**Supplementary Methods**

**Sample collection**

Collection of the BarcUVa-Seq samples was performed in the setting of screening colonoscopy on subjects aged 50 years or more at clinics in Spain. Subjects included in BarcUVa-Seq were part of the Colobank study. The target population included men and women who received an indication for colonoscopy at the Bellvitge University Hospital (Hospitalet de Llobregat, Catalonia, Spain) or at the Viladecans Hospital (Viladecans, Catalonia, Spain) after a positive fecal immunochemical test result (FIT; hemoglobin level of 20mg Hb/g) or after recommendation from a medical doctor. From June 2011 to December 2014, participants who agreed to take part in the Colobank study provided informed consent, completed an epidemiological questionnaire, and donated a blood sample. The Colobank study protocol was approved by the Bellvitge University Hospital Ethics Committee (PR073/11 and PR286/15) and followed national and international directives on ethics and data protection. Included subjects had no history of polyps or CRC and no family history of CRC. All subjects were European. Biopsies were obtained from macroscopically normal mucosa in the proximal, transverse, and distal colon and stored in RNAlater stabilization solution prior to freezing at -80C. RNA was extracted from frozen tissue using the mirVana kit (Thermo Fisher Scientific) after homogenization using the Minilys bead mill (Bertin Instruments). The RNA was DNAse-treated and concentrated using the RNA Clean and Concentrator-5 kit (Zymo Research). Paired-end sequencing was performed on an Illumina HiSeq 2500 instrument (High Output mode) after library production using the TruSeq Stranded Total RNA Library Prep Gold kit (Illumina). Raw sequencing data was transferred to file systems at the University of Barcelona (UB).

The search of the GDC was implemented via the GDC data portal (1). On the “Exploration” tab of the website, the following parameters were selected:

• Primary site = Colon + Rectum + Rectosigmoid junction

• Program = TCGA

• Project = TCGA-COAD + TCGA-READ

• Disease Type = Adenomas and Adenocarcinomas

On the “Repository” tab of the website, the following parameters were selected:

• Data Category = Sequencing Reads

• Experimental Strategy = RNA-Seq

The search of the SRA was implemented via the Entrez SRA homepage (2) with the following expression:

• (RNA-seq[All Fields] AND (Colon[Organism] OR colon[All Fields] OR Rectum[Organism] OR rectum[All Fields] OR Colorectal[Organism] OR colorectal[All Fields]) AND tissue[All Fields]) AND Homo sapiens[orgn] AND biomol rna[Properties]

A targeted search of the BioProject archive (3) was performed to verify completeness of the SRA search. The search of the BioProject archive was implemented via the Entrez BioProject homepage (3) with the following expression:

• (RNA-seq[All Fields] AND (Colon[Organism] OR colon[All Fields])) AND org human[Filter]

### Eligibility criteria

Samples of human tissue with an epithelial component obtained from subjects aged 18 years or more were eligible. Samples were considered healthy if histologically normal and obtained from subjects free of known history of inflammatory bowel disease or CRC or inherited predisposition to CRC or polyps. A single GTEx sample from a subject reported to have had a history of a bleeding polyp was included; otherwise, no healthy subjects had known polyps. Samples from subjects with CRC were eligible if the subject was free of known chemotherapy or radiotherapy at the time of resection, but reporting on neo-adjuvant therapies was incomplete for most studies. Tumor samples were eligible if from histologically confirmed primary adenocarcinoma, and tumor-adjacent samples were eligible if from histologically confirmed normal epithelium. In accord with TCGA Network’s approach, colon and rectum samples were pooled (4). GTEx samples from the sigmoid colon, which explicitly exclude an epithelial component (5), were not eligible. From a technical perspective, samples were eligible if they were part of a study of at least 10 samples in which RNA was extracted from flash-frozen tissue and sequenced on Illumina HiSeq 2000, 2500, or 3000 instruments to a minimum depth of 10 million reads. Single- and paired-end libraries as well as libraries prepared with and without poly-adenylated tail selection were eligible.

### Data preparation

Raw sequencing data was extracted from public archives for all GDC and SRA samples. For GDC samples, BAM files were downloaded from the GDC and converted to FASTQ with Biobambam2 (6). The reads were shuffled after conversion in order to permit quantification of transcripts with Salmon’s streaming inference algorithm (7). Paired-end reads were shuffled with fastq-shuffle (8), and single-end reads were shuffled with command-line utilities. For SRA samples, including GTEx, SRA files were downloaded and converted to FASTQ with SRAToolKit (9). For BarcUVa-Seq samples, FASTQ files were downloaded from UB servers. Technical parameters and read quality of all FASTQ files were assessed with FastQC (10), and metrics were aggregated with MultiQC (11).

Annotation files on phenotypic and technical parameters for each sample were extracted in two phases. First, raw annotations were downloaded from the GDC, SRA, and UB servers. Second, fields relevant to analysis were identified in each set of source documents, and data was extracted, aggregated, and harmonized with a combination of custom command-line utilities and Python scripts that will be made available on GitHub.

**Expression quantification**

A GRCh38 reference transcriptome was downloaded from Ensembl (12) and used to generate a quasi-mapping transcriptome index with the Salmon ‘index’ command with the ‘--type quasi’ flag set (7). A minimum k-mer threshold of 25 was chosen in order to accommodate the range of read lengths in the pooled dataset. Salmon 0.11.2 in mapping-based mode with default parameters and the ‘--gcBias’ flag was used for alignment-free, transcript-level quantification. The *tximport* package was used to convert transcripts-per-million estimates to gene-level counts (13).

**Study design**

Scatterplots of the first two principal components of the 500 most variable genes demonstrated clustering influenced primarily by dataset and library format (Supplemental digital content 3: Figure SF1A). Batch-adjustment appeared to reduce the effect of dataset but did not reduce the effect of library format (Supplemental digital content 3: Figure SF1B). Therefore, single-end libraries were reserved as a validation cohort (Cohort C, see Figure 1).

A set of 834 independent healthy (HLT), normal-adjacent-to-tumor (NAT), and tumor (TUM) paired-end samples was allocated to Cohort A for independent testing (Table 1B), and a separate set of 15 NAT-TUM paired-end matched samples (20% of all NAT-TUM paired-end matched samples) was allocated to Cohort B for repeated-measures testing (Table 1C). The 275 independent single-end samples were allocated to Cohort C for independent testing (Table 1D). Within- and between-cohort independence of samples was ensured by careful NAT-TUM matched sample allocation as described below. No subject was duplicated across any analysis cohorts, and only Cohort B included samples from the same subject more than once.

To ensure within and between cohort independence of samples in Cohorts A and B, NAT-TUM pairs were allocated as follows: The 15 pairs reserved for repeated-measures testing were excluded from Cohort A. The remaining pairs were split and, for TCGA pairs, the NAT sample was always assigned to Cohort A; for non-TCGA pairs, the NAT or the TUM sample was assigned to Cohort A at random. Unassigned NAT and TUM samples were excluded from all analyses. In order to preserve such strict independence, 60 samples from duplicated subjects were excluded from all analyses, which accounts for the difference in total samples between Table 1A and Tables 1B and C. For all NAT-TUM pairs in Cohort C, the NAT or the TUM sample was assigned at random.

### Normalization and adjustment for batch effects

Between sample count normalization was performed with the median of ratios method as implemented in *DESeq2* (14). The effects of confounding variables (i.e. variables besides the phenotype of interest) were estimated in Cohorts A and C with factor analysis as implemented in *SVA* (15). The cumulative effects of all potentially confounding factors, including technical heterogeneity across datasets as well as age and race differences across samples, were considered latent factors. This approach was taken for two reasons: 1) some datasets contained samples of a single phenotype, rendering the effects of phenotype and a potential adjustment variable for dataset linearly dependent and therefore not amenable to reliable simultaneous estimation and 2) some datasets did not include sample-level demographic annotations, which would have necessitated the exclusion of samples with missing data from analysis, reducing power for discovery. Therefore, batch effects were estimated with a full model using phenotype (i.e. healthy, tumor-adjacent, tumor) alone without adjustment variables and a null model with an intercept term alone without adjustment variables to explain gene expression. Predictors of gene expression besides phenotype were derived from the residuals matrix of the full model fit. These predictors were ultimately represented as surrogate variables.

To choose an appropriate number of surrogate variables to estimate, two complementary approaches were used. First, the two methods implemented in *SVA*, the Buja-Eyuboglu (BE) method and the Leek method, were used to estimate the number of significant latent factors. Second, dimension reduction was employed to visually assess the influence of latent factors. For dimension reduction, principal component (PC) analysis as implemented in *DESeq2* and t-distributed stochastic neighbor embedding (tSNE) as implemented in *Rtsne* were used (16). Five surrogate variables were included in downstream analysis based on the BE latent factor estimate of two and evaluation of scatter plots of the first two PCs of global gene expression (Supplemental digital content 3: Figure SF2), which demonstrated removal of three additional surrogate variables resulted in more homogenous mixing of datasets. For dimension reduction and all other visualization methods, unless otherwise indicated, batch-adjusted counts were used. To derive batch-adjusted counts, normalized counts were transformed with the variance stabilizing transformation (VST) in *DESeq2*, and then *limma* was used to estimate effect sizes for surrogate variables and subtract them from transformed counts (17).

### Regression models

*DESeq2*, which fits negative binomial regression models, was used for all statistical tests of differential gene expression (DGE). Cohort A was used for the primary analysis of DGE across phenotypes. Five surrogate variables and phenotype were included as covariates. No additional adjustment variables were used to model gene expression in Cohort A. Contrasts were set to make the following three comparisons: 1) normal-adjacent-to-tumor (NAT) versus healthy (HLT), 2) tumor (TUM) versus NAT, and 3) TUM versus HLT. Cohort B was used to validate the modeling of batch effects in Cohort A. Repeated-measures regressions on the NAT-TUM pairs with subject identifier and sample phenotype as covariates were performed. No additional adjustment variables were used to model gene expression in Cohort B; confounding factors were controlled by the paired design. Cohort C was used to validate DGE results from Cohort A. DGE was tested using the same methods employed for Cohort A. For all regressions, genes were pre-filtered to include only those with a minimum count of 10 in at least 33% of samples. Differential expression was defined with an absolute fold change threshold of two, and the Benjamini-Hochberg false discovery rate (FDR) was set to 5%.

### Gene set over-representation

To identify biological processes in DGE results, gene lists were tested for over-representation in Molecular Signatures Database hallmark gene sets (18) using *fgsea* (19), which implements pre-ranked Gene Set Enrichment Analysis (GSEA) (20). Genes were ranked by test statistics generated by *DESeq2*.

**Transcription factor prediction**

To identify potential drivers of DGE results, gene lists were used to predict the transcription factors most likely driving their differential expression using binding analysis for regulation of transcription (BART), which is a computational method for predicting factors that regulate a given gene set (21).

**Patterns in gene-level variation across phenotypes**

Among differentially expressed genes between NAT samples and samples from the other two phenotypes, the following four expression patterns were defined: gradient, TUM-associated, NAT-specific, and HLT-associated. The gradient set included genes that were differentially expressed in the same direction across NAT and HLT samples and across TUM and NAT samples, where NAT sample expression levels were between HLT and TUM sample levels (Figure 4A *RPS21* and *NBEAL1*). The TUM-associated set included genes that were differentially expressed in NAT relative to HLT samples but not relative to TUM samples (Figure 4A *YWHAE* and *RUFY2*). The NAT-specific set included genes that were differentially expressed in NAT relative to HLT and TUM samples but not between HLT and TUM samples (Figure 4A *B3GNT8* and *WDR90*). The HLT-associated set included genes that were differentially expressed in NAT relative to TUM samples but not relative to HLT samples (Figure 4A *UGP2* and *SF3B3*). Hierarchical clustering was performed on all genes in a given pattern or the top 500 most confidently estimated genes in that pattern (based on FDR-adjusted p values in NAT vs HLT comparisons) when pattern sets exceeded 500 genes.

**Age as a potential confounder**

Given the substantial difference in mean age across phenotypes in Cohort A (Table 1B), scatterplots and a histogram of age-expression correlations for all field effect genes were generated. A weak but positive correlation (Pearson r of approximately 0.2) between age and expression was observed (Supplemental digital content 3: Figure SF3, SF4A). Since age was also positively correlated with the NAT and TUM phenotypes (Supplemental digital content 3: Figure SF4B), linear models of expression versus age and phenotype were fit, and test statistics for the predictors were compared (Supplemental digital content 3: Figure SF4C). These models used normalized, transformed, and batch- adjusted counts to represent expression, so linear models were appropriate. Boxplots of test statistics demonstrated the minor and inconsistent effect of age on expression and highlighted the strong, consistently positive effect of phenotype.

**Spatial limits of field effect**

Of the datasets in which distance from tumor for NAT samples was reported, all reported only a minimum distance from tumor at or beyond which NAT samples were obtained; some did not report distances at all. A qualitative evaluation of the scope of field effect was attempted by plotting gene counts across varying minimum distances from tumor for five of the 20 validated field effect genes, but no meaningful patterns were observed (Supplemental digital content 3: Figure SF5).

### Field effect can influence utility of controls

Based on the observation that important biological pathways were dysregulated in NAT samples in TUM-like patterns, the possibility that some TUM-specific molecular features could be masked by the field effect was tested. TUM-specific DGE results generated using first HLT samples and then NAT samples as controls were compared. Overall results were similar (Pearson correlation for LFC r=0.75, p<2.20e-16), as previously demonstrated (22). Nevertheless, a subset of genes uniquely identified as differentially expressed (DE) with HLT samples as controls was observed (green points in Figure 5A). In total, 3856 genes dysregulated in TUM relative to HLT samples were not identified as DE when NAT samples were used as controls. The smaller number of NAT samples may have accounted for some of the difference, but the field effect likely also contributed. To assess the significance of the masking due to field effect, GSEA among the 3856 masked TUM-specific genes was performed. The three gene sets with highest NES were the same as those identified in enrichment analysis of TUM-specific DGE using NAT samples as controls (Figure 5B and Supplemental digital content 2: Table ST8; compare with Figure 3B and Supplemental digital content 2: Table ST5B). In fact, seven of the eight sets with highest NES were the same in both analyses. The consistency suggested that although particular TUM-specific genes may be missed using NAT samples as controls, underlying biological processes can still be accurately characterized. On the other hand, oxidative phosphorylation and fatty acid metabolism sets, which had two of the lowest NES in the TUM verses NAT comparison (NES=-1.75, p-adj=2.20E-03; NES=-1.44, p-adj=1.74E-02, respectively) were significantly enriched in TUM samples in the analysis of masked genes (NES=2.09, p-adj=9.44E-04; NES=1.71, p-adj=3.48E-02, respectively). The divergence suggested that the field effect could mask important TUM-specific metabolic features.

Finally, in order to identify novel TUM-specific genes unique to comparisons using HLT samples as controls, a subset of 62 relatively under-annotated genes was identified by filtering the TUM versus HLT DGE results not intersected by the TUM verses NAT DGE results for gene symbols including “orf” or beginning with “LOC”. The direction of variation was confirmed in Cohort C for 23 of these novel TUM-specific genes (Table 2B). Another 28 had detectable but statistically indeterminate expression, as indicated by a test statistic of zero, and 11 were not detected. None had an opposite direction of variation. Among validated genes, none had a convincing association with stage (Supplemental digital content 3: Figure SF6), but *C9orf16* was previously shown to be prognostic in CRC (Figure 5C,D), and *C7orf50* was shown to be prognostic in pancreatic cancer (23). In order to test whether the prognostic utility of *C9orf16*, which had previously been demonstrated with a log-rank test comparing univariable Kaplan-Meier estimates, was confounded by age or stage, a Cox proportional hazards model was fit using the original dataset downloaded from the Human Protein Atlas(23) and the *survival*(24) package in R. In the model, *C9orf16* expression (high versus low), age, and tumor stage were included as covariates. Age, stage III, stage IV, and *C9orf16* expression were all significant predictors of survival (p=1.32e-05, p=0.00206, p=1.56e-07, p=0.00278, respectively). For completeness, the *survminer* (25) package was used to generate stratified Kaplan-Meier curves for each stage as well as Cox fits wherein each stage was held constant and age was set to the median in order to demonstrate the effect of expression on survival (Supplemental digital content 3: Figures SF7, SF8).

**Supplementary Figure Legends**

**SF1: Persistent batch effect of library format.** **A)** Scatterplots of first two PCs for normalized but batch-unadjusted counts across all independent samples eligible for analysis and colored by dataset, library format, and sex to highlight likely drivers of observed clustering. **B)** Same as **(A)** but for batch-adjusted counts using five surrogate variables in the model. Library format is a persistent driver of batch effect.

**SF2: Choice of five surrogate variables.** **A)** Scatterplots of first two PCs for normalized but batch-unadjusted counts across all samples in Cohort A and colored by several factors to highlight likely drivers of observed clustering. **B)** Same as **(A)** but for batch-adjusted counts using two surrogate variables in model. **C)** Same as **(A)** but for batch-adjusted counts using five surrogate variables in the model.

**SF3: Expression versus age.** Scatterplots of adjusted counts versus age in Cohort A for the 20 field effect genes validated in Cohort C. Each point represents a single sample, and points are colored by the phenotype of that sample to demonstrate the influence of phenotype on the correlation with age.

**SF4: Age and phenotype as predictors of expression.** A) Histogram of genewise Pearson correlation coefficients between age and expression for 1000 randomly sampled genes from the Cohort A expression matrix of adjusted counts (blue) overlaid with the same histogram for the 20 field effect genes (red). B) Histograms of age stratified by phenotype in Cohort A. C) Boxplots of the linear model test statistics for covariates used to estimate adjusted counts for the 20 field effect genes in Cohort A.

**SF5: Spatial dimension of field effect. A)** Boxplots of adjusted counts for arbitrary subset of validated field effect genes across different distances from tumor. Counts from NAT and TUM samples from Cohort A. **B)** Boxplots of adjusted counts for same genes as **(A)** across different distances from tumor but from NAT and TUM samples from Cohort C.

**SF6: Expression versus stage.** Boxplots of unadjusted but normalized and transformed counts versus stage in Cohort A for the 23 novel TUM-specific genes validated in Cohort C. Each point represents a single sample. Transformation was executed with variance stabilizing transformation of DESeq2. Unadjusted counts shown due to possibility of removing effect of stage on expression during batch adjustment procedure.

**SF7: Kaplan-Meier survival.** Survival curves stratified by stage to demonstrate the consistent effect of expression across stage. Curves not adjusted for age. Blue curve represents low-expression group; gold curve represents high-expression group.

**SF8: Cox proportional hazards survival.** Survival curves inferred from the Cox model fit and plotted with age and stage held constant to demonstrate the effect of expression across stage. Curves based on a model adjusted for age and stage. Blue curve represents low-expression group; gold curve represents high-expression group.

**Supplementary Table References**

In Table ST1C from Supplemental digital content 2, Table ST1, the Publications field refers to the following references:

Peters *et al*., 2017 (26)

Knight *et al*., 2016 (27)

Kim *et al*., 2014 (28)

Purcell *et al*., 2017 (29)

Druliner *et al*., 2016 (30)

Li *et al*., 2018 (31)

Delker *et al*., 2014 (32)

Hanley *et al*., 2017 (33)

Cima *et al*., 2016 (34)

Zhu *et al*., 2018 (35)

Lee *et al*., 2016 (36)

**Supplementary References**

1. Genomic Data Commons Data Portal [Internet]. 2019 [cited 10/20/2018]. Available from: <https://portal.gdc.cancer.gov/>.

2. SRA [Internet]. 2019 [cited 1/21/2019]. Available from: <https://www.ncbi.nlm.nih.gov/sra>.

3. BioProject [Internet]. NCBI. 2019 [cited 1/21/2019]. Available from: <https://www.ncbi.nlm.nih.gov/bioproject>.

4. Network TCGA, Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330 EP.

5. BBRB-PR-0004-W1-G3 GTEx Organ Retrieval, Dissection, and Preservation Details Table [Internet]. NCI. 2015 [cited 10/1/2018]. Available from: <https://biospecimens.cancer.gov/resources/sops/docs/GTEx_SOPs/BBRB-PR-0004-W1-G3%20GTEx%20Organ%20Retrieval,%20Dissection,%20and%20Preservation%20Details%20Table%20.pdf>.

6. Jackson DaW, Andrew and Tischler, German. Biobambam. Online: Sanger; 2019.

7. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods. 2017;14(417).

8. Forster F. fastq-shuffle. GitHub: chloroExtractorTeam; 2018.

9. Team STD. SRAToolKit. Online: NCBI; 2019.

10. Andrews S. FastQC. 0.11.8 ed. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>: Babraham Bioinformatics; 2019.

11. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-8.

12. Ensembl. GRCh38 cDNA Reference Sequences. In: Ensembl, editor. <ftp://ftp.ensembl.org/pub/current_fasta/homo_sapiens/cdna/>: Ensembl; 2019.

13. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 1; peer review: 2 approved]. F1000Research. 2015;4(1521).

14. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550.

15. Leek J, Storey J. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics. 2007;3(9).

16. Krijthe J. Rtsne. Online: GitHub; 2019.

17. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47-e.

18. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25.

19. Sergushichev A. Fast Gene Set Enrichment Analysis (fgsea). 1.1.3 ed. CT Lab GitHub Repository (<https://github.com/ctlab>): GitHub; 2016.

20. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.

21. Wang Z, Civelek M, Miller CL, Sheffield NC, Guertin MJ, Zang C. BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. Bioinformatics. 2018;34(16):2867-9.

22. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nature Communications. 2017;8.

23. Uhlen M, Zhang C, Lee S, Sjostedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357(6352).

24. Therneau TM. A Package for Survival Analysis in R. R package version 3.1-11 ed. <https://CRAN.R-project.org/package=survival>: <https://CRAN.R-project.org/>; 2020.

25. Kassambara A, Kosinski M, Biecek P, Fabian S. survminer. 0.4.6 ed. <https://CRAN.R-project.org/package=survminer>: <https://CRAN.R-project.org/>; 2019-09-03.

26. Peters LA, Perrigoue J, Mortha A, Iuga A, Song W-m, Neiman EM, et al. A functional genomics predictive network model identifies regulators of inflammatory bowel disease. Nature Genetics. 2017;49:1437.

27. Knight JM, Kim E, Ivanov I, Davidson LA, Goldsby JS, Hullar MA, et al. Comprehensive site-specific whole genome profiling of stromal and epithelial colonic gene signatures in human sigmoid colon and rectal tissue. Physiol Genomics. 2016;48(9):651-9.

28. Kim SK, Kim SY, Kim JH, Roh SA, Cho DH, Kim YS, et al. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. Mol Oncol. 2014;8(8):1653-66.

29. Purcell RV, Visnovska M, Biggs PJ, Schmeier S, Frizelle FA. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. Sci Rep. 2017;7(1):11590.

30. Druliner BR, Ruan X, Johnson R, Grill D, O'Brien D, Lai TP, et al. Time Lapse to Colorectal Cancer: Telomere Dynamics Define the Malignant Potential of Polyps. Clin Transl Gastroenterol. 2016;7(9):e188.

31. Li M, Zhao LM, Li SL, Li J, Gao B, Wang FF, et al. Differentially expressed lncRNAs and mRNAs identified by NGS analysis in colorectal cancer patients. Cancer Med. 2018;7(9):4650-64.

32. Delker DA, McGettigan BM, Kanth P, Pop S, Neklason DW, Bronner MP, et al. RNA sequencing of sessile serrated colon polyps identifies differentially expressed genes and immunohistochemical markers. PLoS One. 2014;9(2):e88367.

33. Hanley MP, Hahn MA, Li AX, Wu X, Lin J, Wang J, et al. Genome-wide DNA methylation profiling reveals cancer-associated changes within early colonic neoplasia. Oncogene. 2017;36(35):5035-44.

34. Cima I, Kong SL, Sengupta D, Tan IB, Phyo WM, Lee D, et al. Tumor-derived circulating endothelial cell clusters in colorectal cancer. Sci Transl Med. 2016;8(345):345ra89.

35. Zhu Y, Lu H, Zhang D, Li M, Sun X, Wan L, et al. Integrated analyses of multi-omics reveal global patterns of methylation and hydroxymethylation and screen the tumor suppressive roles of HADHB in colorectal cancer. Clin Epigenetics. 2018;10:30.

36. Lee JR, Kwon CH, Choi Y, Park HJ, Kim HS, Jo HJ, et al. Transcriptome analysis of paired primary colorectal carcinoma and liver metastases reveals fusion transcripts and similar gene expression profiles in primary carcinoma and liver metastases. BMC Cancer. 2016;16:539.