COADREADx: A comprehensive algorithmic dissection unravels salient biomarkers and actionable insights into the discrete progression of colorectal cancer

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

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

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

[*]Corresponding author

E-mail:apalania@scbt.sastra.edu

Abstract

Colorectal cancer is a common condition with an uncommon burden of disease, heterogeneity in manifestation, and no definitive treatment in the advanced stages. Against this backdrop, renewed efforts to unravel the genetic drivers of colorectal cancer progression are paramount. Early-stage detection contributes to the success of cancer therapy and increases the likelihood of a favorable prognosis. Here, we have executed a comprehensive computational workflow aimed at uncovering the discrete stagewise genomic drivers of colorectal cancer progression. Using the TCGA COADREAD expression data and clinical metadata, we constructed stage-specific linear models as well as contrast models to identify stage-salient differentially expressed genes. Stage-salient 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, STAC2, UCHL1, KCNG1 (stage-I salient), KLHL34, LPHN3, GREM2, ADCY5, PLAC2, DMRT3 (stage-II salient), PIGR, HABP2, SLC26A9 (stage-III salient), GABRD, DKK1, DLX3, CST6, HOTAIR (stage-IV salient), and CDH3, KRT80, AADACL2, OTOP2, FAM135B, HSP90AB1 (top linear model genes). In particular the study yielded 31 genes that are progression-significant such as ESM1, DKK1, SPDYC, IGFBP1, BIRC7, NKD1, CXCL13, VGLL1, PLAC1, SPERT, UPK2, and interestingly three members of the LY6G6 family. Significant monotonic linear model genes included HIGD1A, ACADS, PEX26, and SPIB. The stage-salient genes were benchmarked using normals-augmented dataset, and cross-referenced with existing knowledge. In addition, the signature of a multicellular immuno-cyte community specific to colorectal cancer relative to normal tissue was identified. The candidate biomarkers were used to construct the feature space for learning an optimal model for the digital screening of early-stage colorectal cancers. A feature space of just seven

biomarkers, namely ESM1, DHRS7C, OTOP3, AADACL2, LPHN3, GABRD, and LPAR1, was sufficient to optimize a RandomForest model that achieved >98% balanced accuracy (and performant recall) on blind validation with external datasets. Survival analysis yielded a panel of three stage-IV salient genes, namely HOTAIR, GABRD, and DKK1, for the design of an optimal multivariate model for patient risk stratification. Integrating the above results, we have developed COADREADx, a web-server for assisting the screening and prognosis of colorectal cancers. COADREADx has been deployed at: https://apalanialab.shinyapps.io/coadreadx/ for academic research and further refinement.

Introduction

Colorectal adenocarcinoma (COADREAD), or colorectal cancer, is the third most commonly diagnosed cancer in males and the second in females, with an estimated 1.9 million cases and 930 000 deaths occurring in 2020 (compared to 1.4 million cases and 693,000 deaths in 2012) [1]. There are numerous lifestyle and environmental drivers of colorectal cancer in addition to family history, making the bulk of its incidence sporadic [2]. The main lifestyle and environmental influences include a lack of balanced diet [3], physical inactivity, obesity [4], consumption of alcohol and tobacco [5], etc. Familial forms of colorectal cancer include familial adenomatous polyposis (FAP) and hereditary nonpolyposis colorectal cancer (HNPCC), also called Lynch syndrome. Genetic susceptibility in FAP is associated with mutations in the APC tumor suppressor gene (TSG) [6], while HNPCC is associated with mutations in the genes MSH2 and MLH1 involved in the DNA repair pathway [2]. Since survival rates in colorectal cancer plummet with late-stage of presentation, effective surveillance via access to screening

models is necessary. Early-stage diagnosis of colorectal cancer is essential to secure an advantageous prognosis, which could help in clinical management of the disease.

The Cancer Genome Atlas (TCGA) research network has found mutational and integrative signatures in the multidimensional COADREAD dataset [7], but so far our knowledge with respect to the stage-wise progression of colorectal cancer has been incomplete and inadequate. Given that the AJCC staging of colorectal cancer is based on histopathology (viz. the TNM staging) [8], we studied the evidence for a molecular basis of cancer progression in discrete stages, and developed data-driven workflows for discerning the molecular signatures of colorectal cancer through RNA-Seq transcriptomics. We extended the protocol introduced in Sarathi and Palaniappan [9], and identified stage-salient biomarkers. In addition, a new class of biomarkers called progression-significant DEGs, which are genes with a significant monotone trend of differential expression, were also identified. It is noted that the early-stage (i.e, stage-I and stage-II salient) biomarkers could be useful in development of diagnostics and prognostic models, whereas progression-significant biomarkers could pinpoint potential therapeutic targets to halt or reverse the course of cancer (before it does metastasize to a point of no return). A network analysis grounds the findings in a larger context, lending more evidence for the molecular origins of stage-wise discrete cancer progression. It is known that gene expression profiles of certain markers define cell-type identity [10], and even tissue microenvironment [11], it is reasonable to suppose that a community structure of cell-types drives colorectal cancer progression. Molecular gene signatures characterize the cell composition of the tumor, and it could be argued that the tumor progression through stages is in part or whole determined by the complex and collective changes in gene expression. Based on the above results, we have developed models for the early-stage screening as well as risk stratification of colorectal cancer.

These models are bundled into COADREADx, a pilot tool for the digital diagnostic and prognostic screening of colorectal cancers. A user-friendly interface to COADREADx is available at: https://apalanialab.shinyapps.io/coadreadx/ for academic use. All original datasets used in the study were obtained from the public-domain, and all the intermediate results generated from the study are available as Supplementary Information (doi: 10.6084/m9.figshare.20489211.v4).

Materials and methods

The workflow is summarized in Fig 1 and discussed in detail below. The identification of stage-salient biomarkers follows the computational protocol developed earlier in our lab [9].

Data preprocessing

Normalized and $\log_2$-transformed Illumina HiSeq RNA-Seq gene expression data for Colorectal Adenocarcinoma (COADREAD) processed by the RSEM pipeline [12] were obtained from TCGA via the firebrowse.org portal [13]. 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 [14].

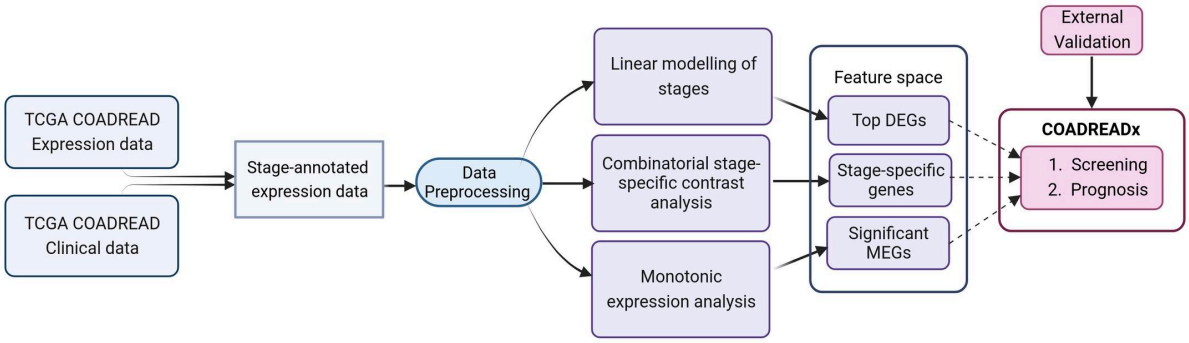

Fig 1. Study design for the dissection of discrete stage-wise progression of colorectal cancer. The identified candidate biomarkers could be used to train machine learning classifiers for the screening and prognosis of colorectal cancers.

Linear modelling

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 [15]. The following linear model was fit for each gene's expression based on the design matrix shown in Fig 2A:

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

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 [16]. 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 [17]. The linear trends across cancer stages for the top significant genes were visualized using boxplots to ascertain the regulation status of the gene relative to the control.

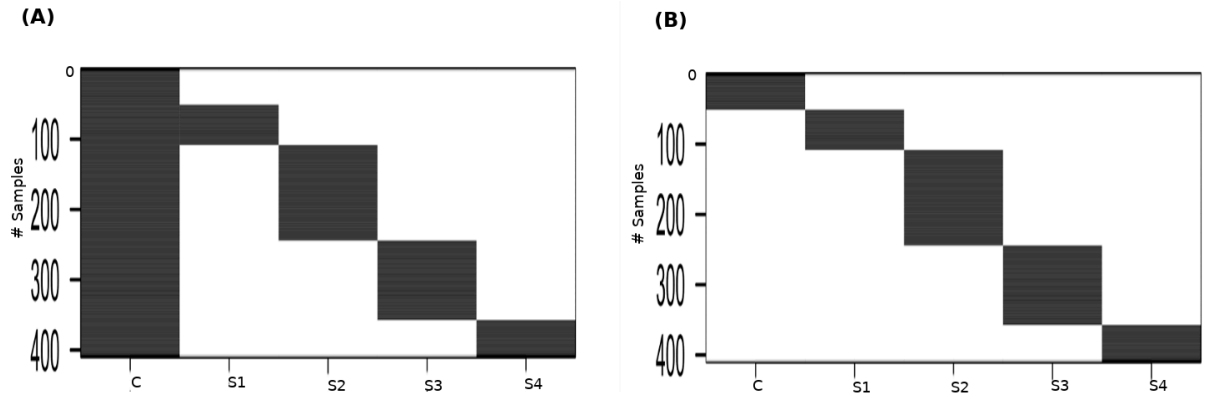

Fig 2. Design matrices (A) for the linear modeling ; and (B) for performing the between-stages contrast analysis. C: Control, S1: Stage-I, S2: Stage-II, S3: Stage-III, S4: Stage-IV.

Pairwise contrasts

To perform contrasts, a slightly modified design matrix shown in Fig 2B was used, which would give rise to the following linear model of expression for each gene:

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \qquad \ldots\ldots (2)$$

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 identified genes whose $|lfc| > 2$ for any stage. For the genes that passed, 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 expression matrix.

Clinical annotation	STAGE - CONTROL			
	I	II	III	IV
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 four-pronged criteria to establish the salience 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. Six 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 any sparsity of progression-significant genes salient to any stage,we defined the "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 modelling of the annotated expression matrix.

Clinical annotation	BETWEEN STAGES:					
	(I, II)	(I, III	(II, III)	(I, IV)	(II, IV)	(III, IV)
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, trend of expression in sync 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:

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

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.

Models for cancer screening and prognosis

(i) Validation of biomarkers with normals-augmented dataset

To study the reliability of findings when a reasonable number of controls are used, we augmented the TCGA cohort with the COADREAD dataset from RNAseqDB [18] that couples TCGA data with 339 normals from the Genotype-Tissue Expression (GTEx) database [19]. The consolidated dataset was subjected to the same biomarker protocol to identify stage-salient genes, and the results compared with those obtained with the TCGA dataset.

(ii) Development of diagnostic model:

The different classes of biomarkers discussed above, including stage-salient genes and monotonically expressed genes, could be used as the feature space to train machine learning (ML) algorithms to solve the binary classification problem of cancer v/s normal samples [20]. Towards this, we split the TCGA dataset in the ratio 0.8:0.2 stratified on the outcome class ('cancer' or 'normal'), and extracted the features of interest. To reduce the dimensionality of the feature space, feature selection techniques such as Boruta [21] and recursive feature elimination [22] were applied to the train dataset and a consensus reduced feature space was obtained. Different ML algorithms were trained on this feature space and hyperparameters optimized by cross-validation. The performance of the ML algorithms was evaluated on the holdout testset to determine the best ML model. The best-performing ML model was then validated on external out-of-domain cohorts.

(iii) Development of Prognostic model:

To study the prognostic significance of the identified stage-salient genes, we used the patient 'OS' ('Overall Survival') attribute in the clinical metadata of the TCGA cohort. Survival analysis was performed according to the protocol outlined in Muthamilselvan and Palaniappan [23]. Univariate Cox regression analysis of the top stage-salient genes was executed to screen the prognostically significant ones, using the R survival library [24]. Genes with p-value < 0.05 were regarded as candidate genes for building a multivariate Cox regression model. This was done using backward variable selection based on the model's Akaike Information Criterion (AIC) metric [25]. The procedure yielded an optimal prognostic signature of size n, given by the following equation:

$$\text{Risk score} = \beta_1 * \text{gene1} + \cdots\cdots + \beta_i * \text{gene}_i + \cdots\cdots + \beta_n * \text{gene}_n$$

where the β_i are the coefficients for the expression of the i^{th} gene. The median risk score from the above distribution was used to classify TCGA COADREAD patients into high-risk and low-risk groups, as implemented in R survminer library [26]. Kaplan–Meier analysis was then performed to assess significance in survival rate variations between the high-risk and low-risk groups, and thereby qualify the biomarker signature.

Benchmarking

Principal component analysis (PCA) was performed using prcomp in R. We used the rand function to choose 100 random genes. In order to visualize significant outlier genes 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-salient differentially expressed genes were visualized using R heatmap and clustered using hclust. Novelty of the identified stage-salient genes was ascertained by screening against curated databases, including the Cancer Gene Census (CGC; cancer.sanger.ac.uk) [27], Network of Cancer Genes NCG7.0 [28], and the Clinical Trials Registry (www.clinicaltrials.gov). STRINGdb was used to translate the findings into network-level insights [29]. To perform immuno-cyte infiltration analysis, we used Cibersort and estimated the proportion of tumor-infiltrating immune cells in TCGA COADREAD samples based on gene expression signatures [10, 30]. Cibersort's inbuilt LM22 signature estimated the proportion of 22 standard immune cell types; setting the number of permutations to 100 allowed the calculation of sample-wise statistical significance with respect to the estimated values. The immuno-cyte patterns of significant samples were analyzed to provide a snapshot of immune ecotypes at play in significant tumor and normal samples, which would increase our basic

understanding of colorectal cancer pathologies and advance rational therapies. The cell-type correlation matrix computed from the proportions of cell-types across significant samples was used to identify substantial co-occurrence patterns. The relative abundance of immunocytes between tumor and normal samples was compared to pinpoint significant differentially elevated or depressed tumor-infiltrating immune cells.

Results

The gene expression matrix from TCGA consisted of 20,502 genes x 428 samples. Upon data pre-processing, the gene expression matrix consisted of 18,212 genes x 409 samples, with an additional vector denoting the sample stage. This dataset is made available as Supplementary File S1. Table 3 shows the distribution of TCGA samples with the corresponding AJCC staging. Table 4 shows the summary of demographic characteristics, where it is seen that the average age was ~ 65 years, and average BMI was ~ 29, hinting at the etiological roles of ageing and obesity.

Table 3. 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	

Table 4. Statistical summary of clinical meta-data associated with the TCGA COADREAD transcriptome.

Characteristic		Control	STAGE OF CRC				NA	Overall
			I	II	III	IV		
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	11	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

Numeric attributes presented as mean ± standard deviation. Nominal attributes (gender and vital status) presented as counts (number of samples). Body mass index (BMI) was calculated only for those instances with both height and weight data.

After preprocessing with voom in *limma,* [14], the dataset yielded 9433 significant genes (adj. $P < 1E\text{-}5$) in the linear modeling, suggesting the existence of a linear trend in their expression across cancer stages. Such an observation could be explained by cancer hallmarks that typically worsen with progression, for e.g., genome-wide instabilitya cancer hallmark, [31]. Some top-ranked upregulated genes from the linear modeling included CDH3, KRT80, ETV4 and ESM1. CDH13 was notably a top upregulated gene obtained from the linear modeling of hepatocellular carcinoma (only after GABRD and PLVAP) in an earlier analysis [9]; these observations point to a consistent role for members of the cadherin gene family in cancer

progression in gastrointestinal cancers. The top downregulated genes included OTOP2, OTOP3, AADACL2 and DHRS7C. Table 5 shows the log-fold changes of the top ten genes in with respect to normal samples,and 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 the linear trend identified needs to be validated with a model for monotonic expression (see Methods), and some stage-specific genes might exhibit maximal differential expression in stages other than stage 4 (Fig 4).

Table 5. Stage-wise lfc, and inferred regulation status of the top ten genes from the linear modelling analysis, ranked by adjusted p-value 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
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
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
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
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

A mixture of both upregulated and downregulated genes was obtained, shown separately here.

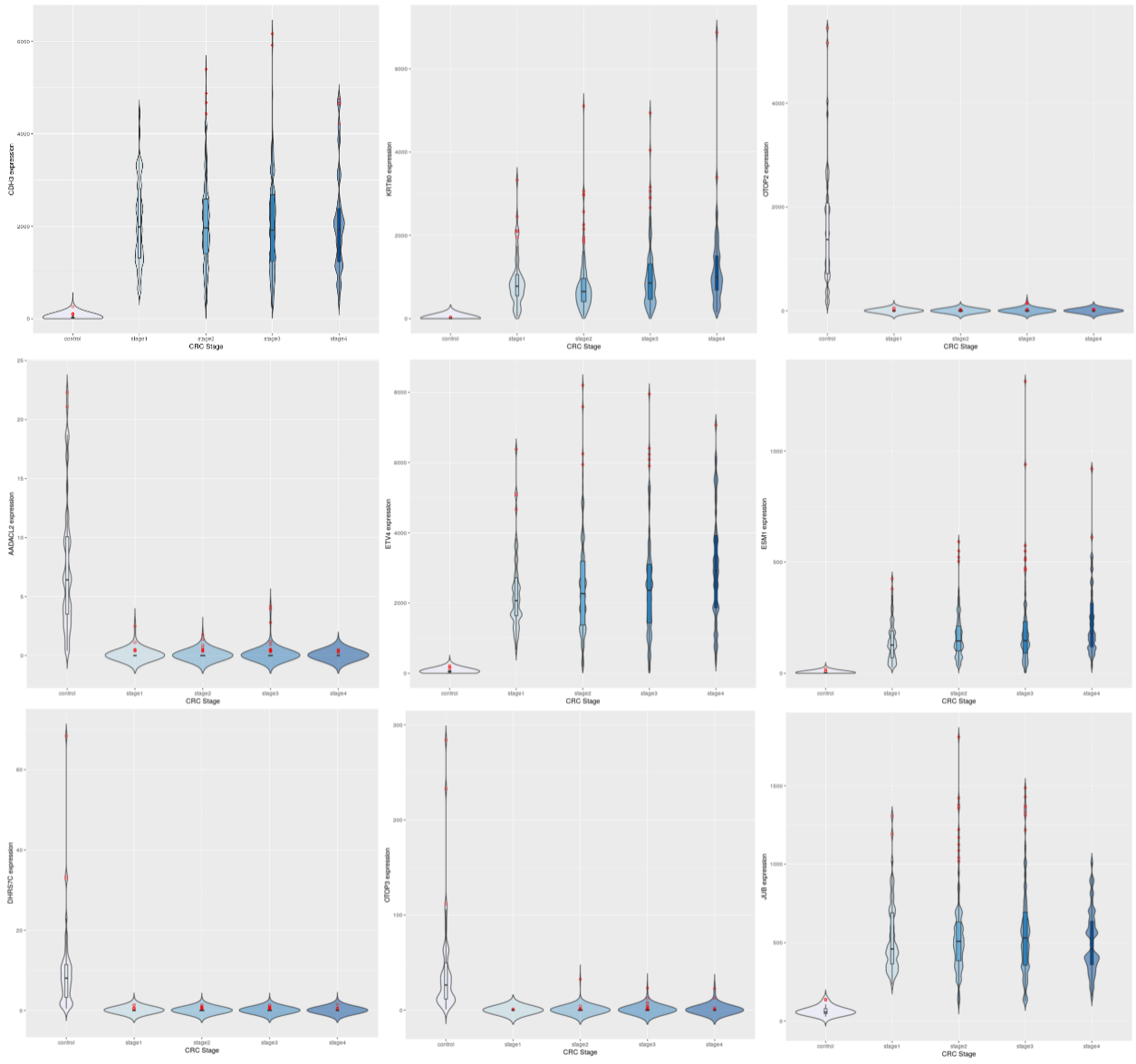

Fig 3. Expression trends of the top 9 DEGs from the linear modeling. Row-wise: CDH3, KRT80, OTOP2, AADACL2, ETV4, ESM1, DHRS7C, OTOP3, JUB. It can be observed that some genes are downregulated to near-zero expression as CRC progresses (notably OTOP2, OTOP3, AADACL2 and DHRS7C).

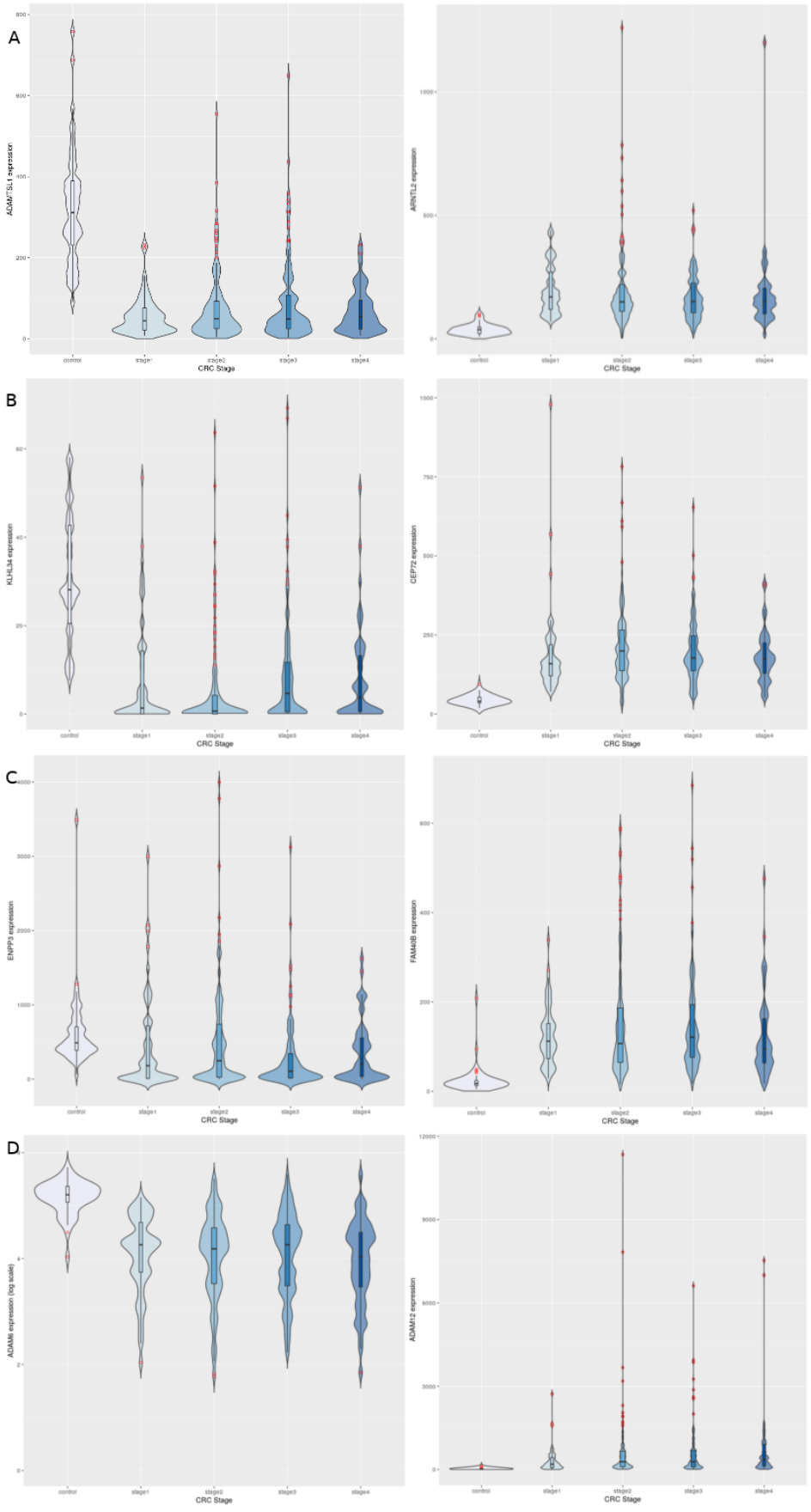
A
ADAM7-SL1 expression
ABIN1L2 expression
B
KLHL34 expression
CEP72 expression
C
ENPP3 expression
FAM49B expression
D
ADAM6 expression (log scale)
ADAM12 expression
control
stage1
stage2
stage3
stage4
CRC Stage

Fig 4. Illustration of the dichotomous expression trends of stage-salient genes (namely, consistent upregulation, and consistent downregulation in cancer samples relative to controls). Each stage is represented by one upregulated gene (column 1) and one downregulated gene (column 2). A: Stage-I: ADAMTSL1 & ARNTL2; B: Stage-II: KLHL34 & CEP72; C: Stage-III: ENPP3 & FAM40B; D: Stage-IV: ADAM6 & ADAM1. Note that the expression of ADAM6 is provided in $\log_{10}$ units.

The samples were visualized using a PCA of the top 100 genes from the linear model (Fig 5A). Separate and distinct clusters of the controls and cancer samples suggested that considerable changes in gene expression in cancer samples. Hence linear modeling yields cancer-specific genes (Supplementary File S2). In contrast, the PCA plot of randomly sampled 100 genes (Fig 5B) failed to distinguish the cancer and control samples, highlighting the significance of linear models in the analysis of cancers.

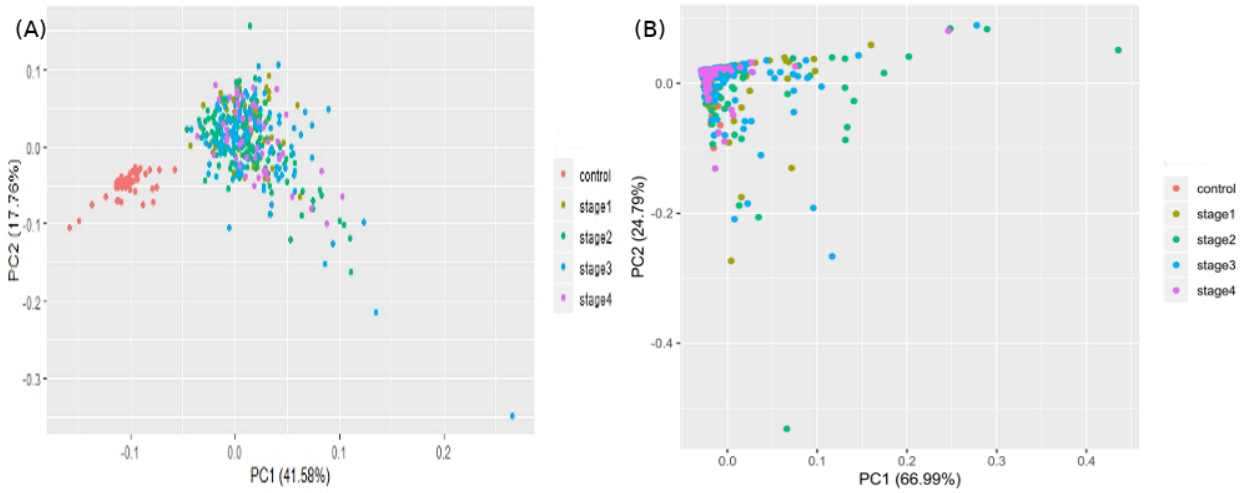

Fig 5. Visualizing the samples in the space of the top two principal components of: (A) top 100 genes of the linear model; and (B) 100 randomly chosen genes.

Differences in gene expression constitute the basis of cell-type identities, and it may not be surprising that gene expression differences drive cancer progression through the AJCC stages. In the first pass, we eliminated 15,970 genes with |lfc|<2 in all stages (Table 1). We binned the remaining genes into different partitions, to obtain stage-specific genes of varying sizes (Fig 6). To establish salience, we applied the second contrast (Table 3) and checked for filter criteria (ii) - (iv) stated in the Methods section. Genes that passed all filters were identified as stage-salient DEGs. This process yielded 71 stage-I salient, 2 stage-II salient, 0 stage-III salient and 59 stage-IV salient genes (Supplementary File S3).

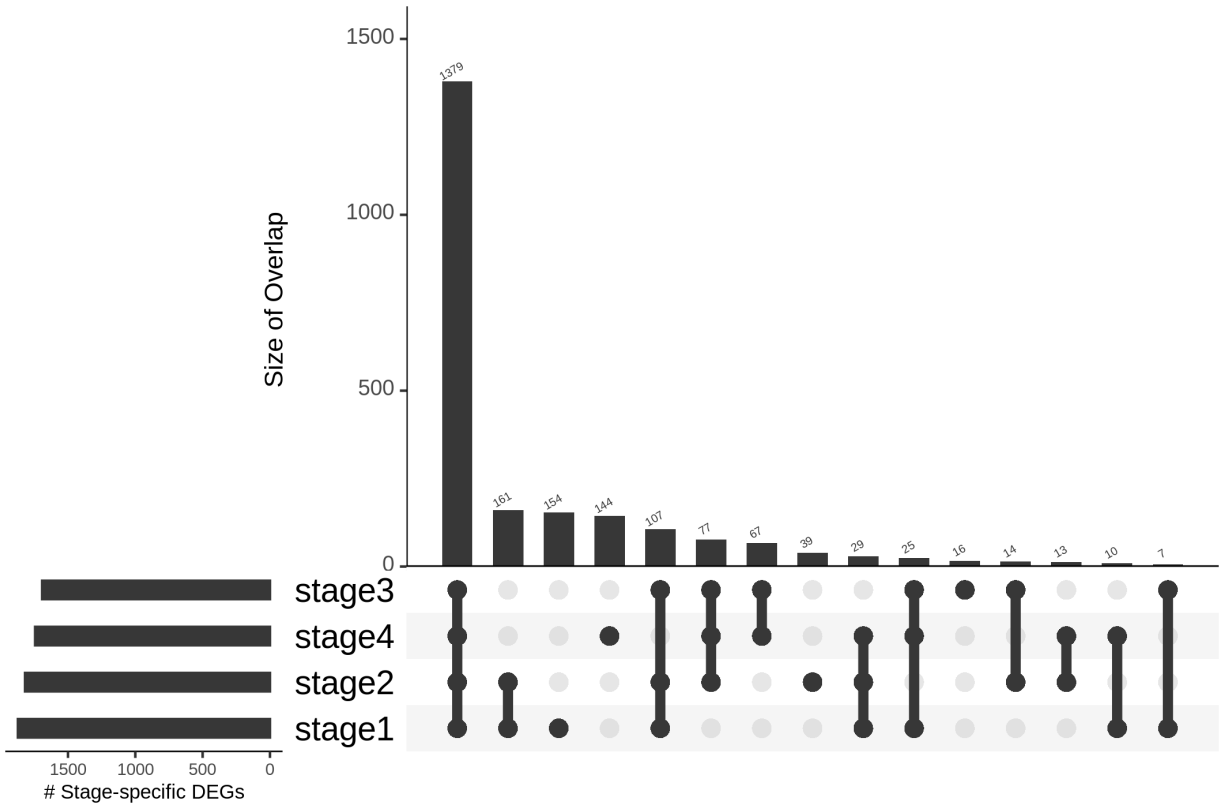

Fig 6. Distribution of genes based on stage-specificity. Of the 2242 DEGs, 1379 appear significant in *all* the stages. It can be clearly seen that the early-stages (stages 1 and 2) share

fewer DEGs with the late-stages (stages 3 and 4), flagging extra factors necessary for cancer progression to metastasis.

Considering 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 stage-salient 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-salient DEGs are presented in Supplementary File S3. It is significant that GABRD emerges as a stage-IV salient gene in COADREAD, reinforcing its identification as a stage-IV salient gene in hepatocellular carcinoma [9], and suggesting a driver role in the metastasis of gastrointestinal cancers more generally.

Table 6. Top ten stage-salient DEGs in each stage, ordered by significance.

Rank	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

Visualizing the lfc expression of stage-salient genes revealed systematic progressive expression across stages (Fig 7). The heatmap was clustered using stage-wise expression differences w.r.to controls 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. Visualizing the clustering of these 40 genes algorithm (Fig 8), we observed that a lot of the stage 4 genes are proto-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. Even though these observations are selective, it is tempting to visualize the implications for the progression pathway of colorectal cancer – initially disabling the damage-control mechanisms innate to the cell and then progressively spiraling out of control.

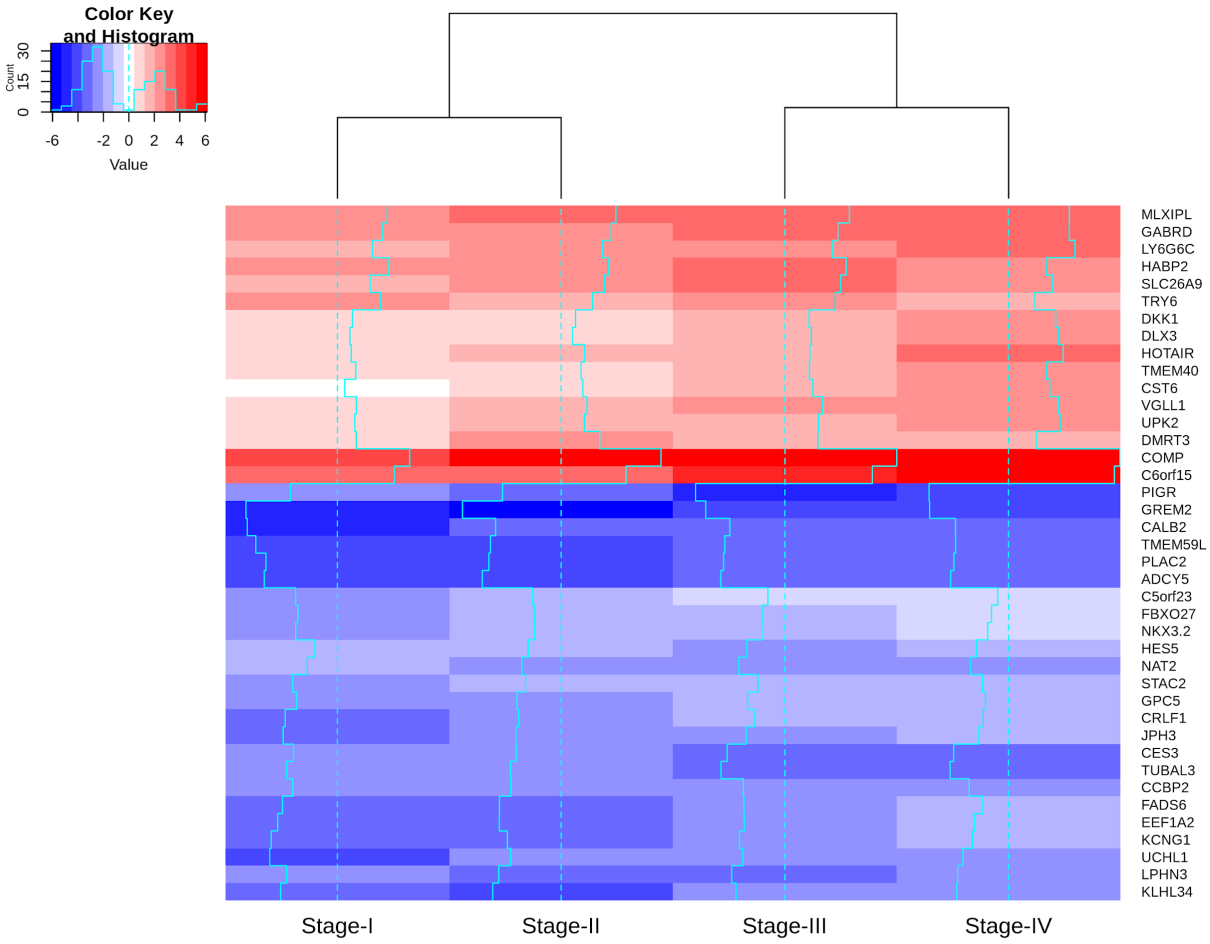

Fig 7. Heatmap of the lfc (w.r.to control samples) of top 40 genes. The expression is increasing on a gradient from blue (downregulated) to red (overexpressed), as shown in the Color Key. Stage-salient genes express maximal salience in one of the stages. It is striking that all the ten stage-IV salient genes show monotonic progressive upregulation (for e.g, GABRD).

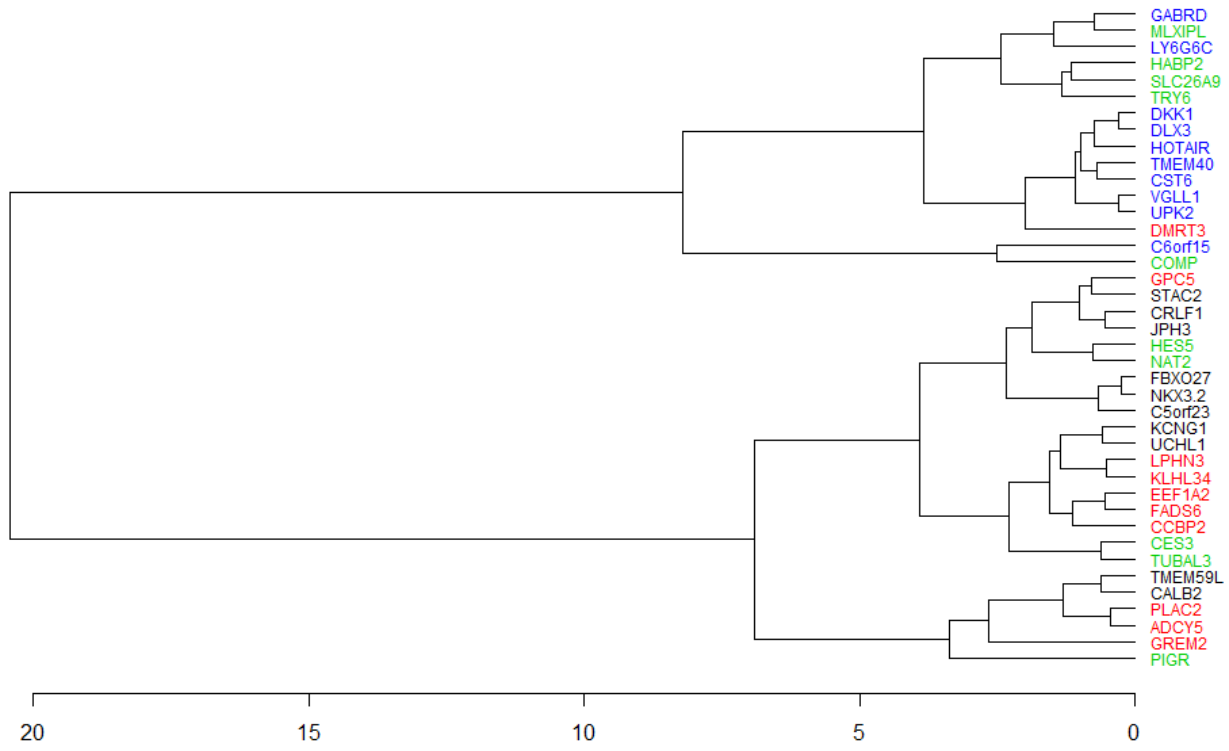

Fig 8. Clustering of 40 stage-salient genes (stage1 – black, stage2 – red, stage3 – green, stage4 – blue). Stage-I and Stage-IV genes do not intermingle in the clustering pattern. There is substantial co-clustering of stage-I with stage-II, and of stage-III with stage-IV. DMRT3 is the only stage-II salient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

The results of the numeric model (eqn.3) sorted by significance are presented in Supplementary File S4. The monotonic analysis yielded 1944 monotonically expressed genes (MEGs; 1389 upregulated and 555 downregulated). These are factors with a constant expression trend agnostic of stage. Applying an adj.p-value cutoff <0.05 yielded 1800 significant MEGs (noted in Supplementary File S5). Examining the overlap of these significant MEGs with stage-salient DEGs yielded 31 progression-significant driver genes (Table 7; expression visualized in Supplementary File S6). As expected, most of these biomarkers (27) are stage-4 salient DEGs, and most of them (27) are also consistently upregulated, signifying unchecked cellular damage

progressing to metastasis. Significant MEGs that are also significant (adj.p-val < 1E-5) in the linear and numeric models (1186 and 997 genes, respectively) are presented in Supplementary File S7. Some of the top 200 genes from the linear model (by adj. p-value) are also significant MEGs; these 18 genes can be found in Supplementary File S8. The intersection between the top 200 genes from the numeric model and the significant monotonically expressed genes yielded 39 genes (presented in Supplementary File S9). A total of 36 genes were found common to the top 200 of both the linear and the numeric (ordinal) models (Supplementary File S10). Three stage-salient DEGs figured in the top 200 genes from the numeric model, namely CES3, LPHN3, and WSCD1. Two of the top 200 genes of the linear model were also stage-salient MEGs, namely GABRD and ESM1.

Table 7. Progression-significant driver genes, obtained by the overlap of significant MEGs with stage-salient DEGs.

S No.	Symbol	Gene	Stage	Status	Adj.p-val
1	ESM1	Endothelial cell-specific molecule 1	IV	UP	3.234E-16
2	GABRD	Gamma-aminobutyric acid receptor subunit delta	IV	UP	7.320E-11
3	LOC283867	Putative Long Intergenic Non-Protein Coding RNA 922	IV	UP	2.628E-10
4	LY6G6E	Lymphocyte antigen 6 family member G6E	IV	UP	1.628E-09
5	LY6G6F	Lymphocyte antigen 6 family member G6F	IV	UP	8.717E-09
6	SPERT	Spermatid-associated protein	IV	UP	3.018E-08
7	LY6G6C	Lymphocyte antigen 6 family member G6C	IV	UP	3.287E-07
8	C2orf48	Uncharacterized protein C2orf48	III	UP	4.499E-07

9	TH	Tyrosine 3-monooxygenase	IV	UP	6.419E-07
10	NKD1	Protein naked cuticle homolog 1	IV	UP	5.896E-06
11	VGLL1	Transcription cofactor vestigial-like protein 1	IV	UP	2.085E-05
12	PLAC1	Placenta-specific protein 1	IV	UP	2.822E-05
13	COL9A3	Collagen alpha-3(IX) chain	IV	UP	8.310E-05
14	SERPINE1	Plasminogen activator inhibitor 1	IV	UP	1.009E-04
15	DSG3	Desmoglein-3	III	UP	1.039E-04
16	IGFBP1	Insulin-like growth factor-binding protein 1	IV	UP	5.645E-04
17	HOTAIR	HOX antisense intergenic RNA	IV	UP	6.808E-04
18	ISM2	Isthmin-2	IV	UP	1.377E-03
19	LOC100133545	C6orf15	IV	UP	1.471E-03
20	DLX3	Homeobox protein DLX-3	IV	UP	1.561E-03
21	C6orf15	Uncharacterized protein C6orf15	IV	UP	4.187E-03
22	KRTAP3.1	Keratin-associated protein 3-1	IV	UP	7.076E-03
23	UPK2	Uroplakin-2	IV	UP	8.241E-03
24	C7orf52	N-acetyltransferase 16	IV	UP	1.145E-02
25	DKK1	Dickkopf-related protein 1	IV	UP	1.621E-02
26	SPDYC	Speedy protein C	IV	UP	1.653E-02
27	BIRC7	Baculoviral IAP repeat-containing protein 7	III	UP	2.918E-02
28	PIGR	Polymeric immunoglobulin receptor	III	DOWN	1.226E-26
29	ADH6	Alcohol dehydrogenase 6	IV	DOWN	6.270E-15
30	ATOH1	Protein atonal homolog 1	IV	DOWN	7.378E-07

31	CXCL13	C-X-C motif chemokine 13	IV	DOWN	4.675E-06

31 genes sorted by the direction of fold-change (up- or downregulation) and corrected significance from the numeric model are shown. Only four genes in this group are monotonically downregulated, namely PIGR, ADH6, ATOH1, and CXCL13, while all the rest are potential proto-oncogene MEGs. It is seen that there are four stage-III salient DEGs (PIGR, DSG3, C2orf48, BIRC70) while all the rest are stage-IV salient DEGs.

Normals-augmented validation

To examine any negative results with the inclusion of more controls in teasing out stage-specific markers, we augmented the dataset using RNAseqDB, which added 339 normal colorectal samples. We noted that the RNAseqDB preprocessing protocol eliminated non-coding transcripts from consideration, ignoring possible expression salience of non-coding RNA biomarkers like HOTAIR. Application of our whole protocol to this controls-augmented dataset yielded a linear model, 1925 stage-specific DEGs (755 stage-I, 418 stage-II, 163 stage-III and 589 stage-IV), and 105 stage-salient markers (40 stage-I, 6 stage-II, 2 stage-III and 57 stage-IV). These are presented in Supplementary File S11. We found a substantial consensus of stage-salient genes between the two datasets, with 70 biomarkers in common (Table 8; highlighted in Supplementary File S11). Notably six of the top stage-I salient genes and nine of the top stage-IV salient genes were identified as salient to the respective stages with the normals-augmented dataset as well, providing robust validation for these biomarkers.

In addition, we identified a colonic cancer dataset with stage-annotation from the Gene Expression Omnibus (GEO) database [32], namely GSE39582, provided by the Carte d'identité

des tumeurs, Ligue Nationale contre le Cancer, France [33]. The dataset had a large number of stage-II (271) and stage-III samples (210), relative to stage-I (38) and stage-IV (60) samples. However, only two normal samples were available, so the dataset was augmented with 308 normal colonic tissue samples from the GTEx. The augmented dataset was subjected to batch correction using ComBat [34], and antilog_2 was taken to obtain the necessary counts for input to voom and the protocol described in the Methods was applied. The results are presented in Supplementary File S12. Five stage-IV salient genes, namely CYP24A1, FGF19, NKD1, COL9A3, and EDNRA are common to both the analyses. In addition, six stage-I salient genes, namely CPXM2, NPR3, PALM, PRDM6, TAGLN, and TPM2 are identified as stage-IV salient here. However the concordance between the markers from the reference TCGA dataset and GSE39582 is not extensive, and merits discussion. Foremost, GSE39582 is limited to colon cancer samples, which might differ in some features from rectal cancers, thereby missing some variation that is captured in the TCGA COADREAD dataset. Second, we would like to note that out-of-domain cohorts might be sensitive to distribution shifts in gene expression, which require measurement calibration with an adequate number of normals from the same (new) cohort. Since there were few normal samples in the original GSE39582 dataset, this might significantly skew the extension of gene signatures established with the reference TCGA cohort. The addition of 308 normal colonic samples available in the GTEx does not mitigate this issue, since (i) these are from an entirely different cohort, and (ii) normal rectal tissue samples remain unaccounted for. In addition, the applicability of candidate biomarker signatures to new cohorts might be bounded by bioinformatic problems pertaining to data curation and processing. The contrarian findings prompted us to seek robust validation of the models developed below.

Table 8. Comparison of the stage-wise salient biomarkers identified with the TCGA and the RNAseqDB datasets.

Stage	No. of stage-salient biomarkers		Size of consensus	Top-10 overlap
	TCGA	RNAseqDB		
I	71	40	25	CALB2, STAC2, UCHL1, KCNG1
II	10	10	5	KLHL34, LPHN3
III	17	10	7	HABP2, SLC26A9
IV	59	57	33	UPK2, LY6G6C, C6orf15, DLX3, CST6, VGLL1

The pval_pdt measure was applied to identify the top ten stage-2 salient and stage-3 salient genes. A substantial stage-wise consensus could be observed. The intersection of the top-10 stage-salient genes in each dataset is shown as 'Top-10 overlap.'

Development of a diagnostic aid for colorectal cancer screening

We combined the 157 stage-salient genes, top ten genes from linear modeling, and the 18 genes that were both linear and monotonically expressed into a single expression feature-space of 185 genes. The TCGA dataset was split into a train dataset of 287 cancer and 41 normal samples, and a holdout testset of 71 cancer and 10 normal samples. Application of the feature selection techniques yielded a consensus feature space of just seven essential features, viz. four of the top ten linear modelling genes (ESM1, DHRS7C, OTOP3, AADACL2), two stage-salient genes (stage-2 salient LPHN3 and stage-4 salient GABRD) and one linearly monotonic gene (LPAR1). Using these features, four different ML models were trained, and hyperparameters optimized. The models were ranked on their performance on the training and holdout test sets (Table 9), and

the Random Forest and 2-layer Neural Network models were identified for blind external validation.

Table 9: A summary of the models used for building a classifier capable of discriminating between cancer and normal samples based on the expression of seven features: ESM1, DHRS7C, OTOP3, AADACL2, LPHN3, GABRD, and LPAR1.

S.No	Classifier	Hyperparameters of interest	Optimal hyperparameters	Performance (bal. acc.)	
				Training	Testing
1	SVM (radial kernel)	cost, gamma	0.5, 0.1	99.97	100
2	Random Forest	ntree (#trees in the forest), mtry (#candidate variables randomly sampled for splitting)	500, 2.83	100	100
3	Neural Networks (1-layer)	size, decay	1, 1	99.97	100
4	Neural Networks (2-layer)	#units in hidden layer 1, #units in hidden layer 2	4,1	100	100

Performance in terms of balanced accuracy (average of the accuracy on either class) is reported. All models achieved ‘perfection’ on the holdout testset, with marginal performance difference on the training set itself.

Two external datasets were chosen for blind validation: (i) Rectal_cancer_MSK [35] with 113 cancer samples, obtained from https://www.cbioportal.org/ ; and (ii) 308 normal colon samples from the GTEx. It is noted that the microarray-based GEO datasets benchmarked in our study,

namely GSE25071, GSE21510, and GSE39582 were limited in the coverage of the gene-space, lacking expression values for some of the seven features used in the models, and not further considered. The hyperparameter-optimized Random Forest and 2-layer neural network models were re-built on the full TCGA dataset and evaluated on the external datasets (Table 10). All the cancer samples were correctly predicted by the Random Forest model, yielding 'perfect' recall. There were just eleven misclassified instances out of the 421 samples in the combined external dataset, and all such instances were normal colon tissue samples, leading to a balanced accuracy of 98.27%. The Random Forest model outperformed the 2-layer Neural network model on all the metrics considered, including sensitivity, specificity, F1-score, and Mathews correlation coefficient (MCC).

Table 10: Blind evaluation of the best-performing ML models on external independent datasets.

S.No	Model	Bal. acc.	Specificity	Precision	Recall	F1-score	MCC
1	Random forest	98.27	96.43	91.13	100	95.36	93.74
2	Neural network (2layer)	96.15	93.18	84.21	99.12	91.06	87.98

The Random Forest model was clearly superior to the Neural Network 2-layer model on the external validation. Bal. acc. refers to balanced accuracy (average of sensitivity (recall) and specificity).

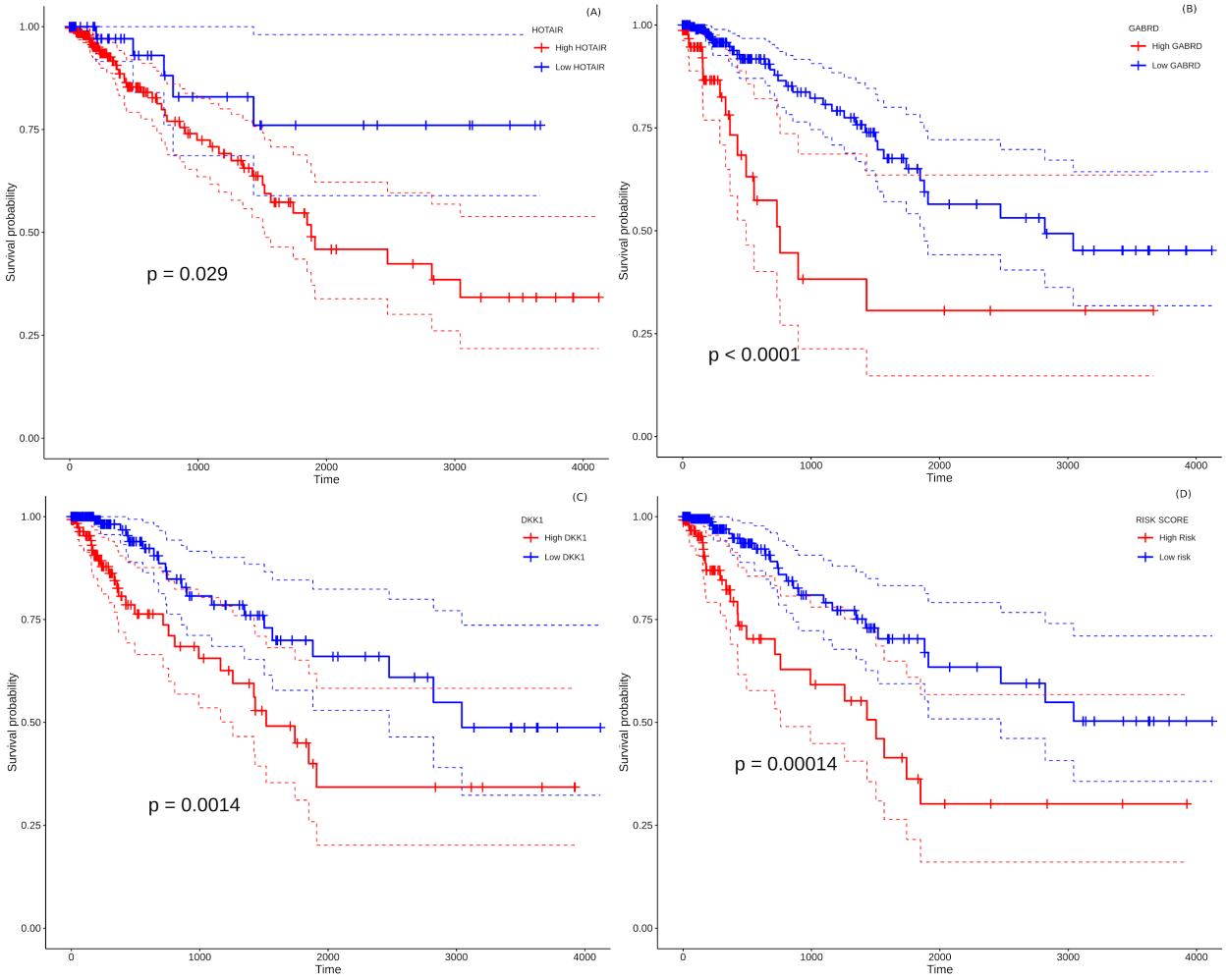

Fig 9. **Survival analysis of prognostically significant stage-salient genes.** Univariate Cox regression analysis of (A) HOTAIR , (B) GABRD, (C) DKK1; and (D) construction of optimal multivariate panel comprising the above biomarkers. Over-expression of the prognostic biomarkers has a significant effect on the survival probabilities ($P < 0.05$), and elevates the patient risk. Red - high-risk, blue - low-risk; colored dashed lines represent corresponding 95% confidence intervals.

Development of a prognostic model for colorectal cancer

All the 157 stage-salient genes were subjected to univariate Cox regression analysis, and the significant biomarkers ($P < 0.05$) are presented in Supplementary File S13. Of the top stage-salient genes, five emerged significant, namely JPH3, HOTAIR, CST6, GABRD, and DKK1 (all $P < 0.03$). HOTAIR, CST6, GABRD, and DKK1 are stage-IV salient, while JPH3 is stage-I salient (Fig 12). Multivariate Cox regression analysis with feature selection yielded an optimal panel of three genes, namely HOTAIR, GABRD, and DKK1, with a model p-value ~ 5e-04, and individual significances ~ 0.0086, 0.0053, and 0.0238, respectively (i.e, all p-values < 0.05). The multivariate risk model was given by:

Risk-score = 0.14872 * HOTAIR + 0.4423 * GABRD + 0.10877 * DKK1

The hazard rate for all the prognostic factors significantly exceeded 1.0, indicating that the constituents of the biomarker panel elevated the prognostic risk, suggesting possible oncogenic roles in line with their overexpression. The distribution of risk scores yielded a median maxstat value of 2.74 for patient risk stratification. Further, the Kaplan-Meier curve of the multivariate model suggested that the high-risk group was significantly associated (p-value <0.0014) with a poorer overall survival than the low-risk group (Fig. 12d). The model yielded an acceptable Concordance index (C-index) ~ 0.71±0.05, suggesting further application as a novel prognostic panel [36-38]. It is significant (and perhaps not surprising) that the identified prognostic panel is entirely composed of stage-IV salient biomarkers, suggesting that the distance to metastasis is the single dominant factor in the stratification and determination of prognosis of colorectal cancer.

Discussion

To clarify the sum of findings from our 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 10) showed that APC was downregulated in the cancer phenotype, flagging its key role as a known TSG.

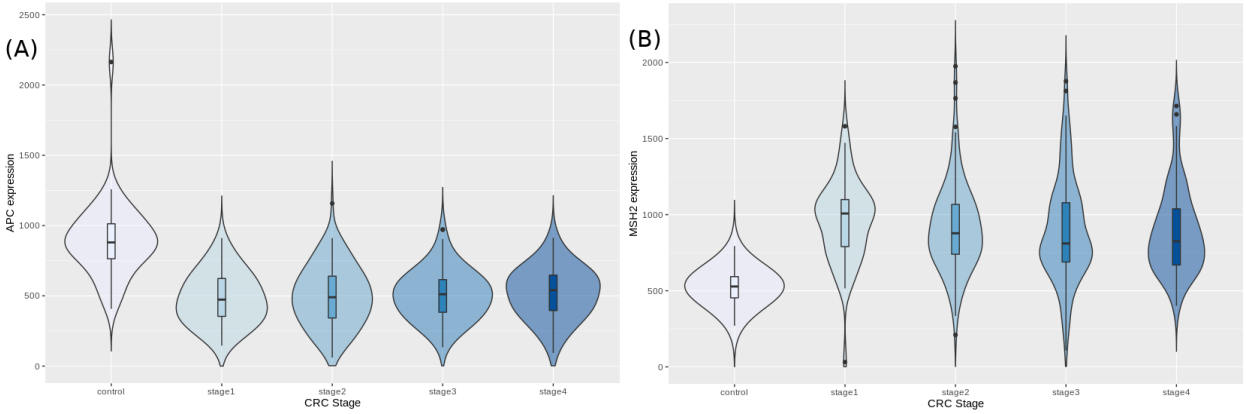

Fig 10. Expression trends of (A) APC and (B) MSH2, known genetic factors of familial CRC.

We then looked at the hub-driver genes identified in a previous study of CRC network analyses [39], 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 11A), reinforcing its role as a hub driver gene for stage 2 progression. EIF2B5 is an oncogene with maximal expression in stage 3 (Fig 11B), again according with its identified role as a major hub driver gene for progression to advanced stages of colorectal cancer.

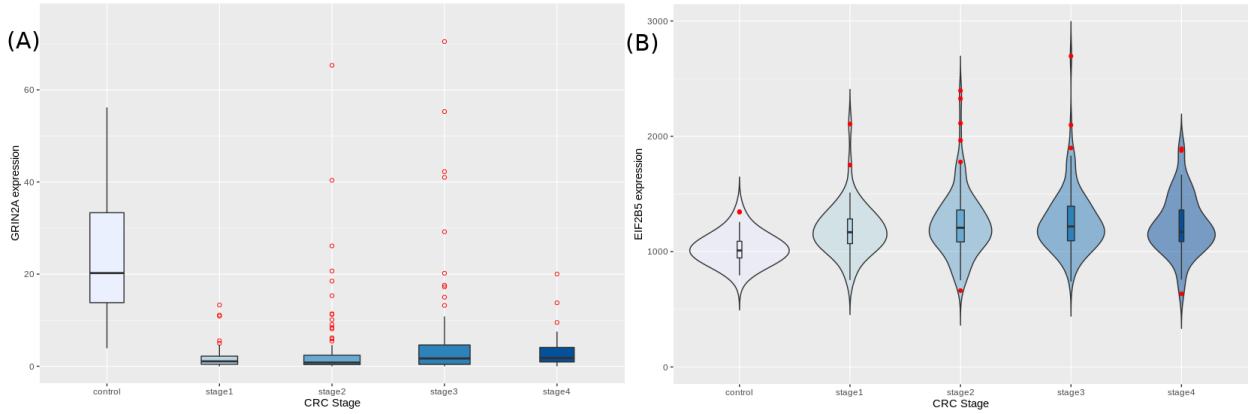

Fig 11. Expression trends of Candidate Hub-Driver genes (A) GRIN2A and (B) EIF2B5.

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) [40], 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 [41] and Advanced Colorectal Cancer [42]. 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 [43], 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* [44]. 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 [45]. 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 [46]. 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 [47].

ESM1 was found to be clearly overexpressed in clear cell renal cell carcinoma [48], and is also one among the 59 stage-4 salient genes from our study. Moreover, ESM1 is also an MEG identified in our study, placing it as a very significant driver of CRC progression. DHRS7C has been recently implicated in signaling pathways involved in glucose metabolism [49]. It exerts its effects via mTORC2, a complex known to be at the heart of metabolic reprogramming [50]. 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 [51]. 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 [52].

Comparing the transcriptomic stage-specific patterns of colorectal cancer samples identified here with their methylomic stage-specific patterns [53], we uncovered interesting connections. Some of the stage-salient genes here were also identified as stage-specific differentially methylated genes, namely: BAI3, TPM2, ZSCAN18, ZNF415 (Stage-I); PLAC2, DMRT3 (Stage-II) ; PIGR, TUBAL3 (Stage-III); and CST6 (Stage-IV). GABRD was earlier found to be significantly differentially methylated in all stages except stage-IV, suggesting that methylation precedes the stage-4 salient change in gene expression observed in this study. In the other direction, GPX3, identified as a stage-I salient gene here, was detected as differentially methylated in stage-2, suggesting the interpretation that change in its expression is necessary for cancer metastasis and mesenchymal transition. The details for the above analysis are presented in Supplementary File S14.

Table 11. Summary of top 200 genes of the linear model documented in the CGC.

Gene Symbol	Illustrative tumors	Documented role	Status
ETV4	Ewing sarcoma, prostate carcinoma	oncogene, fusion	UP
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
FAM135B	small cell lung cancer, esophageal cancer	oncogene	DOWN
FEV	Ewing sarcoma	oncogene, fusion	DOWN
CDH10	Melanoma	TSG	DOWN
PHOX2B	Neuroblastoma	TSG	DOWN
CTNND2	prostate adenocarcinoma, GIST (gastrointestinal stromal tumor)	oncogene	DOWN

These are cancer driver genes with known experimental evidence. 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 regulation tractable to experimental investigation.

Stage-1 salient DEGs

The genes CALB2 and TMEM59L cluster together in Fig 8, showing the least expression in stage-I, suggesting the hypothesis that they are tumor suppressor genes whose expression is required to prevent tumorigenesis. This is supported by evidence in literature, specifically 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 [54]. Another study found that heterozygosity in SNP513 of Intron 9 of the gene CALB2 might be a predictive marker for CRC [55]. It has also been noted that increased TMEM59L expression was a pro-apoptotic indicator of cell death during oxidative stress in neuronal cells [56]. Regarding SOX2 and SOX10, it is noteworthy that the Cancer Genome Atlas Network observed SOX9 as a novel gene with significant recurrent mutations in COADREAD [7].

Stage-2 salient 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 [57]. CCBP2, encoding the Chemokine decoy receptor D6, has an inhibitory effect on breast cancer malignancies due to its action to sequester pro-malignant chemokines [58]. The lncRNA PLAC2 induces cell cycle arrest in glioma by binding to Ribosomal Protein RP L36 in a mechanism involving STAT1 [59]. GPC5 was found to be overexpressed in the lung cancer phenotype [60], in lymphoma, and in gastric cancer [61]. The work by Wang et al. [61] also showed that the overexpression of miR-217 impaired GPC5-induced promotion of proliferation and invasion in GC cells.

Stage-3 salient DEGs

Copy number polymorphisms of TRY6 gene have been found in Breast Cancer [62]. HABP2 gene overexpression has been observed in lung adenocarcinoma and has been proposed as a novel biomarker for the same [63].

Stage-4 salient DEGs

The lncRNA HOTAIR was found to be significantly overexpressed in HCC, and a potential biomarker for lymph node metastasis in HCC [64], and later implicated in different cancers [65]. Another widely-cited study [66] showed that enforced HOTAIR gene expression in epithelial cancer cells induces chromatin reprogramming and an increased metastatic state, while 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 salient significantly monotonically expressed biomarker. GWAS analysis identified a strong association of C6orf15 with occurrence of follicular lymphoma [67]. Promoter methylation of cell free DNA of the CST6 gene was found to be a potential plasma biomarker for Breast Cancer [68]. Expression of VGLL1 and its intronic miRNA miR-934 are associated with sporadic and BRCA1-associated triple negative basal-like breast carcinomas [69]. Expression of DKK1, an inhibitor of osteoblast differentiation, was found to be associated with the presence of bone lesions in patients with multiple myeloma [70]. TMEM40 has been found to be a potential biomarker in patients with Bladder cancer, serving as an oncogene and a possible therapeutic target [71]. The emergence of the C,E,and F members of the Lymphocyte Antigen 6 (LY6) family [72,73] as monotonically expressed proto-oncogenes holds promise for immunotherapy. There is a substantial evidence base for GABRD [74], which is a key component of both the screening and prognostic models developed here. Consistent expression trends in GABRD and other stage-salient MEG DEGs provide unmistakable evidence for the existence of molecular signatures in CRC progression.

Benchmarking with curated databases

We found 13 of the top 200 genes from the linear model documented in the CGC v84 as known cancer genes (Table 11). Two genes, MACC1 and SALL4, were specifically documented for colorectal cancer. HSP90AB1 had been earlier identified as a top MEG in HCC [9]. Screening the 157 stage-salient genes against the NCG7.0, which is a curated database of cancer drivers and healthy drivers, yielded 28 genes, of which eight were in the top 40 stage-salient genes (Supplementary File S15). All the hits were documented to carry mutations in their coding region (vs noncoding region). Three were *canonical* oncogene drivers, namely HOXC11, SOX2, and KCNJ5, while the rest 25 are putative oncogenes and putative tumor suppressors in almost equal measure. Two stage-salient genes, namely CNTN1 and BAI3 (ADGRB3) were documented as putative tumor suppressor genes involved in gastric adenocarcinoma, providing specific support for our findings. PIGR is identified as an essential healthy driver [75], signifying that mutations in this gene confer an exceptional protective effect, and its down-regulation could drive tumorigenic processes. Intriguingly, the stage-salient genes C5ORF23 (NPR3), SOX2, and KCNJ5 are the only instances where the documentation is dissonant with our primary findings; these three were marked as putative oncogenes, though they are identified as down-regulated here. Further investigations in this direction are warranted to set the literature straight. Documentary evidence for drugs targeting any of these genes is absent, emphasizing the value of the present study in pinpointing novel candidates for diagnosis, therapy and prognosis. To perform a systematic analysis of therapeutic interventions based on these targets, we consulted ClinicalTrials.gov for clinical trials targeting stage-salient genes. Ten genes from the top stage-salient genes are being pursued in clinical trials, either as target or endpoint, colorectal or other cancers. Details of clinical trials along with the current status/phase of each trial are

provided in Supplementary File S16. DKK1 and HOTAIR are the only stage-4 salient genes implicated as targets/endpoints in clinical trials. DKK1 is involved in three clinical trials for colorectal and gastric cancers. HOTAIR is the target of a clinical trial for thyroid cancer (NCT03469544) [76]. HOTAIR is documented in the NONCODE database (http://www.noncode.org/) as disease-associated, specifically with colorectal cancer (ID: NONHSAG011264.3), validating its role in oncogenic processes. It is notable that GABRD is not a target in any of the registered clinical trials, flagging a prime potential interest for future efforts. LPHN3, a stage-2 salient gene, is targeted in four clinical trials aimed against metastatic colorectal cancer, to explore possible therapeutic efficacy in thwarting cancer progression prior to irreversible outcomes. FADS6 (a stage-II salient gene) is an endpoint in a clinical trial to treat colorectal adenomatous polyps, which is a precursor to malignant lesions. CALB2 and C5orf23 (NPR3) are each involved in one clinical trial related to colorectal cancer. Some stage-salient genes are being pursued in treatment of cancers in other cell types/tissues, underlining the role played by certain genes in contributing to general cancer hallmark processes [31]. Specifically NAT2 is a target in nine different clinical trials against diverse cancers, significantly highlighting its essential role in driving hallmark processes in unrelated cancers.

Insights from Network Analysis

Stage-wise network analysis of colorectal cancer progression has shed light on certain genes potentially underlying progression [77]. The strength of the computational evidence for the candidate biomarkers identified herein urged a network analysis to examine the findings in a larger context. The intersection between the sets of all stage-salient biomarkers and the significant MEGs might highlight monotonically enriched pathways essential to the

pathophysiology of colorectal cancers. Hence the 31 stage-salient MEGs were chosen to reconstruct the STRING network, with 50 interactors in the first shell and 10 interactors in the second shell. This yielded a PPI with 235 edges with an extremely significant enrichment p-value < 1.0e-16 (Fig 12). A Gene Ontology [78] analysis of this reconstructed network showed enrichment for the Wnt-Frizzled-LRP5/6 complex component at p-value < 1E-04. An analysis with KEGG [79] showed enrichment for 2-oxocarboxylic acid metabolism at p-value ~ 0.001, indicating a Warburg-shift in metabolism. An analysis with Reactome [80] showed significant enrichment of SMAD2/3 and SMAD4 MH2 Domain Mutants in Cancer (p-value < 0.01). These observations *in toto* provide striking evidence for the involvement of these biomarkers in driving CRC progression.

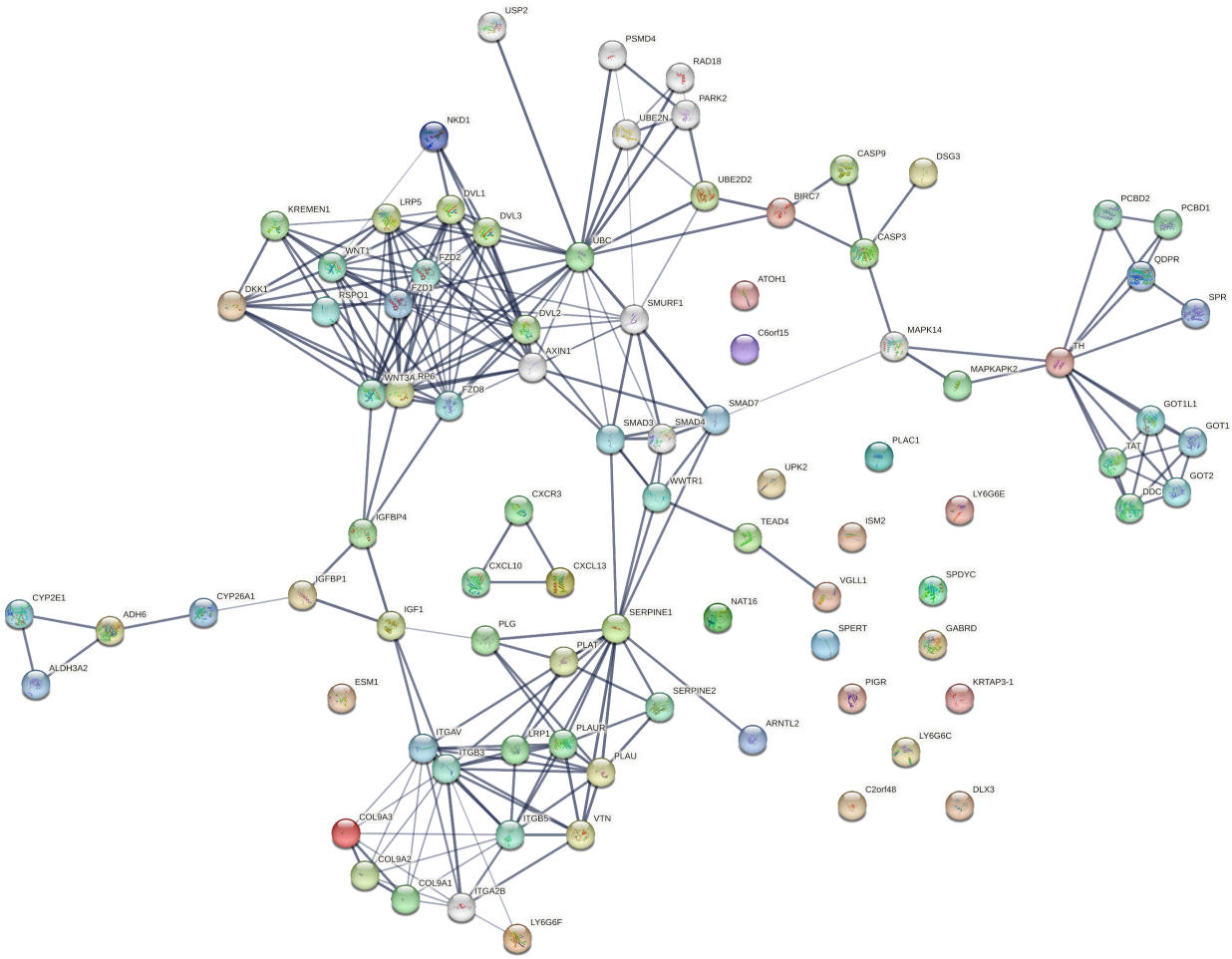

Fig 12. Network reconstruction of perturbed pathways with monotonic expression enrichment based on the seed set of stage-salient MEGs in TCGA COADREAD. Evidence from known interactions (curated databases, experimentally determined) or predicted from gene neighborhood, gene fusions or gene co-occurrence were used in identifying edges. Colored nodes indicate query proteins and first shell of interactors, whereas white nodes indicate second shell of interactors.

Isolated nodes in the network included GABRD, DLX3, ISM2, LY6G6C & E, SPYDC, UPK2, C2orf48, PIGR, KRTAP3-1, C7orf52 (NAT16), SPERT, and PLAC1. All the isolated nodes are proto-oncogenic (see Table 7), hence could provide targets for inhibition in personalized cancer

medicine. An outlier component (not in the giant connected component) was made of the CXCL chemokine family, stemming from CXCL13 - a recently recognized immune checkpoint with a key role in tumor progression [81,82]. This component could constitute a novel target for upregulation in CRC immunotherapy. A drug-repurposing search with the DrugGeneBadger [83] for each of the 31 stage-salient MEGs yielded drugs (small molecules with q-values < 0.05) to pharmacologically alter the expression of these identified biomarkers. The search revealed that curcumin is effective against at least 13 of these targets, and piperlongumine against eight of these targets. Six biomarkers (HOTAIR, ISM2, KRTAP3-1, SPDYC, LY6G6F, and NKD1) found no drug available in LINCS1000 [84] to modulate their expression, and these constitute potential novel targets for drug discovery against metastatic transition in CRC.

A network specific to colon cancer could be obtained using the results for GSE39582. Among the 503 Stage-IV salient genes, 262 were also monotonically significant (Supplementary File S17). We reconstructed a StringDB network seeded with these 262 monotonic stage-salient genes. The resulting interaction network with 316 nodes and 521 edges was significant (p-value ~ 1e-15). Fig. 13 shows the giant connected component of this network; the full network is available in Supplementary File S17. Enrichment analysis of the network with Gene Ontology indicated significance for Arp2/3 complex-mediated actin nucleation (p-value ~ 1e-4), which is known to contribute to invasive colorectal cancer [85]. A KEGG analysis showed enrichment for oxidative phosphorylation (p-value ~ 1e-20), with a prominent clustering of NDUF and COX gene families. A Reactome analysis showed a minor enrichment of enzymatic protein conjugation processes (UBE2I, UBA2, SAE1) that monitor intracellular proteins and cell states (p-value ~ 0.02). These findings indicate an enrichment of proliferation-independent

metabolism-rewiring pathways necessary for colorectal cancer progression, and could be contrasted with the analyses in Marisa et al. [33].

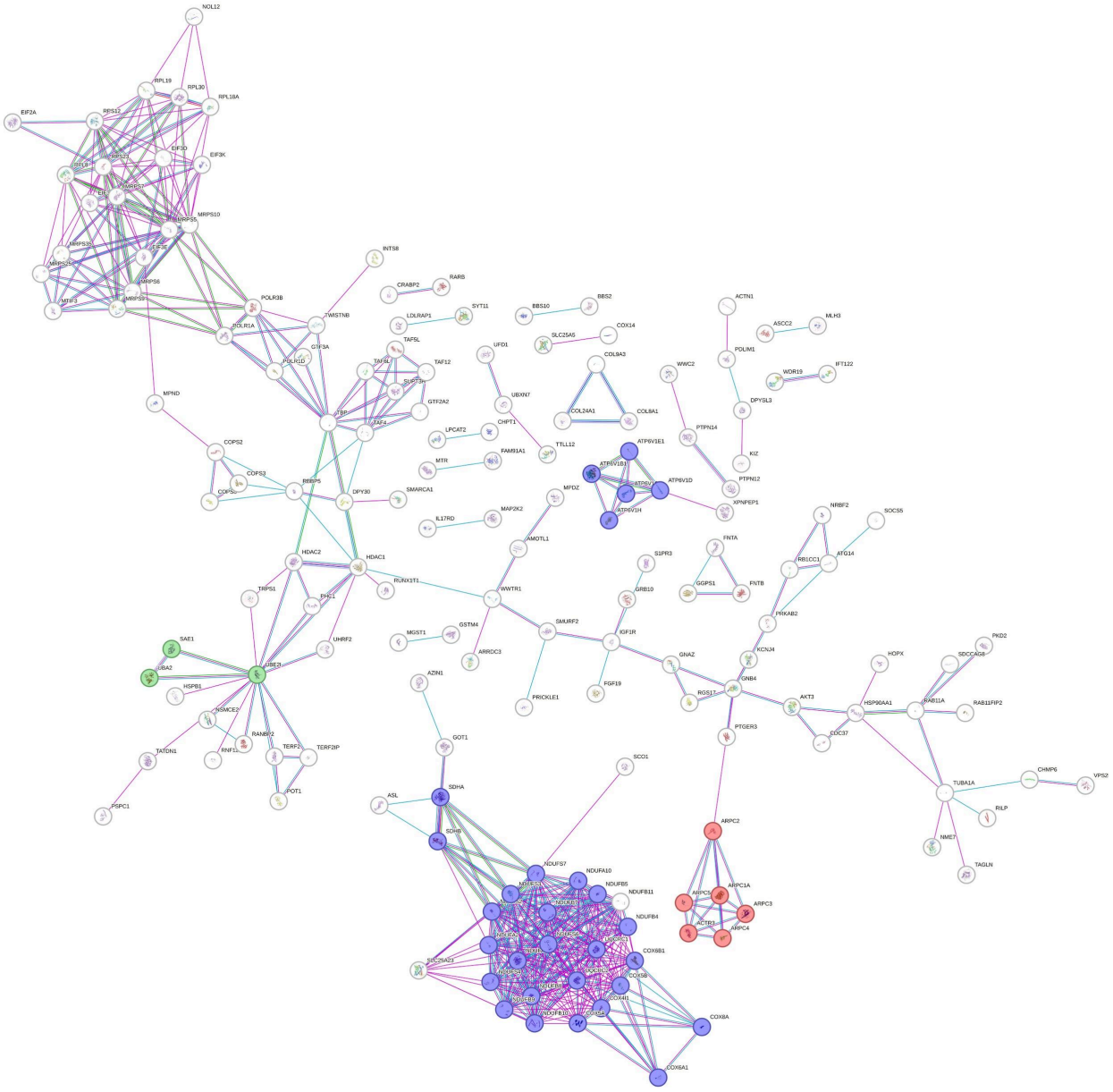

Fig 13. Network reconstruction of perturbed pathways with monotonic expression enrichment based on the seed set of stage-salient MEGs in GSE39582. Evidence from known interactions (curated databases, experimentally determined) or predicted from gene neighborhood, gene fusions or gene co-occurrence were used in identifying edges. Only the giant

connected component is shown. A clustering of enriched pathways could be seen: oxidative phosphorylation involving NDUF, COX, and ATP6V families (purple); ARPC complex (red); and ubiquitin conjugating system (green).

Immune-cell infiltration analysis

Deconvolution of the TCGA samples based on the LM22 immno-cyte signature with 100 permutations yielded 107 samples with significance ($P < 0.05$), including eleven controls. These samples, with their TCGA identifiers, are presented in Supplementary File 18. The significant samples were analyzed for the relative abundance of the 22 immune cell types. A heatmap of the sample-wise immune cell-type proportions was generated (Supplementary File 19 Fig.A), and the clustering patterns of the cell-types across samples was visualized using a dendrogram. We observed the following clusters: mast cells resting and plasma cells; mast cells activated and neutrophils; T cells CD8, T cells follicular helper, and macrophages M1; T cells CD4 memory resting and B cells naive. The macrophages M0 and M2 were clear outgroups in the dendrogram. A normalized stacked bar chart of the sample-wise immnuo-cyte fractions revealed substantial variations in immune cell-type composition between normal and cancer samples (Supplementary File 19, Fig.B). To investigate further, we analyzed the differences in distribution of cell proportions between normal and tumor samples for each immune cell-type (Supplementary File 19, Fig.C; data presented in Supplementary File S18). Eight of the 22 immunocyte types showed significant distribution differences (adj. $P < 0.05$). Specifically, we found the infiltration of four immune cell-types preferentially enriched in tumor samples, namely macrophages M0, T cells CD4 memory activated, mast cells activated, and neutrophils, while four other immune cell-types were preferentially depleted in tumor samples, namely macrophages M2, T cells CD4 memory

resting, mast cells resting, and plasma cells. In particular, macrophages M0 exhibited both the largest effect size (> 2.0) and the greatest significance (< 1E-07) of infiltration in tumor samples. The preferential enrichment of mast cells activated and T cells CD4 memory activated versus the preferential depletion of mast cells resting and T cells CD4 resting suggested that tumorigenesis activates resting immune cell-types, potentiating their infiltration of the tumor microenvironment. To integrate these observations, we computed the correlation matrix of the immune cell-types based on their sample-wise proportions over both normal and tumor samples (Supplementary File 19, Fig.D). The largest positive correlations were exhibited by T cells follicular helper with T cells CD8 (Pearson's $\rho \sim 0.52$), and with macrophages M1 (Pearson's $\rho \sim 0.45$), reinforcing their clustering in the dendrogram. Intriguingly, the largest negative correlation (in magnitude) was exhibited by macrophages M0 and T cells CD4 memory resting (Pearson's $\rho \sim -0.51$) (Supplementary File 19, Fig.D). Given that macrophages M0 are preferentially enriched in tumor samples whereas T cells CD4 resting and mast cells resting (Pearson's $\rho \sim -0.47$ with macrophages M0) are both preferentially depleted, these observations cohere and could hold preliminary significance for immunotherapy. Discovery of multicellular community structures could pave the way for personalized immunotherapy in CRC treatment [11, 86].

COADREADx

Based on the external validation, the Random Forest model was identified as the best model for screening early-stage cancer. Coupled with the prognostic model, these could aid the risk stratification of patient samples. With this application in mind, we have deployed COADREADx, an experimental web service for the screening of patient samples as 'cancer' or 'normal', and subsequent prognostication in the case of 'cancer'. COADREADx has been

implemented using R-Shiny (https://shiny.rstudio.com/), and is available for academic use at: https://apalanialab.shinyapps.io/coadreadx/. A help document with sample input files for different use-cases, and a companion how-to video have been made available on the landing page. To aid the effective interpretation of COADREADx predictions, the prediction probability for the predicted diagnostic class is provided, yielding a level of confidence in the prediction. Similarly the risk stratification of 'cancer' samples is accompanied by the quantile of the estimated risk-score as well as its fold-change from the median value of the risk score distribution. These values suggest the strength of evidence for the predicted risk class.

In summary, we have performed a novel de novo analysis of the TCGA COADREAD gene expression dataset, and identified multiple interesting classes of biomarkers. The biomarkers have been validated with alternative datasets, network analysis and immune cell infiltration analysis. Some of the biomarkers could suggest novel hypotheses for targeted therapy and immunotherapy. Using purifying techniques, we have carved feature spaces from these biomarkers to build screening and prognostic models of colorectal cancer. The screening model has been externally validated, while the prognostic model has been bootstrapped for confidence. Both the models have been deployed as a web-server, COADREADx, which has been configured to return confidence estimates for all its predictions. Phenomena of distribution drift and shift in new samples and out-of-domain cohorts challenge the applicability of COADREADx, which might need refinement in the light of such data. Enabling risk stratification is vital to treatment strategy and clinical management of the cancer. Thus experimental validation and further improvement of COADREADx is necessary to demonstrate its clinical utility for screening and prognosis purposes. It is reckoned that the availability of such software-as-medical-devices could

ease the accessibility to effective surveillance technologies for early detection of colorectal cancer [20].

Conclusions

We have executed multiple workflows towards computational validation of stage-salient signatures of colorectal cancer progression. We have identified stage-agnostic progression-significant monotonically expressed biomarkers. Modulating the expression of progression-significant biomarkers (for e.g, by inhibiting the overexpressed ones or activating the expression of downregulated ones) represents a promising potential strategy to effectively intervene in the progression of colorectal cancer. The candidate biomarkers identified have been benchmarked against curated databases and the literature. A binary classification model for early-stage screening of colorectal cancer was created using seven consensus biomarkers (namely ESM1, DHRS7C, OTOP3, AADACL2, LPHN3, GABRD, and LPAR1), and yielded > 98% balanced accuracy on external validation. A survival analysis protocol yielded a prognostic panel of three stage-IV salient genes (namely HOTAIR, GABRD, and DKK1) for patient risk stratification, suggesting that high-risk prognosis is entirely dependent on the oncogenic expression of these metastasis-salient genes, and inviting experimental confirmation. By benchmarking our findings in multiple ways, we have evaluated the assumptions underlying our computational models. The weight of the evidence presented herein suggests the central role of molecular factors in cancer progression. In summary, we have developed a set of tools for colorectal cancer screening and prognosis, COADREADx, based on the candidate biomarkers identified in our study. COADREADx is available for academic use at: https://apalanialab.shinyapps.io/coadreadx/. Our work provides a pilot study for further

exploration of signature panels on the overall path to securing the best possible intervention for the condition. The hypothesis-agnostic overall study design provides a framework for the investigation of other cancers, and more generally, conditions that are progressive (and degenerative).

Acknowledgments

We would like to thank the reviewers for helpful comments. We are grateful to SASTRA deemed University for resources, infrastructure, and support. Computing in our lab is also supported on a generous grant from Google TPU Research Cloud (TRC).

References

1. Morgan E, Arnold M, Gini A, Lorenzoni V, Cabasag CJ, Laversanne M, et al. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. Gut. 2023;72(2): 338–344.
2. Haggar FA, & Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clinics in colon and rectal surgery. 2009; *22*(4): 191.
3. Willett WC. Diet and cancer: an evolving picture. JAMA. 2005; 293(2): 233–234
4. de Jong AE, Morreau H, Nagengast FM, Mathus-Vliegen EM, Kleibeuker JH, Griffioen G, et al. Prevalence of adenomas among young individuals at average risk for colorectal cancer. Am J Gastroenterol. 2005;100(1): 139–143.
5. Zisman AL, Nickolov A, Brand RE, Gorchow A, Roy HK. 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. Overview of the epidemiology of colorectal cancer. Dis Colon Rectum 1997;40(4): 483–493.
7. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487: 330–337. doi: 10.1038/nature11252.
8. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. 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. 2017;67(2): 93-99.
9. Sarathi A, Palaniappan A. Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma. BMC Cancer. 2019;19(1): 663. doi: 10.1186/s12885-019-5838-3.
10. Chen B, Khodadoust MS, Liu CL, Newman AM, & Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods in molecular biology (Clifton, N.J.). 2018;1711: 243–259. doi: 10.1007/978-1-4939-7493-1_12.
11. Luca BA, Steen CB, Matusiak M, Azizi A, Varma S, Zhu C, et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. Cell. 2021;184(21): 5482–5496.e28. doi: 10.1016/j.cell.2021.09.014.
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. doi: 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. doi: 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. Wang Q, Armenia J, Zhang C, Penson AV, Reznik E, Zhang L, Minet T, et al. Unifying cancer and normal RNA sequencing data from different sources. Sci Data. 2018;5: 180061. doi: 10.1038/sdata.2018.6.
19. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S,et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45: 580–585. doi:10.1038/ng.2653.
20. Muthamilselvan S, Ramasami Sundhar Baabu P, Palaniappan A. Microfluidics for Profiling miRNA Biomarker Panels in AI-Assisted Cancer Diagnosis and Prognosis. Technology in Cancer Research & Treatment. 2023; 22: 15330338231185284. doi: 10.1177/15330338231185284.
21. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010;36(11): 1–13.
22. Kuhn, Max. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008;28(5): 1–26.

23. Muthamilselvan S, Palaniappan A. CESCProg: a compact prognostic model and nomogram for cervical cancer based on miRNA biomarkers. PeerJ. 2023;27:11:e15912. doi: 10.7717/peerj.15912.
24. Therneau, Terry M., and Thomas Lumley. "Package ‘survival’." R Top Doc 128.10 (2015): 28-33.
25. Gerds TA, Scheike TH, and Andersen PK. (2012) Absolute risk regression for competing risks: interpretation, link functions, and prediction. Statistics in medicine, 31(29):3921–3930.
26. Kassambara A, Kosinski M, Biecek P, Fabian S. (2017). Package ‘survminer’. Drawing Survival Curves using ‘ggplot2’(R package version 0.3.1).
27. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4: 177–83.
28. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A,et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 2019;20: 1. doi: 10.1186/s13059-018-1612-0.
29. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49(D1): D605-12.
30. Newman AM, Steen CB, Liu C, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37: 773–782. doi: 10.1038/s41587-019-0114-2.

31. Hanahan, D., & Weinberg, R. A. Hallmarks of cancer: the next generation. Cell. 2011;144(5): 646–674. doi: 10.1016/j.cell.2011.02.013.

32. Barret T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2013;41(D1): D991–D995. doi: 10.1093/nar/gks1193.

33. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5): e1001453.

34. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics (Oxford, England). 2012;28(6): 882–883. doi: 10.1093/bioinformatics/bts034.

35. Chatila WK, Kim JK, Walch H, Marco MR, Chen CT, Wu F, et al. Genomic and transcriptomic determinants of response to neoadjuvant therapy in rectal cancer. Nat Med. 2022;28(8): 1646–1655.

36. Svoboda M, Slyskova J, Schneiderova M, Makovicky P, Bielik L, Levy M, et al. HOTAIR long non-coding RNA is a negative prognostic factor not only in primary tumors, but also in the blood of colorectal cancer patients. Carcinogenesis. 2014;35(7): 1510–1515. doi: 10.1093/carcin/bgu055.

37. Niu G, Deng L, Zhang X, Hu Z, Han S, Xu K, et al. GABRD promotes progression and predicts poor prognosis in colorectal cancer. Open medicine (Warsaw, Poland). 2020;15(1): 1172–1183. doi: 10.1515/med-2020-0128.

38. Sui Q, Zheng J, Liu D, Peng J, Ou Q, Tang J, et al. Dickkopf-related protein 1, a new biomarker for local immune status and poor prognosis among patients with colorectal liver Oligometastases: a retrospective study. BMC cancer. 2019;19(1); 1210. doi: 10.1186/s12885-019-6399-1.

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

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

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

42. Hibi K, Goto T, Mizukami H, Kitamura YH, Sakuraba K, Sakata M, et al. Demethylation of the CDH3 gene is frequently detected in advanced colorectal cancer. Anticancer research. 2009;29(6): 2215-2217.

43. Zhao Y, Huang X, Zhang Z, Li H, & Zan T. The Long Noncoding Transcript HNSCAT1 Activates KRT80 and Triggers Therapeutic Efficacy in Head and Neck Squamous Cell Carcinoma, Oxidative Medicine and Cellular Longevity 2022;2022: 4156966, doi: 10.1155/2022/4156966

44. Qu H, Su Y, Yu L, Zhao H, & Xin C. Wild - type p53 regulates OTOP 2 transcription through DNA loop alteration of the promoter in colorectal cancer. FEBS open bio. 2019;9(1): 26–34.
45. Pande M, Joon A, Brewster AM, Chen WV, Hopper JL, Eng C, et al. Genetic susceptibility markers for a breast-colorectal cancer phenotype: Exploratory results from genome-wide association studies. PLoS One. 2018;13(4):e0196245. doi: 10.1371/journal.pone.0196245.
46. Aytes A, Mitrofanova A, Kinkade CW, Lefebvre C, Lei M, Phelan V et al. 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. 2013; 110(37): E3506-E3515.
47. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic acids research. 2016;45(D1): D777-D783.
48. Leroy X, Aubert S, Zini L, Franquet H, Kervoaze G, Villers A, et al. Vascular endocan (ESM‐1) is markedly overexpressed in clear cell renal cell carcinoma. Histopathology. 2010;56(2): 180-187.
49. Ruiz A, Dror E, Handschin C, Furrer R, Perez-Schindler J, Bachmann C, et al. Over-expression of a retinol dehydrogenase (SRP35/DHRS7C) in skeletal muscle activates mTORC2, enhances glucose metabolism and muscle performance. Sci Rep. 2018;8: 636. doi: 10.1038/s41598-017-18844-3.

50. Masui K, Cavenee WK, Mischel PS. mTORC2 in the center of cancer metabolic reprogramming. Trends Endocrinol Metab. 2014;25(7): 364-73. doi: 10.1016/j.tem.2014.04.002.
51. Liang XH, Zhang GX, Zeng YB, Yang HF, Li WH, Liu QL, et al. LIM protein JUB promotes epithelial–mesenchymal transition in colorectal cancer. Cancer science. 2014;105(6): 660-666.
52. Lee D, Xu IMJ, Chiu DKC, Lai RKH, Tse APW, Li LL, etal. Folate cycle enzyme MTHFD1L confers metabolic advantages in hepatocellular carcinoma. The Journal of clinical investigation. 2017;127(5): 1856-1872.
53. Muthamilselvan S, Raghavendran A, & Palaniappan A. Stage-differentiated ensemble modeling of DNA methylation landscapes uncovers salient biomarkers and prognostic signatures in colorectal cancer progression. PloS one. 2022;17(2): e0249151. doi: 10.1371/journal.pone.0249151.
54. Stevenson L, Allen WL, Proutski I, Stewart G, Johnston L, McCloskey K, et al. Calbindin 2 (CALB2) regulates 5-fluorouracil sensitivity in colorectal cancer by modulating the intrinsic apoptotic pathway. PLoS One. 2011;6(5): e20276.
55. Vonlanthen S, Kawecki TJ, Betticher DC, Pfefferli M, & Schwaller B. Heterozygosity of SNP513 in intron 9 of the human calretinin gene (CALB2) is a risk factor for colon cancer. Anticancer research.2007;27(6C): 4279-4288.
56. Zheng Q, Zheng X, Zhang L, Luo H, Qian L, Fu X, et al. The neuron-specific protein TMEM59L mediates oxidative stress-induced cell death. Molecular neurobiology. 2017;54(6): 4189-4200.

57. Ha YJ, Kim CW, Roh SA, Cho DH, Park JL, Kim SY, et al. Epigenetic regulation of KLHL34 predictive of pathologic response to preoperative chemoradiation therapy in rectal cancer patients. *I*nternational Journal of Radiation Oncology* Biology* Physics. 2015;91(3): 650-658.
58. Yang C, Yu KD, Xu WH, Chen AX, Fan L, Ou ZL et al. Effect of genetic variants in two chemokine decoy receptor genes, DARC and CCBP2, on metastatic potential of breast cancer. PloS one.2013; 8(11): e78901.
59. Hu YW, Kang CM, Zhao JJ, Nie Y, Zheng L, Li HX, et al. 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.2018;22(1): 497-510.
60. Li Y, and Yang P. GPC5 gene and its related pathways in lung cancer. Journal of thoracic oncology. 2011;6(1): 2-5.
61. Wang H, Dong X, Gu X, Qin R, Jia H, & Gao J. The microRNA-217 functions as a potential tumor suppressor in gastric cancer by targeting GPC5. PLoS One. 2015;10(6): e0125474.
62. Wagner K, Grzybowska E, Butkiewicz D, Pamula-Pilat J, Pekala W, Tecza K, et al. High-throughput genotyping of a common deletion polymorphism disrupting the TRY6 gene and its association with breast cancer risk. BMC genetic. 2007;8(1): 41.
63. Wang KK, Liu N, Radulovich N, Wigle DA, Johnston MR, Shepherd FA, et al. Novel candidate tumor marker genes for lung adenocarcinoma. Oncogene. 2002;21(49): 7598.

64. Geng YJ, Xie SL, Li Q, Ma J, & Wang GY. Large intervening non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression. Journal of International Medical Research. 2011;39(6): 2119-2128.
65. 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.
66. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464(7291): 1071–1076. doi: 10.1038/nature08975.
67. Skibola CF, Bracci PM, Halperin E, Conde L, Craig DW, Agana L, et al. Genetic variants at 6p21. 33 are associated with susceptibility to follicular lymphoma. Nature genetics. 2009;41(8): 873.
68. Chimonidou M, Tzitzira A, Strati A, Sotiropoulou G, Sfikas C, Malamos N, et al. CST6 promoter methylation in circulating cell-free DNA of breast cancer patients. Clinical biochemistry. 2013;46(3): 235-240.
69. Castilla MÁ, López-García MÁ, Atienza MR, Rosa-Rosa JM, Díaz-Martín J, Pecero ML. VGLL1 expression is associated with a triple-negative basal-like phenotype in breast cancer. Endocrine-related cancer. 2014;21(4): 587-599.
70. Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, et al. The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. New England Journal of Medicine. 2003;349(26): 2483-2494.
71. Zhang ZF, Zhang HR, Zhang QY, Lai SY, Feng YZ, Zhou Y, et al. High expression of TMEM40 is associated with the malignant behavior and tumorigenesis in bladder cancer. Journal of translational medicine. 2018;16(1): 9.

72. Loughner CL, Bruford EA, McAndrews MS, Delp EE, Swamynathan S, & Swamynathan SK. Organization, evolution and functions of the human and mouse Ly6/uPAR family genes. Hum Genomics. 2016;10: 10. doi: 10.1186/s40246-016-0074-2.
73. Upadhyay G. Emerging Role of Lymphocyte Antigen-6 Family of Genes in Cancer and Immune Cells. Front Immunol. 2019;10:819. doi: 10.3389/fimmu.2019.00819.
74. 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. doi: 10.1371/journal.pone.0142618.
75. Olafsson S, McIntyre RE, Coorens T, Butler T, Jung H, Robinson PS, et al. Somatic Evolution in Non-neoplastic IBD-Affected Colon. Cell. 2020;182(3): 672-684.e11.
76. Abudoureyimu A, Maimaiti R, Magaoweiya S, Bagedati D, Wen H. Identification of long non-coding RNA expression profile in tissue and serum of papillary thyroid carcinoma. Int J Clin Exp Pathol. 2016;9(2): 1177-1185.
77. Rahiminejad S, Maurya MR, Mukund K, Subramaniam S. Modular and mechanistic changes across stages of colorectal cancer. BMC Cancer. 2022; 22**:** 436. doi: 10.1186/s12885-022-09479-3.
78. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25.
79. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation, Nucleic Acids Research. 2016;44(D1): D457–D462. doi: 10.1093/nar/gkv1070.

80. Sidiropoulos K, Viteri G, Sevilla C, Jupe S, Webber M, Orlic-Milacic M, et al. Reactome enhanced pathway visualization. Bioinformatics. 2017;33(21): 3461-3467. doi: 10.1093/bioinformatics/btx441.
81. Yang M, Lu J, Zhang G, Wang Y, He M, Xu Q, et al. CXCL13 shapes immunoactive tumor microenvironment and enhances the efficacy of PD-1 checkpoint blockade in high-grade serous ovarian cancer. Journal for ImmunoTherapy of Cancer 2021;9: e001136. doi: 10.1136/jitc-2020-001136.
82. Ren J, Lan T, Liu T, Liu Y, Shao B, Men K, et al. CXCL13 as a Novel Immune Checkpoint for Regulatory B Cells and Its Role in Tumor Metastasis. The Journal of Immunology. 2022;208(10): 2425-2435. doi: 10.4049/jimmunol.2100341.
83. 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;35(7): 1247-1248. doi: 10.1093/bioinformatics/bty763.
84. Subramanian A, Narayan R et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6): 1437-1452.e17. doi: 10.1016/j.cell.2017.10.049.
85. Zheng S, Qin F, Yin J, Li D, Huang Y, Hu L, et al. Role and mechanism of actin-related protein 2/3 complex signaling in cancer invasion and metastasis: A review. Medicine. 2023;102(14): e33158. doi: 10.1097/MD.0000000000033158.
86. Ge P, Wang W, Li L, Zhang G, Gao Z, Tang Z, et al. Profiles of immune cell infiltration and immune-related genes in the tumor microenvironment of colorectal

cancer. Biomedicine & pharmacotherapy. 2019;118: 109228. doi: 10.1016/j.biopha.2019.109228.

Supporting Information

S1 File: Dataset

S2 File: Sorted linear model - full information

S3 File: Stage-specific DEGs

S4 File: Sorted numeric model - full information

S5 File: significant MEGs

S6 File: Expression visualization of MEGs

S7 File: MEGs significant in Linear Model and Numeric Model (adj. p-val < 1E-05)

S8 File: MEGs that are in the Linear Model top200

S9 File: MEGs in the Numeric Model top200

S10 File: Overlap between the Top 200 of Linear & Numeric Models

S11 File: Stage-salient genes identified in the normals-augmented RNAseqDB COADREAD data

S12 File: Stage-salient genes identified in an external cohort (GSE39582) for benchmarking

S13 File: Univariate Cox survival analysis of all the 157 stage-salient genes

S14 File: Consensus between the stage-salient genes and the stage-specific DMGs identified using ChAMP software

S15 File: Summary of the stage-salient genes documented in the Network of Cancer Genes

S16 File: Summary of the stage-salient genes that have been used as targets or endpoints in clinical trials, as documented in ClinicalTrials.gov

S17 File: Monotonically expressed genes (MEG) identified in the external dataset GSE39582, and STRINGdb network reconstruction using the overlap between MEGs and stage-4 salient DEGs.

S18 File: Deconvolution results from Cibersort immuno-cyte profiling analysis, yielding significant samples, as well as the raw data used in the related figures.

S19 File: Figures A, B, C, D: Analysis of immune cell-type infiltration between tumor and normals, with visualization of immune ecotypes and differential distribution of immune cell-type populations.