Supplementary materials and methods

Single cell preparation

Fresh tissue of each region was washed with 10 mL cold PBS for three times. Necrotic tissues were excluded artificially under microscope. Then the tissues were cut into small pieces by surgical scissors. 2 mg/mL Collagenase/Dispase (Roche, cat. # 10269638001) was used to digest the tissues into single cells at 37℃ for 30-40 min. Then the single cell suspension was centrifuged at 800g for 5min at 4℃. The cell pellets were washed with cold PBS three times and finally re-suspended in 1% HAS (Human Serum Albumin). Single cells were picked into 2μl scRNA-seq lysis buffer (0.0475% Triton X-100 (Sigma-Aldrich, X100), 0.1 U/μL RNase Inhibitor (Takara), 2.5 μM dNTP mixture (Thermo) and 2.5 μM Transcription reverse (RT) primers) gently with mouth pipette. And the remaining cell pellets were used to extract genomic DNA according to the manufacturer protocol of DNeasy Blood & Tissue Kit (QIAGEN).

Single cell RNA-seq library construction

Single cell RNA-seq libraries were carried out using STRT-seq protocol (refrerence). We modified the protocol to make it compatible to multiplexed single cell sequencing as mentioned in the published papers by our group[1,2]. Briefly, we add 8bp barcode sequence to the RT primer (5'-TCAGACGTGTGCTCTTCCGATCT-XXXXXXXX-NNNNNNNN-T25-3', X represents the 8bp predesigned barcodes, 'N' stands for the UMI (unique molecular identifier). We lysed the single cells at 72℃ for 3min and then add 2.85μl RT Mix (5 U RNase Inhibitor; 40 U SuperScript II reverse transcriptase (Invitrogen); 5× Superscript II first-strand buffer (Invitrogen); 25 mM dithiothreitol; 5 M betaine (Sigma-Aldrich); 30 mM MgCl2 (Sigma-Aldrich); 1.75 μM TSO primer (same sequence as described in STRT protocol)). Reverse transcription was performed in thermocycler using program as follows: at 25 °C for 5 min; at 42 °C for 60 min; at 50 °C for 30 min; at 70 °C for 10 min. Then 7.5μl PCR Mix (6.25 μL 2× KAPA HiFi

Li J, *et al. Gut* 2020; 69:1283–1293. doi: 10.1136/gutjnl-2019-319438

HotStart ReadyMix (KAPA); 300 nM ISPCR oligo primer (5'-AAGCAGTGGTATCAACGCAGAGT-3'); 1 μM 3′ Anchored oligo primer (5'-Oligo-GTGACTGGAGTTCA GACGTGTGCTCTTCCGATC-3')) to amplify the cDNA. The cDNA was amplified at 98℃ for 20 s; 65℃ for 30 s; 72℃ for 5 min; 14 cycle of 98℃ for 20 s, 67℃ for 15 s, 72 ℃ for 5 min; 72 ℃ for 5 min. The cDNA of cells with different barcodes can be pooled together for next steps. The pooled DNA was purified once with DNA Clean & Concentrator-5 (DC2005; Vistech) and twice with 0.8× Ampure XP beads (Beckman, A63882). Next, we use bio- tinylated index primer (5'-/Biotin/CAAGCAGAAGACGG CATACGAGAT-index-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-3') and ISPCR primer to amplify the pooled DNA. Then the biotin-labeled DNA was fragmented into 300-bp using ultra-sonication. The fragmented DNA with biotin was enriched by the affinity interaction between biotin and Dynabeads® MyOne™ Streptavidin C1 beads (Invitrogen, cat. #65602). Then the fragmented DNA inserts were library constructed according to the manufacturer procedures of KAPA Hyper Prep Kits with PCR Library Amplification/Illumina series (KAPA, cat. KK8054) and sequencing on Illumina HiSeq 4000 for obtaining 150 bp paired end reads.

Whole genome sequencing

The extracted genomic DNA for each region was fragmented by sonication into 300-bp and then was used to construct the sequencing library according to the manufacturer procedures of KAPA Hyper Prep Kits with PCR Library Amplification/Illumina series (KAPA, cat. KK8054) and sequenced on Illumina HiSeq 4000 for obtaining 150-bp paired end reads.

Hematoxylin-eosin staining and Immunohistochemistry

The sampled tissues were fixed in neutral buffered formalin and embedded in paraffin. H&E staining and immunohistochemistry staining were performed on 5-um thick slides according to standard procedures. For immunohistochemistry staining,

Li J, *et al. Gut* 2020; 69:1283–1293. doi: 10.1136/gutjnl-2019-319438

secondary horseradish peroxidase (HRP)-conjugated antibodies were used to reveal the color in combination with DAP Peroxidase substrate. The primary antibodies used included Anti-Ki67 antibody (1:3000) (Abcam, ab15580), Anti-TGFBI antibody (1:250) (Abcam, ab170874) and Anti-CD8 antibody (1:200) (Abcam, ab17147). We captured the images with a Nikon Eclipse 90i and the pathological grades of tumors were assessed by two independent pathologists.

Quality control, mapping and quantification of single cell RNA-Seq data

Paired-end sequence data was firstly split according to the cell barcode. Low quality reads, TSO and poly A were removed and then the cleaned reads were mapped to human genome (hg19) by TopHat[3]. Only unique mapping reads were kept for further analysis. By using HTSeq[4], transcript abundance of each gene was calculated and reads with unique UMIs of each genes were used to measure the expression levels of genes. The expression levels of genes were normalized to the transcripts per million (TPM). Cells that expressed less than 1000 genes or showed similar expression pattern with less than 2 samples (pearson correlation < 0.6) were filtered. Genes expressed by less than 5 cells were removed for subsequent analysis.

Quality control and mapping for whole genome sequence

Low quality and adapter contaminate reads were firstly removed by applying quality control pipeline. The cleaned reads were mapped to human genome (hg19) with BWA (v0.7.12) and duplicated reads were marked by picard “MarDuplicateds”. As for copy number variants, mapped reads of each 10M window were calculated and normalized by the total mapped read depths of each samples. For each 10M window, reads were further normalized by dividing the average reads of diploid samples and the results were visualized by R ggplot2 package.

Whole exome sequence and SNV calling

First low quality and adapter contaminate reads were removed by applying quality control pipeline. Clean reads were mapped to human genome (hg19) through BWA.

Li J, *et al. Gut* 2020; 69:1283–1293. doi: 10.1136/gutjnl-2019-319438

GATK (Genome Analysis Toolkit, Version 3.8) was performed on duplicates marked reads. Haplotype[5] and Mutect2[6] were used to call germline and somatic mutations respectively according to the online manual. "PASS" somatic mutations were filtered for further analysis. For each germline mutation, the reads sequenced more than 30X were kept for further analysis. For somatic mutations, WES data of blood from each patient were used as control. cnLOH includes the various loss of heterogeneity (LOH) events including, copy gain, copy loss and copy-neutral, and it was inferred from whole exome sequencing (WES) by using 'Sequenza' tool according to the online manual (https://cran.r-project.org/web/packages/sequenza/vignettes/sequenza.html)[7]. Sanger sequencing was used to verify the germline and somatic mutation sites in APC, as well as the germline mutations in *MUTYH*. The primers used for Sanger sequencing were summarized in Supplementary figure 3.

Prediction of potential pathogenic mutations

For the somatic mutations called by Mutect2, further filtrations were used for prediction of potential pathogenic mutations. Firstly, synonymous mutations and mutations in noncoding regions were excluded. Then, non-synonymous mutations that would cause amino acid changes or have "High effect" on the protein were kept. Finally, the filtered mutation sites on recurrently mutated genes in CRC were supposed to be potential pathogenic mutations (Supplementary table 4).

Tumor phylogenetic tree reconstruction

The somatic mutations of each patient were used for the construction of tumor phylogenetic tree. The 'dist.gene' function of 'ape' package was used to calculated the distance between different lesion of each patient. Then the "nj" function was used to construct the phylogenetic tree, which was finally displayed by running "plot.phylo" function.

tSNE clustering based on the transcription factor regulation network

The python package pySCENIC[8] was used to construct the transcription factor

Li J, *et al. Gut* 2020; 69:1283–1293. doi: 10.1136/gutjnl-2019-319438

regulation network following the online manual (https://github.com/aertslab/pySCENIC)[8]. The cellular regulon enrichment matrix was used to calculate the distance between samples. Non-linear dimensional reduction was analyzed by performing Rtsne package based on the distance matrix. Cell clusters were identified based on the first nine PCA dimensions and annotated by known cell type markers.

PCA and tSNE clustering based on the transcriptome data

The FactorMinR package was used to perform the PCA analysis based on the top 500 high variants genes and the "RunTSNE" function in Seurat package was used to perform the tSNE clustering with parameters dims.use=1:10, do.fast=T.

Identification of differentially expressed genes

The "FindAllMarkers" function of Seurat package[9] were used to identified DEGs among different groups with setting the parameters as "min.pct = 0.25, only.pos = T, logfc.threshold = 1.5".

Pseudotime analysis

Firstly, DEGs between adjacent normal tissue and tumor were identified by Seurat package. Then Monocle2[10] was used to order cells according to the progression of tumorigenesis based on the DEGs between normal and tumor tissues. We first reduced the dimension of our dataset by running "reduceDimension" function and then "orderCells" function was used to order cells along the pseudotime trajectory. The DEGs between epithelial cells from adjacent normal tissue and multi-grade lesions, as well as the TCA related genes (from KEGG https://www.genome.jp/kegg/pathway.html) were plotted with 'plot_pseudotime_heatmap' function in monocle package with setting 'show_rownames = T, num_clusters=5'

Li J, *et al. Gut* 2020; 69:1283–1293. doi: 10.1136/gutjnl-2019-319438

References

1 Dong J, Hu Y, Fan X, *et al.* Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome Biol* 2018;**19**:1–20. doi:10.1186/s13059-018-1416-2

2 Islam S, Kjällquist U, Moliner A, *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* 2011;**21**:1160–7. doi:10.1101/gr.110882.110

3 Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;**25**:1105–11. doi:10.1093/bioinformatics/btp120

4 Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;**31**:166–9. doi:10.1093/bioinformatics/btu638

5 Depristo MA, Banks E, Poplin R, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491–501. doi:10.1038/ng.806

6 Cibulskis K, Lawrence MS, Carter SL, *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;**31**:213–9. doi:10.1038/nbt.2514

7 Favero F, Eklund AC, Joshi T, *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* Published Online First: 2014. doi:10.1093/annonc/mdu479

8 Aibar S, Bravo González-Blas C, Moerman T, *et al.* SCENIC: Single-Cell Regulatory Network Inference And Clustering. *bioRxiv* 2017;:1–41. doi:10.1101/144501

9 Satija R, Farrell JA, Gennert D, *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**:495–502. doi:10.1038/nbt.3192

10 Trapnell C, Cacchiarelli D, Grimsby J, *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**:381–6. doi:10.1038/nbt.2859

179

Li J, *et al. Gut* 2020; 69:1283–1293. doi: 10.1136/gutjnl-2019-319438