A comprehensive algorithmic dissection yields biomarker discovery and insights into the discrete stage-wise progression of colorectal cancer

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

[1]Department of Bioinformatics, School of Chemical and Biotechnology, SASTRA deemed University, Thanjavur Tamilnadu 613401. India

[2]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen. Denmark

[*]Corresponding author: apalania@scbt.sastra.edu

ABSTRACT

Colorectal cancer remains an increasingly common disease with uncommon burden of disease, heterogeneity in manifestation, and no definitive treatment. Against this backdrop, renewed efforts to unravel the genetic drivers of colorectal cancer progression are paramount. Early-stage detection of cancer increases success of treatment as well as prognosis. Here, we have executed a comprehensive computational workflow aimed at uncovering the discrete stagewise genetic drivers of colorectal cancer progression. Using the TCGA COADREAD expression data and clinical metadata, we constructed stage-specific linear models and contrasts to identify stage-specific differentially expressed genes. Stage-specific differentially expressed genes with a significant monotone trend of expression across the stages were identified as progression-significant biomarkers. Among the biomarkers identified are: CRLF1, CALB2 (stage-I specific), GREM2, ADCY5, PLAC2, DMRT3 (stage-II specific), PIGR, SLC26A9 (stage-III specific), GABRD, DLX3, CST6, HOTAIR (stage-IV specific), and CDH3, KRT80, AADACL2, OTOP2, FAM135B, HSP90AB1 (top linear model genes). In particular the study yielded 19 genes that are progression-significant such as CCDC19, SERPINE1, HOXC11, SOX10. The existing literature for many of these biomarkers are discussed and analyzed for encouraging evidence of translational utility that would still need clinical validation. The study yielded many classes of biomarkers, which could serve in signature panels for early-stage colorectal cancer diagnosis as well as establish strategies for therapy. Our work is a concrete step in the direction of establishing the molecular signatures of the discrete progressive stages of colorectal cancer, with the future goal of securing more effective treatment for the condition.

INTRODUCTION

Colorectal adenocarcinoma (COADREAD), or colorectal cancer, is the third most commonly diagnosed cancer in males and the second in females, with an estimated 1.4 million cases and 693,900 deaths occurring in 2012 (Torre et al,2015). Colorectal cancer has more environmental causes than familial, making its incidence more sporadic (Haggar et al., 2009). The main environmental causes include diet (Willett, 2005), physical inactivity and obesity (deJong et al., 2005), consumption of alcohol and tobacco (Zisman et al., 2006) etc. Certain forms of colorectal cancer are familial, and these include familial adenomatous polyposis (FAP) and hereditary nonpolyposis colorectal cancer (HNPCC), also called Lynch syndrome. FAP is caused by mutations in the APC tumor suppressor gene (TSG) (Wilmink, 1997), while HNPCC is associated with mutations in the genes MSH2 and MLH1 involved in the DNA repair pathway (Haggar et al., 2009). Early diagnosis of colorectal cancer is essential to secure a favorable prognosis.

The Cancer Genome Atlas (TCGA) research network have analyzed the multidimensional COADREAD dataset and found mutational and integrative signatures (The Cancer Genome Atlas Network, 2012), but much remains to be known with respect to the stage-wise progression of colorectal cancer. While the AJCC staging of colorectal cancer is based on histopathology (the TNM staging) (Amin et al., 2017), the research presented here is open to the hypothesis of a molecular basis for cancer progression in discrete stages, and is aimed at identifying the molecular signatures of the discrete progression of colorectal cancer through RNA-Seq transcriptomics. Towards this, we have designed a comprehensive workflow that uses the TCGA data, and follows and extends the protocol introduced in Sarathi and Palaniappan (2019). In addition to discovering the stage-specific biomarkers, this study introduces a new class of biomarkers termed progression-significant DEGs, which are DEGs with a significant monotone trend of expression (up or down). It is noted that the early-stage (i.e, stage-I and stage-II specific) biomarkers could serve in a diagnostic setting and prognostic analysis, whereas progression-significant biomarkers could constitute effective therapeutic targets to halt or reverse the course of cancer (before it can metastasize or reach a point of no return). The findings are benchmarked against the "hub-driver" genes identified in an earlier study from our lab (Palaniappan et al, 2016). Finally a network analysis grounds the findings in a larger context, lending more evidence for the molecular origins of stage-wise discrete cancer progression. This departs and detracts from the thesis advanced in Huo et al (2017) that the stage-wise colorectal cancer progression lacks any molecular basis. The preprocessed and annotated dataset and all the results generated are made available as Supplementary Information at 10.6084/m9.figshare.20489211 .

MATERIALS AND METHODS

The workflow is summarized in Fig. 1 and discussed in detail below. The identification of stage-specific biomarkers follows the computational protocol developed earlier in our lab (Sarathi, A., & Palaniappan, A. 2018).

DATA PREPROCESSING

Normalized and $\log_2$-transformed Illumina HiSeq RNA-Seq gene expression data for Colorectal Adenocarcinoma (COADREAD) processed by the RSEM pipeline (Li and Dewey, 2011) were obtained from TCGA via the firebrowse.org portal (Broad Institute TCGA Genome Data Analysis Center, 2016). 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. To annotate the stage information of the cancer samples, we obtained the corresponding clinical dataset from firebrowse.org 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 cancer staging is encoded in the attribute "pathologic_stage" of the clinical data. The sub-stages (A,B,C) were collapsed into the parent stage, resulting in four stages of interest (I, II, III, IV). We retained a handful of clinical variables related to demographic features, namely age, sex, height, weight, and vital status. Using this merged dataset, we filtered out genes that showed little change in expression across all samples (defined as $\sigma < 1$). We also removed cancer samples that were missing stage annotation (value 'NA' in the "pathologic stage") from our analysis. The data pre-processing was done with R (www.r-project.org) and the final data set was processed through voom in *limma* to prepare for linear modeling (Law et al., 2014).

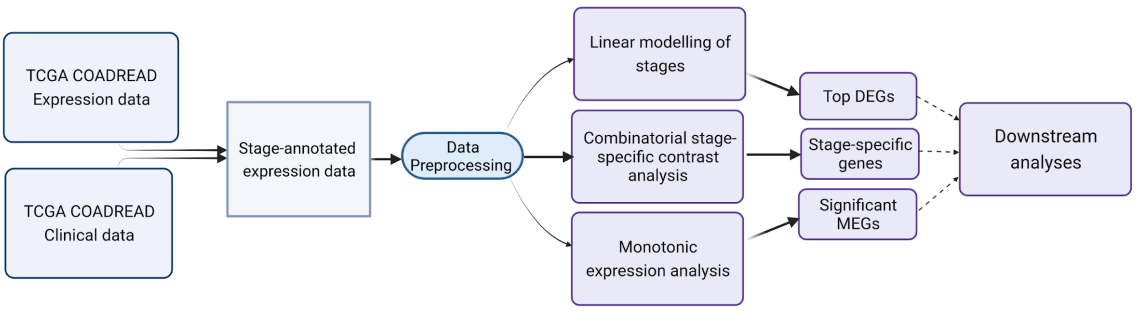

Figure 1. Representation of the study design for the dissection of discrete stage-wise progression of colorectal cancer. Created with BioRender.com

LINEAR MODELING

Linear modeling 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 (Ritchie et al., 2015). The following linear model was fit for each gene's expression based on the design matrix shown in Fig. 2:

$$y = \alpha + \beta_1^{'} x_1 + \beta_2^{'} x_2 + \beta_3^{'} x_3 + \beta_4^{'} x_4 \qquad \ldots. (1)$$

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 stagewise log fold-change

(lfc) coefficients relative to controls. The linear model was subjected to empirical Bayes adjustment to obtain moderated t-statistics (McCarthy and Smyth, 2009). 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 of Hochberg and Benjamini (1990). 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.

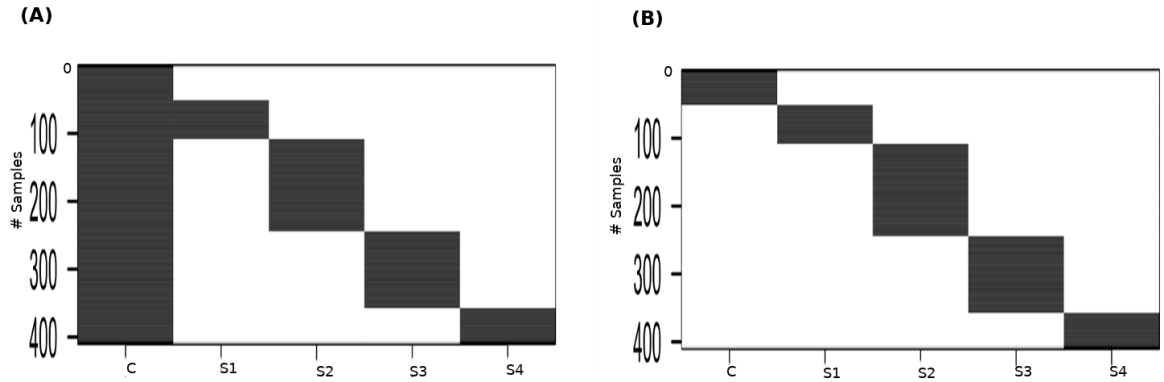

Figure 2. Design matrices (A) for the linear modeling ; and (B) for performing the between-stages contrast analysis.

PAIRWISE CONTRASTS

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

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \qquad \text{….. (2)}$$

where the controls themselves constitute 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 contrast to identify genes differentially expressed with respect to the control. Genes could be differentially expressed in any combination of the stages . In the first pass, we removed genes whose |lfc| < 2 for all stages. For the remaining genes, we identified the stage that showed the highest |lfc| for each gene and assigned the gene as specific to that stage for the rest of our analysis.

Table 1. Coefficients of the contrasts matrix for stage-control modeling of the annotated expression matrix.

Clinical annotation	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

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

(ii)-(iv) P-value of the contrast with respect to other stages < 0.05. Ten such contrasts are possible.

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

To deal with possible sparsity of progression-significant genes specific to any stage,we defined another metric, "pval_pdt", of a given gene in a certain stage as the product of the p_values of its expression contrast in that stage vs each of the other stages (eg, pval_pdt of gene x in stage 1 is (pval(gene x in st1 vs st2))*(pval(gene x in st1 vs st3))*(pval(gene x in st1 vs st4)).

Table 2. Coefficients of the contrasts matrix for between-stages modeling of the annotated expression matrix.

Clinical annotation	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

MONOTONIC EXPRESSION

The linear model in eqn. (1) would not be sufficient to identify genes with an monotonic, stage-ordered trend of expression. But addressing this question is necessary to assess monotonic changes of gene expression with disease progression, which could uncover stage-agnostic expression of progression-significant driver genes. Towards this end, we used a model of gene expression where the cancer stage was treated as a numeric variable:

$$y = aX + b \qquad (3)$$

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

the following patterns of monotonic expression 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 was evaluated using the adj. p-values from the model given by eqn. (3) to yield genes with significant monotonic patterns of expression.

POST PROCESSING

Principal component analysis (PCA) was performed using prcomp in R. We used the rand function to choose 100 random genes. In order to visualize outlier genes that are significant with a large effect size, volcano plots were obtained by plotting the ($-\log_{10}$)-transformed p-value vs. the log fold-change of gene expression. Heat maps of significant stage-specific differentially expressed genes were visualized using R `heatmap` and clustered using `hclust`. Novelty of the identified stage-specific genes was ascertained by screening against the Cancer Gene Census v84 (Futreal et al., 2004). Finally the findings are situated in the network context using STRING db (Szklarczyk et al, 2021).

RESULTS

The TCGA expression data consisted of expression values of 20,502 genes in 428 samples. After the completion of data pre-processing, we obtained a final dataset of expression data for 18,212 genes across 409 samples annotated with the corresponding sample stage. This dataset is made available as 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 those instances in the clinical data that had both height and weight. The average age of onset of COADREAD was around 65 years, and the average BMI was about 29, indicating a possible link of colorectal cancer with ageing and obesity.

Table 3. Statistical summary of clinical meta-data associated with the TCGA COADREAD transcriptome. Number of samples in each gender and vital status categories is also given.

Characteristic		Control	Stage 1	Stage 2	Stage 3	Stage 4	NA	Overall
Number of Samples		51	57	136	113	52	19	428
Age (years)		69.1 ± 14.1	65.8 ± 12.6	66.7 ± 12.9	63.1 ± 13.2	60.6 ± 13.3	65.4 ± 12.2	65.1 ± 13.3
Weight (kg)		79.3 ± 25.3	83.9 ± 19.4	78.3 ± 23.3	81.3 ± 20.2	82.2 ± 17.4	83.6 ± 26.2	80.7 ± 21.5
Height (cm)		169.3 ± 9.5	172.1 ± 11.0	167.0 ± 13.0	169.0 ± 11.0	172.0 ± 11.1	170.9 ± 12.3	169.2 ± 11.7
BMI (kg/m^2)		27.4 ± 7.0	28.5 ± 6.1	29.7 ± 25.1	28.3 ± 6.3	28.7 ± 5.6	28.1 ± 6.0	28.8 ± 15.3
Gender	Male	23	34	72	61	30	11l	231
	Female	28	23	64	52	22	8	197
Vital Status	Alive	44	55	122	100	36	15	372
	Dead	7	2	14	13	16	4	56

Table 4. Sample distribution in the various stages of colorectal cancer in the TCGA dataset.

TCGA Stage	TNM classification	# Cases	
1	T1a N0 M0	56	57
1A	T1b N0 M0	1	
2	T2 N0 M0	18	136
2A	T2 N0 M0	110	
2B	T2 N0 M0	6	
2C	T2 N0 M0	2	
3	T3 N0 M0	9	113
3A	T4 N0 M0	10	
3B	-	59	
3C	-	35	
4	-	27	52
4A	T(any) N1 M0	23	
4B	T(any) N(any) M1	2	
CONTROL	-	51	
NA	-	19	

The dataset was processed through `voom` in *limma* to prepare for linear modeling (Law et al., 2014). At an adjusted p-value cutoff of 10E-5, 9433 genes were significant in the linear modeling, suggesting the existence of a linear trend in their expression across cancer stages. Such an observation was not entirely surprising since one of the hallmarks of cancer phenotype is genome-wide instability, which would likely worsen with progression (Hanahan and Weinberg, 2011). Some top-ranked upregulated genes from the linear modeling include CDH13,KRT80,ETV4,ESM1. It is significant that CDH13 was the top upregulated gene obtained from the linear modeling of hepatocellular carcinoma as well (only after GABRD and PLVAP); such a consistent result might point to a very key role for CDH13 in cancer progression in colorectal and hepatocellular carcinomas, which are both gastrointestinal cancers. The top downregulated genes include OTOP2,OTOP3,AADACL2,DHRS7C. Table 5 shows the lfc for each stage with respect to normal samples of the top ten genes (ranked by adjusted p-value), along with their inferred regulation status. Boxplots of the expression of the top 9 genes indicated a progressive net increase in expression across cancer stages relative to control for up-regulated genes, while repressed expression across cancer stages relative to control was the hallmark of downregulated genes (Fig. 3). A constant trend of regulation across stages underscores the stage-specific basis of cancer progression. It is noted that a given gene might have maximal differential expression in any stage, not necessarily stage 4, and the linear trend does not invariably suggest a monotonic direction of expression across stages (Fig. 4).

Table 5. Stage-wise log-fold changes (lfc) in the expression of the top ten genes from the Linear Modelling analysis. The overall significance of the gene-wise linear model is also indicated, along with the inferred regulation status of the gene in cancer (relative to normal samples). Both upregulated and downregulated genes are seen in the top of the linear model.

Gene	Stage I lfc (β_1)	Stage II lfc (β_2)	Stage III lfc (β_3)	Stage IV lfc (β_4)	Adj. p-val	Regulation Status
CDH3	6.5572	6.4729	6.4325	6.4874	1.06E-156	UP
KRT80	6.8613	6.6695	6.9847	7.2830	4.39E-143	UP
OTOP2	-9.9507	-10.030	-9.9761	-9.9196	4.62E-139	DOWN
AADACL2	-3.3481	-3.4103	-3.3285	-3.3960	4.99E-131	DOWN
ETV4	5.6165	5.5937	5.5175	5.8992	8.28E-131	UP
ESM1	5.7276	5.9611	5.9339	6.4049	2.56E-130	UP
DHRS7C	-3.4279	-3.5170	-3.5209	-3.5196	3.14E-130	DOWN
OTOP3	-5.3795	-5.2544	-5.1438	-5.1531	1.80E-125	DOWN
JUB	3.1785	3.1473	3.1536	3.0750	7.78E-102	UP
MTHFD1L	2.6099	2.5692	2.5300	2.5766	2.10E-100	UP

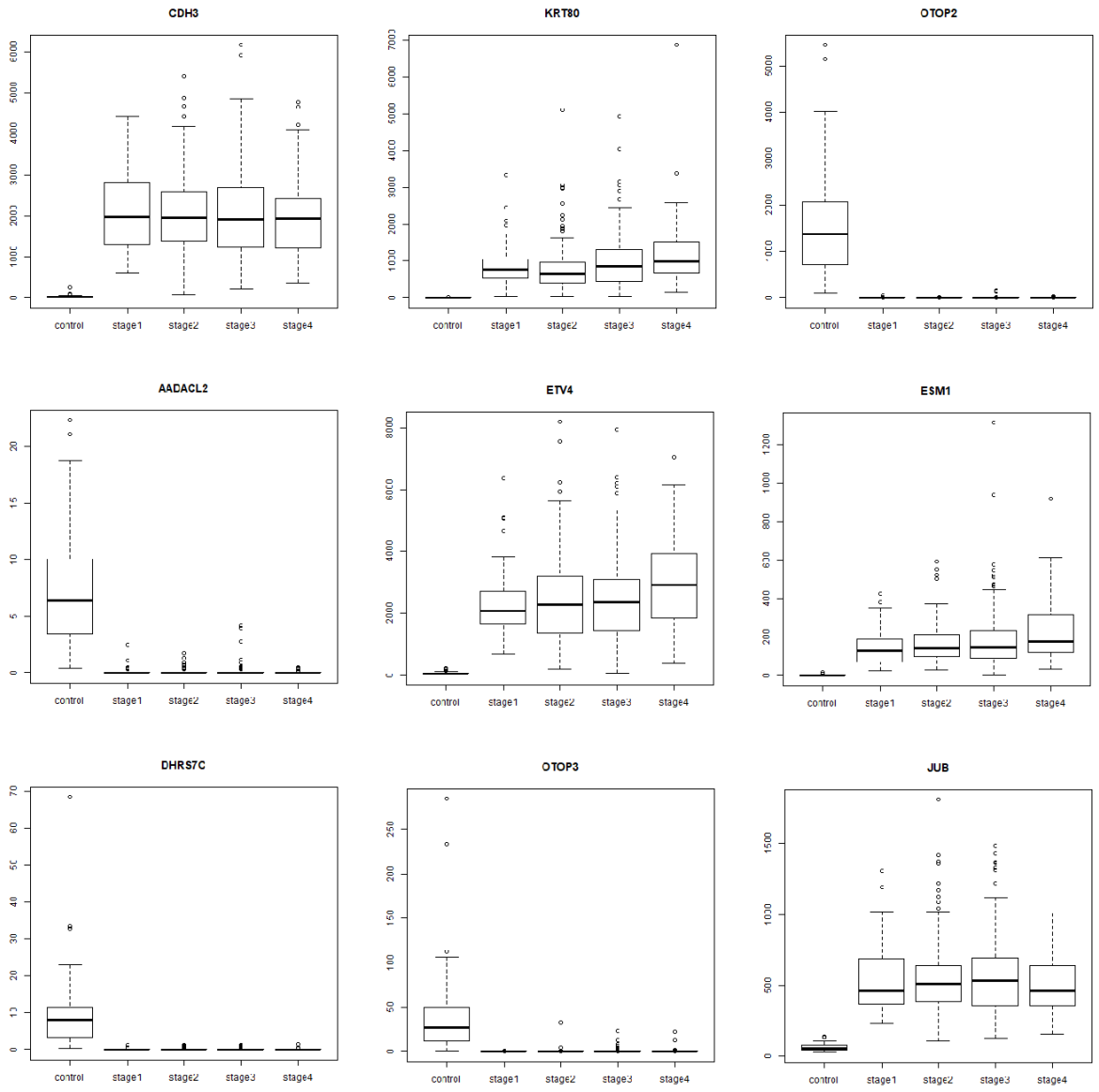

Figure 3. Expression trends of the top 9 DEGs from the linear modeling.

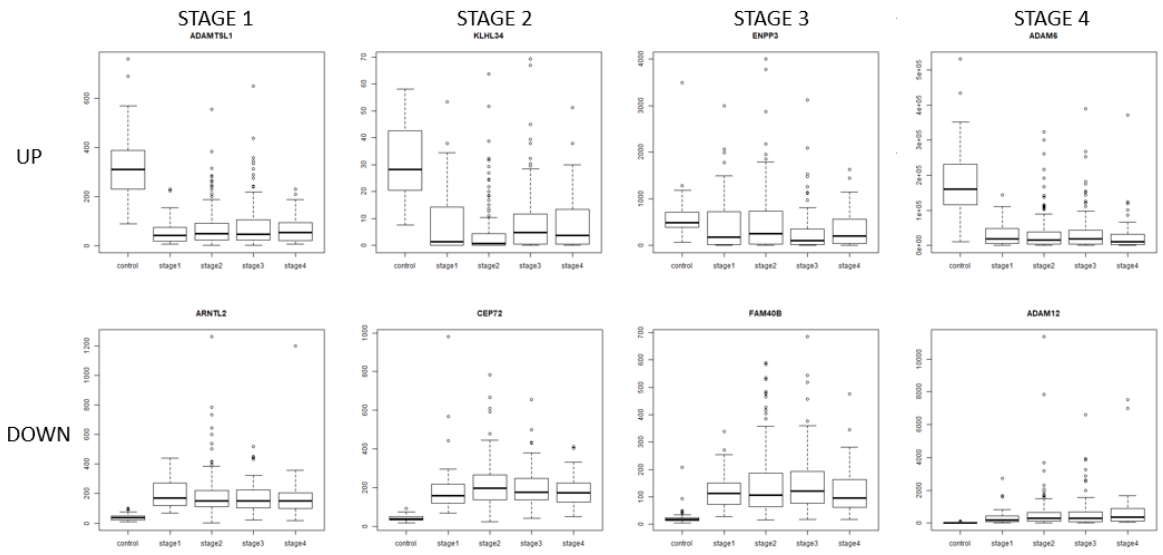

Figure 4. Expression trends of genes across cancer stages to illustrate the salience of stage. Two expression trends, one upregulated and the other downregulated with respect to controls, to illustrate the salience of each stage are shown.

The samples were visualized using a PCA of the top 100 genes from the linear model (Fig. 5 (A)). A clear separation of the controls and the cancer samples could be seen, suggesting the considerable extent of differential expression of these genes in cancer samples. Hence linear modeling yields cancer-specific genes versus normal controls, and the results for all the genes, including the top 100, are provided in order in Supplementary File S2. The PCA plot of 100 randomly sampled genes (Fig. 5 (B)) failed to show any separation whatsoever of the cancer and control samples, highlighting the significance of linear models in the analysis of cancers.

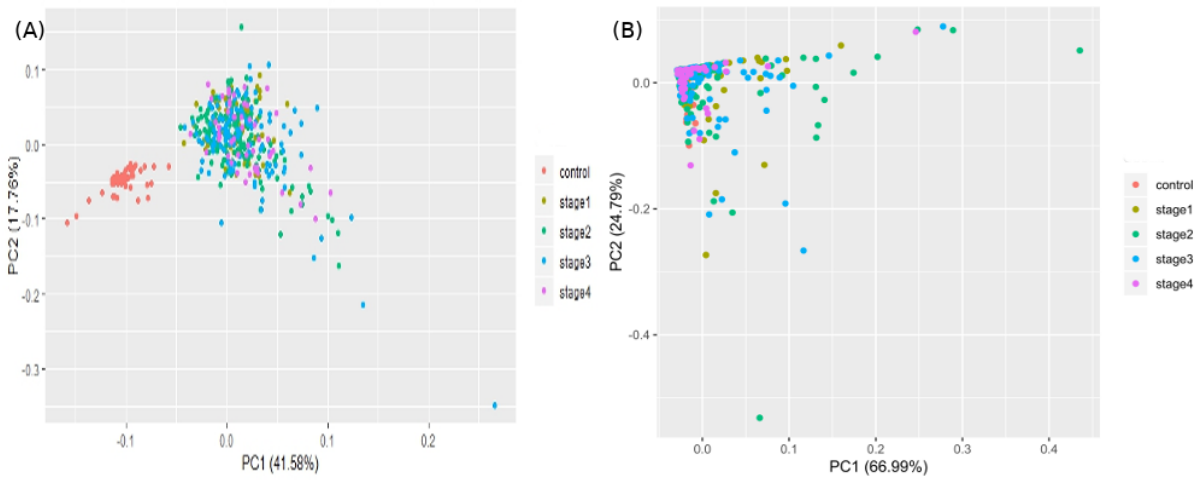

Figure 5. Plot of the top two principal components of top 100 genes from the linear model genes (left) and 100 random genes (right)

The results from the linear modeling were in contrast with those obtained by Huo *et al*. (2017) and were most likely driven by the inclusion of 51 controls (normal samples) in our study (and exclusion of the same in the Huo *et al.* 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. In the first pass, we eliminated 15,970 genes with |lfc|<2 in all stages. To identify the stage-specific differentially expressed genes, we applied the first contrast matrix (Table 1). We then binned all the genes into different partitions of varying sizes, to obtain stage-specific genes (Fig. 6). To establish their significance, we applied the second contrast (Table 3) to identify genes that passed all the filter criteria described in the Methods section. Such genes were retained as significant stage-specific differentially expressed genes. This process yielded 71 stage-I specific, 2 stage-II specific, 0 stage-III specific and 59 stage-IV specific genes (Supplementary File S3).

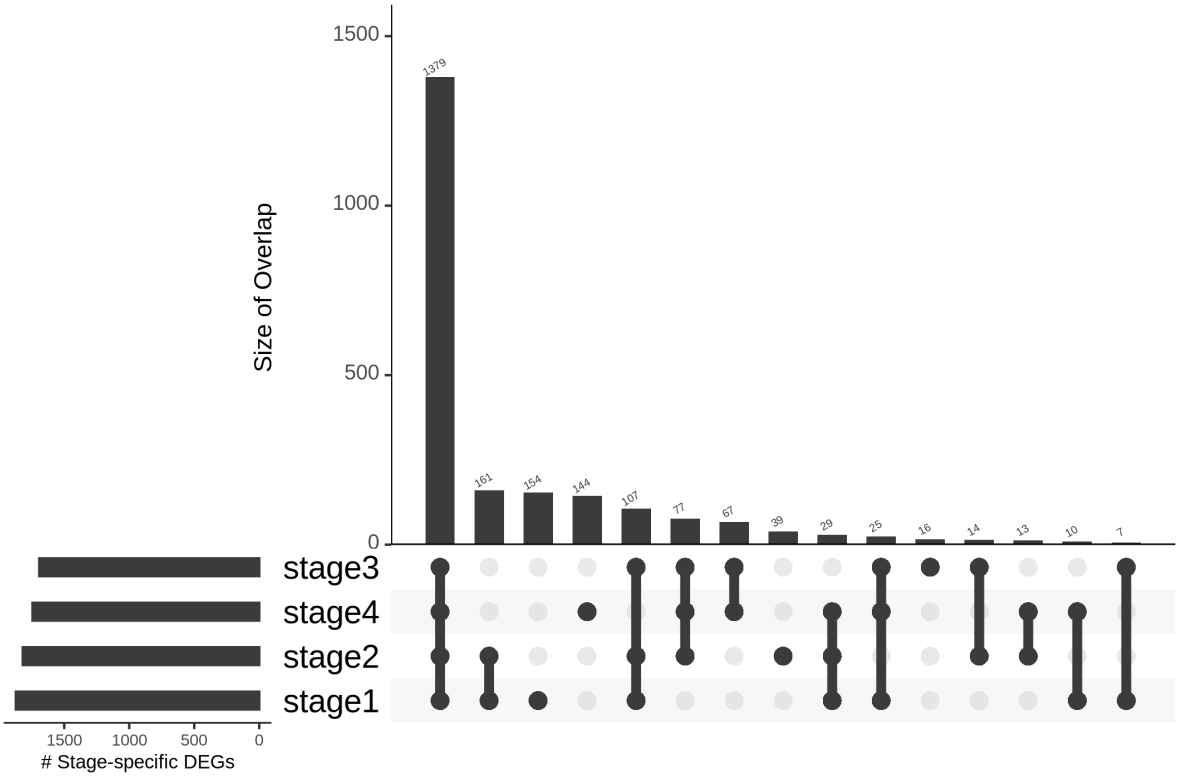

Figure 6. Distribution of genes based on stage-specificity. Of the 2242 DEGs, 1379 appear significant in *all* the stages. It can be clearly seen that the early-stages (stages 1 and 2) share fewer DEGs with the late-stages (stages 3 and 4), flagging extra factors necessary for cancer progression to metastasis.

In view of the sparsity of genes passing the filters for stages 2 and 3, we applied the pval_pdt, described in the Methods section, and extracted the top 10 genes for each stage. For stages 1 and 4, all these top 10 genes figured in the 71 and 59 genes that had been identified as significant stage-specific DEGs, respectively. For stage 2, we took the 2 genes that passed the filtering and appended genes with the lowest pval_pdt to obtain 10 genes. For stage 3, we used the 17 genes with pval_pdt < 0.125E-3. The top 10 genes from each stage are shown in Table 6, and the entire set of 157 stage-specific DEGs are presented in Supplementary File S3. It is significant that GABRD emerges as a stage-IV specific gene in COADREAD, reinforcing its identification as a stage-IV specific gene in LIHC (Sarathi and Palaniappan, 2019) and suggesting its driver role in metastasis of other gastrointestinal cancers as well.

Table 6. Top 10 Stage-specific DEGs in each stage, ordered by significance.

S.No.	Stage 1	Stage 2	Stage 3	Stage 4
1	CALB2	FADS6	PIGR	UPK2
2	TMEM59L	EEF1A2	MLXIPL	HOTAIR
3	JPH3	KLHL34	TUBAL3	LY6G6C
4	STAC2	DMRT3	COMP	C6orf15
5	NKX3.2	GREM2	SLC26A9	DLX3
6	UCHL1	CCBP2	CES3	CST6
7	KCNG1	ADCY5	TRY6	VGLL1
8	CRLF1	PLAC2	HABP2	GABRD
9	C5orf23	GPC5	NAT2	DKK1
10	FBXO27	LPHN3	HES5	TMEM40

A heatmap of the lfc expression of stage-specific genes across different stages was visualized (Fig.7) and revealed systematic variation in expression relative to control on a gradient from blue (downregulated) to red (overexpressed). The map was clustered on the basis of differential expression (i.e, |lfc|) across stages and showed an early-stage (stages 1 & 2) vs late-stage (stages 3 & 4) separation, arguing for the role of progression-significant genes in driving colorectal cancer. The clustering of these 40 genes was also visualized using a hierarchical clustering algorithm to yield interesting observations (Fig. 8).

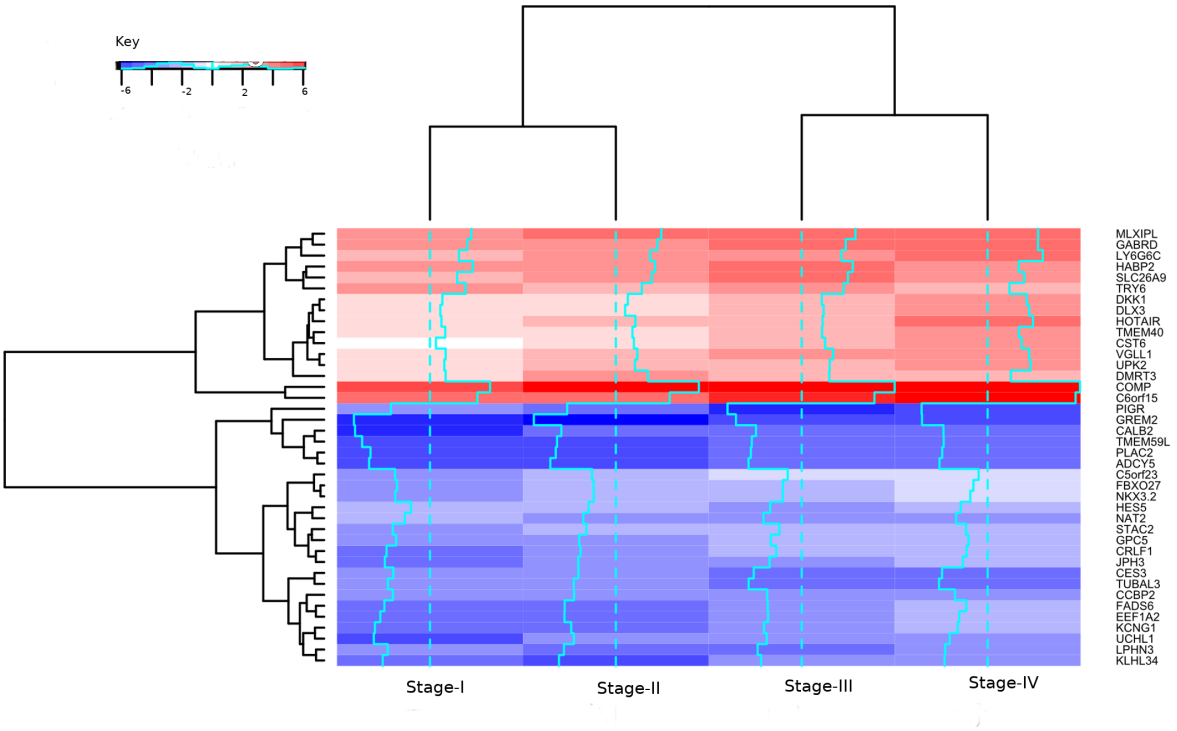

Figure 7. Heatmap of log-fold change (lfc) of top 40 genes. Stage-specific genes express maximal salience in one of the stages. It is striking that all the ten stage-IV specific show progressive upregulation (for e.g, GABRD).

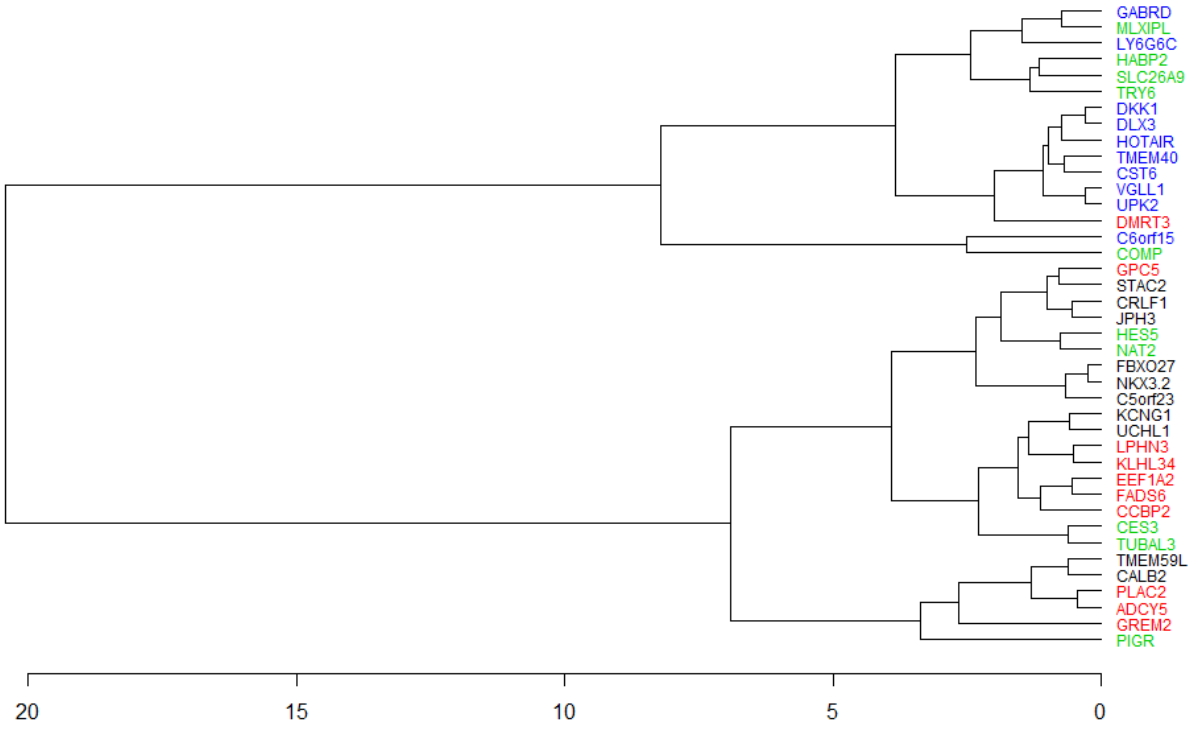

Figure 8. Clustering of 40 stage-specific genes (stage1 – black, stage2 – red, stage3 – green, stage4 – blue). Stage-I and Stage-IV genes do not intermingle in the clustering pattern. There is substantial co-clustering of stages I-II, and stages III-IV. DMRT3 is the only stage-II specific gene to co-cluster with stage-IV specific genes.

We observed that a lot of the stage 4 genes are "oncogenes", steadily over-expressed in the cancer phenotype unto metastasis, whereas most of the early-stage (stages 1 and 2) genes are "tumor suppressor genes", which are differentially down-regulated in the cancer phenotype. This has implications for the progression pathway of colorectal cancer – initially disabling the damage-control mechanisms innate to the cell and progressively spiraling out of control, and could suggest a general hypothesis for the progression of cancers that would need experimental support.

The top 200 genes from the numeric model (eqn.3) are presented in Supplementary File S4. With an adj.p-value <0.05, the monotonic analysis yielded 1389 upregulated and 555 downregulated genes, of which 1058 MEGs are significantly monotonically expressed (noted in Supplementary File S5). Examining the overlap of these significant MEGs with stage-specific DEGs yielded 19 progression-significant driver genes (Table 7). It is especially significant that ten stage-4 specific DEGs are also MEGs, underscoring implications for driving the progression of colorectal cancer to metastasis if left unchecked. The intersection between the top 200 genes from the numeric model and the significant monotonically expressed genes yielded 18 genes (presented in Supplementary File S6).

Table 7. Progression-significant driver genes, obtained by the overlap of significant MEGs with stage-specific DEGs. 19 genes sorted by the corrected significance from the numeric model are shown. Status indicates the direction of the monotone trend across the stages.

SNo.	Gene	Stage	Status	Adj.p-val
1	PIGR	III	DOWN	1.23E-26

2	GREM2	II	UP	4.81E-24
3	SOX10	I	UP	3.59E-19
4	GHR	I	UP	4.42E-15
5	LOC283867	IV	UP	2.63E-10
6	SCG2	I	UP	4.13E-08
7	FGF18	IV	UP	5.35E-07
8	TH	IV	DOWN	6.42E-07
9	ATOH1	IV	UP	7.38E-07
10	FAM124A	I	UP	2.86E-06
11	SSTR2	I	UP	9.97E-06
12	GJC3	IV	UP	1.20E-05
13	CCDC19	III	DOWN	4.34E-05
14	COL9A3	IV	UP	8.31E-05
15	SERPINE1	IV	DOWN	1.00E-04
16	DLX3	IV	DOWN	1.56E-03
17	SLCO1A2	IV	UP	2.23E-03
18	HOXC11	IV	UP	2.82E-03
19	C1orf180	I	UP	1.84E-02

DISCUSSION

To clarify the sum of findings from our modeling studies, we began by looking at the canonical CRC drivers, APC and MSH2, which are both implicated in familial CRC. APC and MSH2 are both significantly differentially expressed (adj.p-values ~ 7.35e-13 and 2.06e-18 respectively). The expression patterns of these two genes (Fig.9) showed that APC was downregulated in the cancer phenotype, flagging its key role as a known TSG .

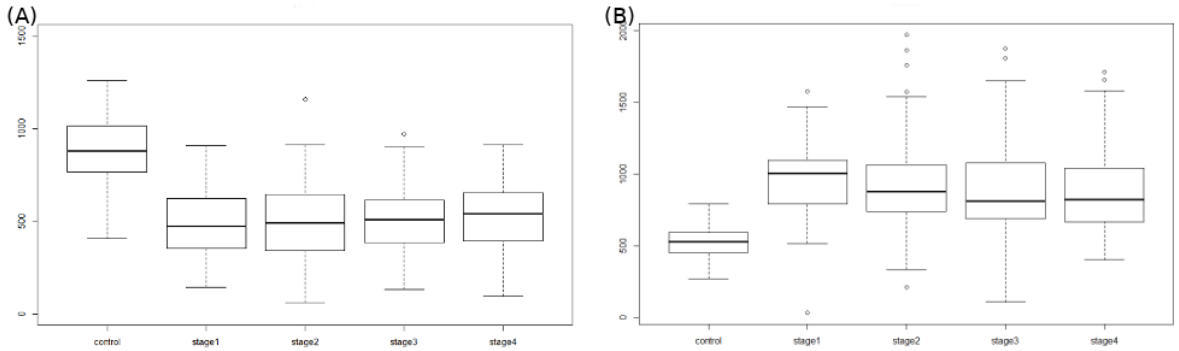

Figure 9. Expression trends of (A) APC and (B) MSH2, known genetic factors of familial CRC

We then looked at the hub-driver genes identified in a previous study of CRC network analyses (Palaniappan et al 2016), and found that GRIN2A and EIF2B5 were significantly differentially expressed in the cancer samples(adj.p-values ~ 2.14e-37 and 2.32e-13, respectively). GRIN2A is a TSG with least expression in stage 2 (Fig. 10(A)), reinforcing its role as a hub driver gene for stage 2

progression. EIF2B5 is an oncogene with maximal expression in stage 3 (Fig. 10(B)), again according with its identified role as a major hub driver gene for progression to advanced stages of colorectal cancer.

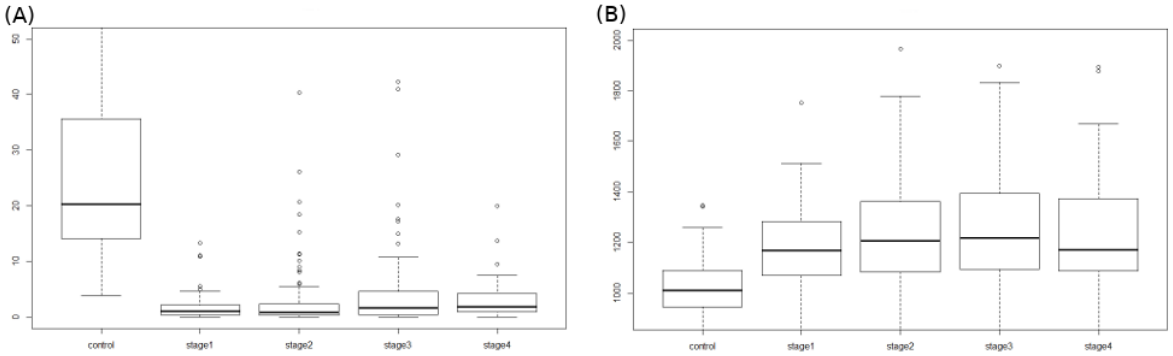

Figure 10. Expression trends of Candidate Hub-Driver genes (A) GRIN2A and (B) EIF2B5.

An analysis of the top genes from our linear model uncovered certain interesting observations. The top gene hit, CDH3 (Cadherin 3 or P-Cadherin), has been found to be overexpressed in a great majority of Pancreatic Ductal Adenocarcinomas (PDACs) (Taniuchi et al. 2005), lending support to its key role in gastrointestinal cancers. Further, hypomethylation of the CDH3 promoter has been found in addition to (and the cause of) increased expression of CDH3 in both Breast Cancer (Paredes et al. 2005) and Advanced Colorectal Cancer (Hibi et al. 2009). This can be due to the fact that over-expression of P cadherin leads to high motility of cells, which enables the cancer cells to metastasize.

There is emerging evidence for the role of KRT80 in head and neck squamous cell carcinoma (Zhao et al, 2022), but it is not a known cancer driver (https://www.intogen.org/search?gene=KRT80). The gene OTOP2 has been identified as a TSG, as it was significantly downregulated in the cancer phenotype. Another independent study also found that wild type p53 regulated OTOP2 transcription in cells, and increased levels of OTOP2 suppressed tumorigenesis *in vitro.* (Qu et al. 2018). OTOP3 belongs to the same family of otopetrin proton channels, but there is no published evidence for its role in any cancer (https://www.intogen.org/search?gene=OTOP3).

AADACL2 is not a known cancer driver, but there is evidence for its role in a comorbid breast-colorectal cancer phenotype (Pande M et al 2018). ETV4, another top candidate in our linear model, has shown significant promise as a therapeutic target. A previous study found that ETV4 knockdown in metastatic murine prostate cancer cells abrogates the metastatic phenotype but does not affect tumor size (Aytes et al. 2013). According to our model, ETV4 shows maximal expression in stage 4 and is concordant with a molecular basis for cancer stages. ETV4 is also a designated cancer gene in the COSMIC census (Forbes et al. 2016).

ESM1 was found to be clearly overexpressed in clear cell renal cell carcinoma (Leroy et al. 2010), and is also one among the 59 stage-4 specific genes from our study. DHRS7C has been recently implicated in signaling pathways involved in glucose metabolism (Ruiz et al, 2018). It exerts its effects via mTORC2, a complex known to be at the heart of metabolic reprogramming (Masui et al, 2014). Mysteriously DHRS7C was seen downregulated in colorectal cancer, given that its upregulation is necessary for glucose uptake. These observations merit experimental investigations to ascertain the precise nature of the molecular biology in question.

Some studies reveal that the LIM-domain-containing JUB serves as an oncogene in CRC by promoting the epithelial-mesenchymal transition (EMT), a critical process in the metastatic transition (Liang et al. 2014). The gene MTHFD1L coding for methylenetetrahydrofolate dehydrogenase 1–like is significantly overexpressed in the colorectal cancer phenotype. Studies show that MTHFD1L contributes to the production and accumulation of NADPH to levels that are sufficient to combat oxidative stress in cancer cells. The elevation of oxidative stress through MTHFD1L knockdown or the use of methotrexate, an antifolate drug, sensitizes cancer cells to sorafenib, a targeted therapy for hepatocellular carcinoma (Lee et al. 2017).

We also found that 13 of the top 200 genes from the linear model were known cancer genes in the Cancer Gene Census (CGC; cancer.sanger.ac.uk) (Table 8). Two genes, MACC1 and SALL4 are already documented genes for colorectal cancer, testifying to our analysis. HSP90AB1 had been earlier identified as a top MEG in HCC (Sarathi and Palaniappan, 2019).

Table 8. Summary of hits in the top 200 of the linear model that are documented with evidence in the CGC. These are known cancer driver genes, of which MACC1 and SALL4 are also known colorectal cancer driver genes. In the case of FAM135B, FEV, CBFB, and CTNND2, the regulatory status inferred here is at odds with the documented cancer role, and could point to anomalous contra-regulation inviting experimental investigation.

Gene Symbol	**Illustrative tumors**	**Documented role**	**Status**
ETV4	Ewing sarcoma, prostate carcinoma	oncogene, fusion	UP
FAM135B	small cell lung cancer, esophageal cancer	oncogene	DOWN
FEV	Ewing sarcoma	oncogene, fusion	DOWN
CDH10	melanoma	TSG	DOWN
CBFB	acute myeloid leukemia	TSG, fusion	UP
KIAA1549	pilocytic astrocytoma	fusion	UP
HSP90AB1	non-Hodgkin's lymphoma	fusion	UP
MACC1	hepatocellular carcinoma, ***CRC***	oncogene	UP
SET	T-cell acute lymphoblastic leukemia	oncogene, fusion	UP
MET	papillary renal, head-neck squamous cell	oncogene	UP
SALL4	***CRC***, breast cancer, prostate cancer, glioblastoma, melanoma	oncogene	UP
PHOX2B	neuroblastoma	TSG	DOWN
CTNND2	prostate adenocarcinoma, GIST (gastrointestinal stromal tumor)	oncogene	DOWN

Stage1 – Specific DEGs

The gene CALB2 and TMEM59L cluster together in the stage-specific dendrogram. Both of them show the lowest expression in stage one, which leads to the hypothesis that their expression is required to prevent tumorigenesis. This is supported by evidence in literature, which observes that cells in which CALB2 is silenced do not respond to 5-flourouracil, a popular treatment for CRC, indicating that CALB2 expression is essential for 5-flourouracil induced apoptosis (Stevensol et al. 2011). Another study found a heterozygosity in SNP513 of Intron 9 of the gene CALB2 might be a predictive marker for CRC (Vonlanthen et al. 2007). Another study noted that increased TMEM59L expression was a pro-apoptotic indicator of cell death during oxidative stress in neuronal cells (Zheng et al. 2017)

Stage2 – Specific DEGs

KLHL34 was found to be hypermethylated in Locally Advanced Rectal Cancer, and knockdown of KLHL34 lowered colony formation, increased cytotoxicity, and increased radiation induced caspase 3 activity in LoVo cells (Ha et al. 2015)
CCBP2, encoding the Chemokine decoy receptor D6, has an inhibitory effect on breast cancer malignancies due to its action to sequester pro-malignant chemokines (Yang et al. 2013). The lncRNA PLAC2 induces cell cycle arrest in glioma by binding to Ribosomal Protein RP L36 in a mechanism involving STAT1 (Hu et al. 2018). GPC5 was found to be overexpressed in the lung cancer phenotype (Li et al. 2011), in lymphoma (Yu et al 2003), and in Gastric Cancer (Wang et al. 2015). The work by Wang et al also showed that the overexpression of miR-217 impaired GPC5-induced promotion of proliferation and invasion in GC cells.

Stage3 – Specific DEGs

Copy number polymorphisms of TRY6 gene have been found in Breast Cancer (Wagner et al. 2007). HABP2 gene overexpression has been observed in lung adenocarcinoma and has been proposed as a novel biomarker for the same (Wang et al. 2002).

Stage4 – Specific DEGs

The lincRNA HOTAIR was found to be significantly overexpressed in HCC, and a potential biomarker for lymph node metastasis in HCC (Geng et al. 2011), and later implicated in different cancers (Hajjari & Salavaty, 2015). Another widely cited study (Gupta et al, 2010) showed that enforced HOTAIR gene expression in epithelial cancer cells induces chromatin reprogramming, and an increased metastatic state, and inhibition of HOTAIR inhibits cancer invasiveness. These accounts of the role of HOTAIR in metastasis accord with our findings that HOTAIR is a stage-4 specific biomarker. GWAS analysis identified a strong association of C6orf15 with occurrence of follicular lymphoma (Skibola et al. 2009). Promoter methylation of cell free DNA of the CST6 gene was found to be a potential plasma biomarker for Breast Cancer (Chimonidou et al. 2013)
Expression of VGLL1 and its intronic miRNA miR-934 are associated with sporadic and BRCA1-associated triple negative basal-like breast carcinomas (Castilla et al. 2014). Expression of DKK1, an inhibitor of osteoblast differentiation, was found to be associated with the presence of bone lesions in patients with multiple myeloma (Tian et al. 2003). TMEM40 has been found to be a potential biomarker in patients with Bladder cancer, serving as an oncogene and a possible therapeutic target. (Zhang et al. 2018)

Network validation

Stage-wise network analysis of colorectal cancer progression has shed light on certain genes potentially underlying progression (Rahiminejad et al, 2022). The strength of the computational evidence for the biomarkers identified herein urged a network analysis to examine the findings in a larger context. We were particularly interested in two types of monotonically enriched pathways. The intersection between the sets of all stage-specific biomarkers and the significant MEGs might indicate certain genes essential to the pathophysiology of colorectal cancers. These 17 genes (barring the un-annotatable C1orf180 and LOC283867) were imported into STRING and an interaction network was constructed with 50 interactors in the first shell and 10 interactors in the second shell. This yielded a PPI with 157 edges with an extremely significant enrichment p-value ~ 2.07E-14 (Fig 11(A)). An analysis of the KEGG (Kanehisa et al, 2016) enrichment for this reconstructed network shows

significant enrichment for signaling pathways known to be involved in cancer (ErbB, JAK-STAT, PI3K-Akt, neurotrophin, pluripotency of stem cells, Ras, T-cell receptor), pluripotency of stem cells, checkpoint pathways, cancer-specific pathways (CML, AML, prostate, endometrial, non-small cell lung cancer, breast cancer, glioma, renal cell), gastrointestinal cancers (gastric, hepatocellular, colorectal), and growth hormone action, all significant at p-value < 4.26E-10.

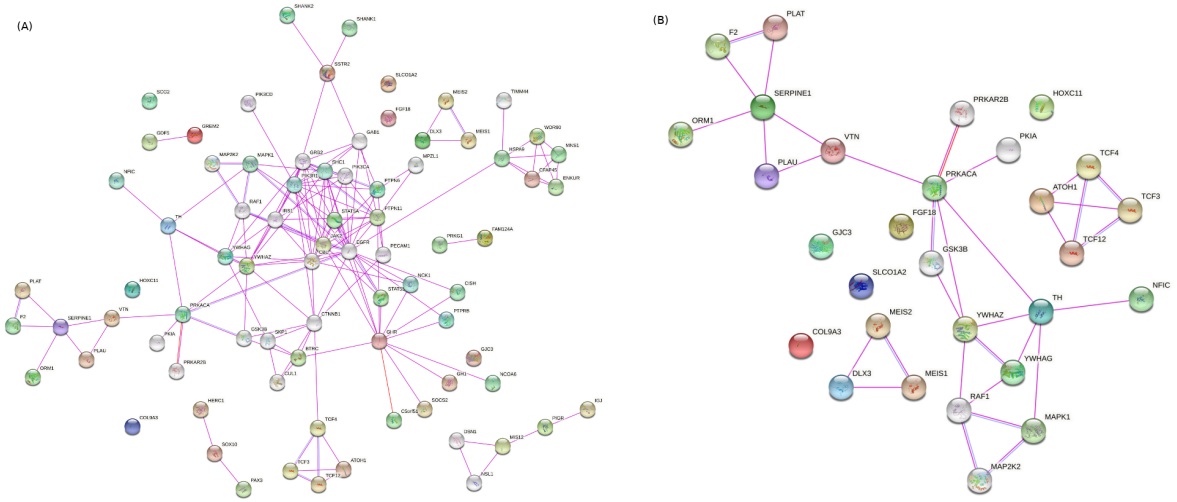

Figure 11. Network reconstruction of perturbed pathways with monotonic expression enrichment. (A) Seed set of all stage-specific DEGs and significant MEGs; and (B) seed set of stage-IV specific DEGs and significant MEGs. Networks were constructed using evidence from known interactions (curated databases, experimentally determined) or predicted from gene neighborhood, gene fusions or gene co-occurrence. Colored nodes indicate query proteins and first shell of interactors, whereas white nodes indicate second shell of interactors.

A more selective set of putative COADREAD driver genes constitutes the intersection between stage-IV specific biomarkers and significant MEGs. These comprise the core set of nine genes driving the cancer inexorably to metastasis. This set was used to construct an interaction network, with 20 interactors in the first shell and 5 interactors in the second shell. This yielded a PPI with 33 edges indicating an enrichment p-value ~ 0.012 (Fig.11(B)). A Gene Ontology (Ashburner, 2000) analysis of this reconstructed network shows enrichment for those components involved in migration, adhesion and invasion (pseudopodium, focal adhesion, blood microparticle, collagen-containing extracellular matrix) at p-value < 0.05. This is striking evidence for the involvement of these biomarkers in driving the metastatic transition. An analysis of the Reactome (Sidiropoulos, 2017) enrichment for this network shows significant disruption of canonical cancer signaling pathways including the MAPK pathway.

Isolated nodes in the above networks included GJC3, SLCO1A2, COL9A3, SCG2, FGF18 and HOXC11. Outliers (not in the giant connected component) in the above networks included FAM124A, GREM2, DLX3, and SOX10. It is noteworthy that the Cancer Genome Atlas Network observed SOX9 as a novel gene with significant recurrent mutations in COADREAD (2012). It is notable that PIGR was also in its own component and could constitute a novel target for upregulation in CRC. A drug-repurposing analysis with the DrugGeneBadger (Wang et al, 2019) yielded drugs (small molecules with q-values < 0.05) for most of these targets indicating that the expression of these identified biomarkers could be altered pharmacologically. It revealed that curcumin is effective against at least five of these targets, and piperlongumine against four of these targets. FAM124A emerged as the only identified

biomarker with no drug able to modulate its expression. Other biomarkers with few drugs targeting their expression included GJC3 and DLX3.

CONCLUSIONS

We have executed multiple workflows with the singular aim of discovering stage-specific signatures of colorectal cancer progression. A data-driven hypothesis-free analysis has uncovered hitherto unreported entirely strange biomarkers towards the early diagnosis of colorectal cancer. In addition to significant stage-specific biomarkers, this work has uncovered stage-agnostic progression-significant monotonically expressed biomarkers. Modulating the expression of progression-significant biomarkers (for e.g, by halting the overexpressed ones or activating the expression of downregulated ones) could turn out to be a robust strategy to arresting the progression of colorectal cancer to the advanced stages including metastasis. The various classes of biomarkers identified have been scrutinized with further analysis and existing literature evidence, and found to be potentially translationally meaningful. Experimental proof could provide the validation necessary to translate these findings to the clinical settings. In addition to advancing novel biomarker signature panels, this computational study has contributed significant evidence for the genetic factors driving the discrete stage-wise progression of colorectal cancer. The study design also provides a framework of broad generality for the investigation of cancers and other progressive conditions.

ACKNOWLEDGEMENTS

This work was supported on grant DST-SERB EMR/2017/000470 to A.P.

REFERENCES

1. Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, *65*(2), 87-108.

2. Haggar, F. A., & Boushey, R. P. (2009). Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, *22*(4), 191.

3. Willett WC.(2005). Diet and cancer: an evolving picture. JAMA;293(2):233–234

4. de Jong AE, Morreau H, Nagengast FM, et al.(2005) Prevalence of adenomas among young individuals at average risk for colorectal cancer. Am J Gastroenterol;100(1):139–143

5. Zisman AL, Nickolov A, Brand RE, Gorchow A, Roy HK.(2006)Associations between the age at diagnosis and location of colorectal cancer and the use of alcohol and tobacco: implications for screening. Arch Intern Med 2006;166(6): 629–634

6. Wilmink ABM.(1997) Overview of the epidemiology of colorectal cancer. Dis Colon Rectum ;40(4):483–493

7. Amin, M.B., Greene, F.L., Edge, S.B., Compton, C.C., Gershenwald, J.E., Brookland, R.K., Meyer, L., Gress, D.M., Byrd, D.R. and Winchester, D.P., (2017). 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: a cancer journal for clinicians, 67(2), pp.93-99.

8. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337 (2012). https://doi.org/10.1038/nature11252

9. Sarathi A, Palaniappan A. Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma. (2019) BMC Cancer 19(1): 663 DOI: 10.1186/s12885-019-5838-3

10. Palaniappan A, Ramar K, Ramalingam S (2016) Computational Identification of Novel Stage-Specific Biomarkers in Colorectal Cancer Progression. PLoS ONE 11(5): e0156665. doi:10.1371/journal.pone.0156665

11. Huo T, Canepa R, Sura A, Modave F, Gong Y (2017) Colorectal cancer stages transcriptome analysis. PLoS ONE 12(11): e0188697. https://doi.org/10.1371/journal.pone.0188697

12. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(1):323.

13. 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.

14. 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.

15. 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.

16. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 2009;25:765–71.

17. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med. 1990;9:811–8

18. 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.

19. Szklarczyk D*, Gable AL*, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P‡, Jensen LJ‡, von Mering C‡. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021 Jan 8;49(D1):D605-12

20. Taniuchi, K., Nakagawa, H., Hosokawa, M., Nakamura, T., Eguchi, H., Ohigashi, H., ... & Nakamura, Y. (2005). Overexpressed P-cadherin/CDH3 promotes motility of pancreatic cancer cells by interacting with p120ctn and activating rho-family GTPases. *Cancer research*, *65*(8), 3092-3099.

21. Paredes, J., Albergaria, A., Oliveira, J. T., Jerónimo, C., Milanezi, F., & Schmitt, F. C. (2005). P-cadherin overexpression is an indicator of clinical outcome in invasive breast carcinomas and is associated with CDH3 promoter hypomethylation. *Clinical cancer research*, *11*(16), 5869-5877.

22. Hibi, K., Goto, T., Mizukami, H., Kitamura, Y. H., Sakuraba, K., Sakata, M., ... & Sanada, Y. (2009). Demethylation of the CDH3 gene is frequently detected in advanced colorectal cancer. *Anticancer research*, *29*(6), 2215-2217.

23. Yixuan Zhao, Xin Huang, Zewei Zhang, Haizhou Li, Tao Zan, "The Long Noncoding Transcript HNSCAT1 Activates KRT80 and Triggers Therapeutic Efficacy in Head and Neck Squamous

Cell Carcinoma", Oxidative Medicine and Cellular Longevity, vol. 2022, Article ID 4156966, 19 pages, 2022. https://doi.org/10.1155/2022/4156966

24. Qu, H., Su, Y., Yu, L., Zhao, H., & Xin, C. (2019). Wild-type p53 regulates OTOP 2 transcription through DNA loop alteration of the promoter in colorectal cancer. *FEBS open bio*.

25. Pande M, Joon A, Brewster AM, Chen WV, Hopper JL, Eng C, Shete S, Casey G, Schumacher F, Lin Y, Harrison TA, White E, Ahsan H, Andrulis IL, Whittemore AS, John EM, Ko Win A, Makalic E, Schmidt DF, Kapuscinski MK, Ochs-Balcom HM, Gallinger S, Jenkins MA, Newcomb PA, Lindor NM, Peters U, Amos CI, Lynch PM. Genetic susceptibility markers for a breast-colorectal cancer phenotype: Exploratory results from genome-wide association studies. PLoS One. 2018 Apr 26;13(4):e0196245. doi: 10.1371/journal.pone.0196245. PMID: 29698419; PMCID: PMC5919670.

26. Aytes, A., Mitrofanova, A., Kinkade, C. W., Lefebvre, C., Lei, M., Phelan, V., ... & Shen, M. M. (2013). ETV4 promotes metastasis in response to activation of PI3-kinase and Ras signaling in a mouse model of advanced prostate cancer. *Proceedings of the National Academy of Sciences*, *110*(37), E3506-E3515.

27. Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., ... & Stefancsik, R. (2016). COSMIC: somatic cancer genetics at high-resolution. *Nucleic acids research*, *45*(D1), D777-D783.

28. Leroy, X., Aubert, S., Zini, L., Franquet, H., Kervoaze, G., Villers, A., ... & Lassalle, P. (2010). Vascular endocan (ESM-1) is markedly overexpressed in clear cell renal cell carcinoma. *Histopathology*, *56*(2), 180-187.

29. Ruiz, A., Dror, E., Handschin, C. et al. Over-expression of a retinol dehydrogenase (SRP35/DHRS7C) in skeletal muscle activates mTORC2, enhances glucose metabolism and muscle performance. Sci Rep 8, 636 (2018). https://doi.org/10.1038/s41598-017-18844-3

30. Masui K, Cavenee WK, Mischel PS. mTORC2 in the center of cancer metabolic reprogramming. Trends Endocrinol Metab. 2014 Jul;25(7):364-73. doi: 10.1016/j.tem.2014.04.002. Epub 2014 May 21. PMID: 24856037; PMCID: PMC4077930

31. Liang, X. H., Zhang, G. X., Zeng, Y. B., Yang, H. F., Li, W. H., Liu, Q. L., ... & Yu, L. N. (2014). LIM protein JUB promotes epithelial–mesenchymal transition in colorectal cancer. *Cancer science*, *105*(6), 660-666.

32. Lee, D., Xu, I. M. J., Chiu, D. K. C., Lai, R. K. H., Tse, A. P. W., Li, L. L., ... & Wong, C. M. (2017). Folate cycle enzyme MTHFD1L confers metabolic advantages in hepatocellular carcinoma. *The Journal of clinical investigation*, *127*(5), 1856-1872.

33. Stevenson, L., Allen, W. L., Proutski, I., Stewart, G., Johnston, L., McCloskey, K., ... & Johnston, P. G. (2011). Calbindin 2 (CALB2) regulates 5-fluorouracil sensitivity in colorectal cancer by modulating the intrinsic apoptotic pathway. *PLoS One*, *6*(5), e20276.

34. Stevenson, L., Allen, W. L., Proutski, I., Stewart, G., Johnston, L., McCloskey, K., ... & Johnston, P. G. (2011). Calbindin 2 (CALB2) regulates 5-fluorouracil sensitivity in colorectal cancer by modulating the intrinsic apoptotic pathway. *PLoS One*, *6*(5), e20276.

35. Vonlanthen, S., Kawecki, T. J., Betticher, D. C., Pfefferli, M., & Schwaller, B. (2007). Heterozygosity of SNP513 in intron 9 of the human calretinin gene (CALB2) is a risk factor for colon cancer. *Anticancer research*, *27*(6C), 4279-4288.

36. Zheng, Q., Zheng, X., Zhang, L., Luo, H., Qian, L., Fu, X., ... & Zhang, M. (2017). The neuron-specific protein TMEM59L mediates oxidative stress-induced cell death. *Molecular neurobiology*, *54*(6), 4189-4200.

37. Ha, Y. J., Kim, C. W., Roh, S. A., Cho, D. H., Park, J. L., Kim, S. Y., ... & Kim, J. C. (2015). Epigenetic regulation of KLHL34 predictive of pathologic response to preoperative chemoradiation therapy in rectal cancer patients. *International Journal of Radiation Oncology* Biology* Physics*, *91*(3), 650-658.

38. Yang, C., Yu, K. D., Xu, W. H., Chen, A. X., Fan, L., Ou, Z. L., & Shao, Z. M. (2013). Effect of genetic variants in two chemokine decoy receptor genes, DARC and CCBP2, on metastatic potential of breast cancer. *PloS one*, *8*(11), e78901.

39. Hu, Y. W., Kang, C. M., Zhao, J. J., Nie, Y., Zheng, L., Li, H. X., ... & Qiu, Y. R. (2018). Lnc RNA PLAC 2 down-regulates RPL 36 expression and blocks cell cycle progression in glioma through a mechanism involving STAT 1. *Journal of cellular and molecular medicine*, *22*(1), 497-510.

40. Li, Y., & Yang, P. (2011). GPC5 gene and its related pathways in lung cancer. *Journal of thoracic oncology*, *6*(1), 2-5.

41. Wang, H., Dong, X., Gu, X., Qin, R., Jia, H., & Gao, J. (2015). The microRNA-217 functions as a potential tumor suppressor in gastric cancer by targeting GPC5. *PLoS One*, *10*(6), e0125474.

42. Wagner, K., Grzybowska, E., Butkiewicz, D., Pamula-Pilat, J., Pekala, W., Tecza, K., ... & Försti, A. (2007). High-throughput genotyping of a common deletion polymorphism disrupting the TRY6 gene and its association with breast cancer risk. *BMC genetics*, *8*(1), 41.

43. Wang, K. K., Liu, N., Radulovich, N., Wigle, D. A., Johnston, M. R., Shepherd, F. A., ... & Tsao, M. S. (2002). Novel candidate tumor marker genes for lung adenocarcinoma. *Oncogene*, *21*(49), 7598.

44. Geng, Y. J., Xie, S. L., Li, Q., Ma, J., & Wang, G. Y. (2011). Large intervening non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression. *Journal of International Medical Research*, *39*(6), 2119-2128.

45. Skibola, C. F., Bracci, P. M., Halperin, E., Conde, L., Craig, D. W., Agana, L., ... & Spinelli, J. J. (2009). Genetic variants at 6p21. 33 are associated with susceptibility to follicular lymphoma. *Nature genetics*, *41*(8), 873.

46. Hajjari M, Salavaty A. HOTAIR: an oncogenic long non-coding RNA in different cancers. Cancer Biol Med. 2015;12(1):1-9. doi:10.7497/j.issn.2095-3941.2015.0006

47. Chimonidou, M., Tzitzira, A., Strati, A., Sotiropoulou, G., Sfikas, C., Malamos, N., ... & Lianidou, E. (2013). CST6 promoter methylation in circulating cell-free DNA of breast cancer patients. *Clinical biochemistry*, *46*(3), 235-240.

48. Castilla, M. Á., López-García, M. Á., Atienza, M. R., Rosa-Rosa, J. M., Díaz-Martín, J., Pecero, M. L., ... & Palacios, J. (2014). VGLL1 expression is associated with a triple-negative basal-like phenotype in breast cancer. *Endocrine-related cancer*, *21*(4), 587-599.

49. Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., & Shaughnessy Jr, J. D. (2003). The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. *New England Journal of Medicine*, *349*(26), 2483-2494.

50. Zhang, Z. F., Zhang, H. R., Zhang, Q. Y., Lai, S. Y., Feng, Y. Z., Zhou, Y., ... & Zhou, J. Y. (2018). High expression of TMEM40 is associated with the malignant behavior and tumorigenesis in bladder cancer. *Journal of translational medicine*, *16*(1), 9.

51. Rahiminejad, S., Maurya, M.R., Mukund, K. *et al.* Modular and mechanistic changes across stages of colorectal cancer. *BMC Cancer* **22,** 436 (2022). https://doi.org/10.1186/s12885-022-09479-3

52. Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, Mao Tanabe, KEGG as a reference resource for gene and protein annotation, Nucleic Acids Research, Volume 44, Issue D1, 4 January 2016, Pages D457–D462, https://doi.org/10.1093/nar/gkv1070

53. 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.

54. Sidiropoulos K, Viteri G, Sevilla C, Jupe S, Webber M, Orlic-Milacic M, Jassal B, May B, Shamovsky V, Duenas C, Rothfels K, Matthews L, Song H, Stein L, Haw R, D'Eustachio P, Ping P, Hermjakob H, Fabregat A. Reactome enhanced pathway visualization. Bioinformatics. 2017 Nov 1;33(21):3461-3467. doi: 10.1093/bioinformatics/btx441. PubMed PMID: 29077811.

55. Wang Z, He E, Sani K, Jagodnik KM, Silverstein MC, Ma'ayan A. Drug Gene Budger (DGB): an application for ranking drugs to modulate a specific gene based on transcriptomic signatures. Bioinformatics. 2019 Apr 1;35(7):1247-1248. doi: 10.1093/bioinformatics/bty763. PMID: 30169739; PMCID: PMC6449747.