting the hippo signaling pathway. Sci Rep. 2017;7:42125. https://doi.org/10.1038/srep42125.
33. Margheri F, et al. GDF5 regulates TGFß-dependent angiogenesis in breast carcinoma MCF-7 cells: in vitro and in vivo control by anti-TGFß peptides. PLoS One. 2012;7(11):e50342. https://doi.org/10.1371/journal.pone.0050342.
34. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 2009;25:765–71.
35. O'Connor KW, et al. PARI overexpression promotes genomic instability and pancreatic tumorigenesis. Cancer Res. 2013;73(8):2529–39. https://doi.org/10.1158/0008-5472.CAN-12-3313.
36. Palaniappan A, Ramar K, Ramalingam S. Computational identification of novel stage-specific biomarkers in colorectal cancer progression. PLoS One. 2016;11(5):e0156665.
37. Peng L, Bian XW, Li DK, Xu C, Wang GM, Xia QY, Xiong Q. Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 Normal tissue controls across 12 TCGA Cancer types. Sci Rep. 2015;5:13413. https://doi.org/10.1038/srep13413.
38. Perz JF, Armstrong GL, Farrington LA, Hutin YJ, Bell BP. The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. J Hepatol. 2006;45(4):529–38.
39. Qiao GJ, Chen L, Wu JC, Li ZR. Identification of an eight-gene signature for survival prediction for patients with hepatocellular carcinoma based on integrated bioinformatics analysis. PeerJ. 2019;7:e6548.
40. Qin CJ, Song XM, Chen ZH, Ren XQ, Xu KW, Jing H, He YL. XRCC2 as a predictive biomarker for radioresistance in locally advanced rectal cancer patients undergoing preoperative radiotherapy. Oncotarget. 2015;6(31):32193.
41. Research Network TCGA. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell. 2017;169(7):1327–41.
42. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses forRNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. https://doi.org/10.1093/nar/gkv007.
43. Saito Y, Desai RR, Muthuswamy SK. Reinterpreting polarity and cancer: the changing landscape from tumor suppression to tumor promotion. Biochim Biophys Acta. 2018;1869(2):103–16. https://doi.org/10.1016/j.bbcan.2017.12.001.
44. Sato N, et al. Activation of WD repeat and high-mobility group box DNA binding protein 1 in pulmonary and esophageal carcinogenesis. Clin Cancer Res. 2010;16(1):226–39. https://doi.org/10.1158/1078-0432.CCR-09-1405.
45. Shi M, Chen MS, Sekar K, Tan CK, Ooi LL, Hui KM. A blood-based three-gene signature for the non-invasive detection of early human hepatocellular carcinoma. Eur J Cancer. 2014;50(5):928–36.
46. Tello D, Balsa E, Acosta-Iborra B, Fuertes-Yebra E, Elorza A, Ordóñez Á, Corral-Escariz M, Soro I, López-Bernardo E, Perales-Clemente E, Martínez-Ruiz A, Enríquez JA, Aragonés J, Cadenas S, Landázuri MO. Induction of the mitochondrial NDUFA4L2 protein by HIF-1α decreases oxygen consumption by inhibiting complex I activity. Cell Metab. 2011;14(6):768–79. https://doi.org/10.1016/j.cmet.2011.10.008.
47. Terashima M, et al. TOP2A, GGH, and PECAM1 are associated with hematogenous, lymph node, and peritoneal recurrence in stage II/III gastric cancer patients enrolled in the ACTS-GC study. Oncotarget. 2017;8(34):57574–82. https://doi.org/10.18632/oncotarget.15895.
48. Timofeeva MN, et al. Recurrent coding sequence variation explains only a small fraction of the genetic architecture of. Colorectal Cancer Sci Rep. 2015;5:16286. https://doi.org/10.1038/srep16286.
49. Tomiyama L, Sezaki T, Matsuo M, Ueda K, Kioka N. Loss of Dlg5 expression promotes the migration and invasion of prostate cancer cells via Girdin phosphorylation. Oncogene. 2015;34(9):1141–9. https://doi.org/10.1038/onc.2014.31.
50. Tsioras K, Papastefanaki F, Politis PK, Matsas R, Gaitanou M. Functional interactions between BM88/Cend1, ran-binding protein M and Dyrk1B kinase affect cyclin D1 levels and cell cycle progression/exit in mouse neuroblastoma cells. PLoS One. 2013;8(11):e82172. https://doi.org/10.1371/journal.pone.0082172.
51. Wang Z, Chen G, Wang Q, Lu W, Xu M. Identification and validation of a prognostic 9-genes expression signature for gastric cancer. Oncotarget. 2017;8(43):73826–36. https://doi.org/10.18632/oncotarget.17764.
52. Wang HK, Liang JF, Zheng HX, Xiao H. Expression and prognostic significance of ECT2 in invasive breast cancer. J Clin Pathol. 2018;71(5):442–5. https://doi.org/10.1136/jclinpath-2017-204569.
53. Wermke M, et al. RNAi profiling of primary human AML cells identifies ROCK1 as a therapeutic target and nominates fasudil as an antileukemic drug. Blood. 2015;125(24):3760–8. https://doi.org/10.1182/blood-2014-07-590646.
54. Xi Y, Formentini A, Nakajima G, Kornmann M, Ju J. Validation of biomarkers associated with 5-fluorouracil and thymidylate synthase in colorectal cancer. Oncol Rep. 2008;19(1), 257–62.
55. Xia HB, Wang HJ, Fu LQ, Wang SB, Li L, Ru GQ, He XL, Tong XM, Mou XZ, Huang DS. Decreased CRHBP expression is predictive of poor prognosis in patients with hepatocellular carcinoma. Oncol Lett. 2018;16(3):3681–9. https://doi.org/10.3892/ol.2018.9073.
56. Xu W, Rao Q, An Y, Li M, Zhang Z. Identification of biomarkers for Barcelona clinic liver Cancer staging and overall survival of patients with hepatocellular carcinoma. PLoS One. 2018;13(8):e0202763.
57. Xu Y, et al. Oxidative stress activates SIRT2 to deacetylate and stimulate phosphoglycerate mutase. Cancer Res. 2014;74(13):3630–42. https://doi.org/10.1158/0008-5472.CAN-13-3615.
58. Xu XR, Huang J, Xu ZG, Qian BZ, Zhu ZD, Yan Q, Cai T, Zhang X, Xiao HS, Qu J, Liu F. Insight into hepatocellular carcinogenesis at transcriptome level by comparing gene expression profiles of hepatocellular carcinoma with those of corresponding noncancerous liver. Proc Natl Acad Sci. 2001;98(26):15089–94.
59. Yang H, Zhang X, Cai XY, Wen DY, Ye ZH, Liang L, Zhang L, Wang HL, Chen G, Feng ZB. From big data to diagnosis and prognosis: gene expression signatures in liver hepatocellular carcinoma. PeerJ. 2017;5:e3089.
60. Yang JD, Roberts LR. Hepatocellular carcinoma: a global view. Nat Rev Gastroenterol Hepatol. 2010;7:448–58. https://doi.org/10.1038/nrgastro.2010.100.
61. Zhan P, Xi GM, Zhang B, Wu Y, Liu HB, Liu YF, Xu WJ, Zhu Q, Cai F, Zhou ZJ, Miu YY. NCAPG2 promotes tumour proliferation by regulating G2/M phase and associates with poor prognosis in lung adenocarcinoma. J Cell Mol Med. 2017;21(4):665–76.
62. Zhang Y, et al. PARI functions as a new transcriptional target of FOXM1 involved in gastric cancer development. Int J Biol Sci. 2018;14(5):531–41. https://doi.org/10.7150/ijbs.23945.
63. Zhang Y, Qiu Z, Wei L, Tang R, Lian B, Zhao Y, He X, Xie L. Integrated analysis of mutation data from various sources identifies key genes and signaling pathways in hepatocellular carcinoma. PLoS One. 2014;9(7):e100854.
64. Zhao Q. RNAi-mediated silencing of praline-rich gene causes growth reduction in human lung cancer cells. Int J Clin Exp Pathol. 2015;8(2):1760.
65. Zhou Z, et al. Methylation-mediated silencing of Dlg5 facilitates bladder cancer metastasis. Exp Cell Res. 2015;331(2):399–407. https://doi.org/10.1016/j.yexcr.2014.11.015.
66. Zhou S, Wang P, Su X, Chen J, Chen H, Yang H, Fang A, Xie L, Yao Y, Yang J. High ECT2 expression is an independent prognostic factor for poor overall survival and recurrence-free survival in non-small cell lung adenocarcinoma. PLoS One. 2017;12(10):e0187356.

Publisher's Note