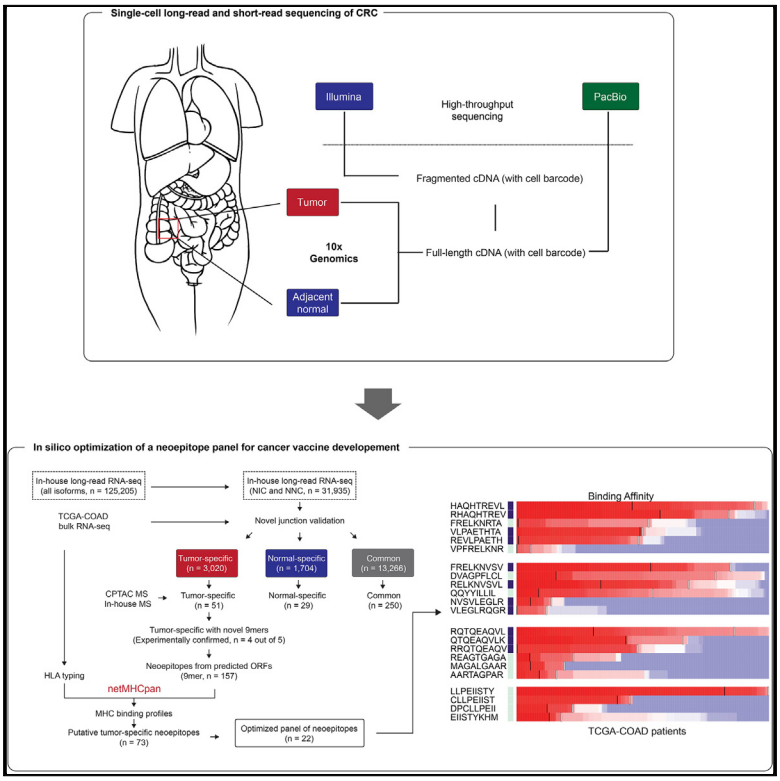
Cell Genomics

Resource

An isoform-resolution transcriptomic atlas of colorectal cancer from long-read single-cell sequencing

Graphical abstract



Highlights

- Provides a single-cell full-length transcriptomic atlas of CRC for future research
- Unveils isoform-level dysregulation in CRC
- Facilitates the development of neoantigen-based cancer vaccines for CRC

Authors

Zhongxiao Li, Bin Zhang, Jia Jia Chan, ..., Ker-Kan Tan, Xin Gao, Yvonne Tay

Correspondence

bin.zhang@kaust.edu.sa (B.Z.), xin.gao@kaust.edu.sa (X.G.), yvonnetay@nus.edu.sg (Y.T.)

In brief

Our study provides a single-cell full-length transcriptomic atlas of colorectal cancer (CRC), identifying tumor-specific/-associated RNA isoforms and highlighting their potential in generating shared neoantigens across CRC patients. This may facilitate the development of universal neoantigen-based cancer vaccines for CRC.

Li et al., 2024, Cell Genomics *4*, 100641
September 11, 2024
https://doi.org/10.1016/j.xgen.2024.100641

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Resource

An isoform-resolution transcriptomic atlas of colorectal cancer from long-read single-cell sequencing

Zhongxiao Li,[1,2,3,12] Bin Zhang,[1,2,3,12,*] Jia Jia Chan,[4,12] Hossein Tabatabaeian,[4,11] Qing Yun Tong,[4] Xiao Hong Chew,[4] Xiaonan Fan,[4] Patrick Driguez,[5] Charlene Chan,[4] Faith Cheong,[4] Shi Wang,[6] Bei En Siew,[7] Ian Jse-Wei Tan,[7,8] Kai-Yin Lee,[7,8] Bettina Lieske,[7,8] Wai-Kit Cheong,[7,8] Dennis Kappei,[4,9,10] Ker-Kan Tan,[7,8] Xin Gao,[1,2,3,*] and Yvonne Tay[4,9,10,13,*]

[1]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia
[2]Center of Excellence for Smart Health (KCSH), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia
[3]Center of Excellence on Generative AI, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia
[4]Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599, Singapore
[5]Core Labs, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia
[6]Department of Pathology, National University Health System, Singapore 119228, Singapore
[7]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore
[8]Division of Colorectal Surgery, University Surgical Cluster, National University Health System, Singapore 119228, Singapore
[9]NUS Centre for Cancer Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore
[10]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore
[11]Present address: Peter MacCallum Cancer Center, Melbourne, Australia
[12]These authors contributed equally
[13]Lead contact
*Correspondence: bin.zhang@kaust.edu.sa (B.Z.), xin.gao@kaust.edu.sa (X.G.), yvonnetay@nus.edu.sg (Y.T.)
https://doi.org/10.1016/j.xgen.2024.100641

SUMMARY

Colorectal cancer (CRC) ranks as the second leading cause of cancer deaths globally. In recent years, short-read single-cell RNA sequencing (scRNA-seq) has been instrumental in deciphering tumor heterogeneities. However, these studies only enable gene-level quantification but neglect alterations in transcript structures arising from alternative end processing or splicing. In this study, we integrated short- and long-read scRNA-seq of CRC samples to build an isoform-resolution CRC transcriptomic atlas. We identified 394 dysregulated transcript structures in tumor epithelial cells, including 299 resulting from various combinations of splicing events. Second, we characterized genes and isoforms associated with epithelial lineages and subpopulations exhibiting distinct prognoses. Among 31,935 isoforms with novel junctions, 330 were supported by The Cancer Genome Atlas RNA-seq and mass spectrometry data. Finally, we built an algorithm that integrated novel peptides derived from open reading frames of recurrent tumor-specific transcripts with mass spectrometry data and identified recurring neoepitopes that may aid the development of cancer vaccines.

INTRODUCTION

The heterogeneity of cells and their composition is critical for cancer progression, patient survival, and response to therapy.[1,2] Single-cell RNA sequencing (scRNA-seq) enables RNA expression profiling for thousands of genes in hundreds and thousands of cells in parallel with facilitating the comprehensive characterization of these variations. scRNA-seq has been performed for colorectal cancer (CRC) to investigate tumor cell heterogeneities, transformation states, spatial organizations, and immune cell infiltrations.[3–6] Based on the gene expression profiles in each cell, scRNA-seq can distinguish malignant from infiltrated cells and group them into subpopulations with distinct molecular properties. Furthermore, the varying compositions of each cell subpopulation for both malignant and infiltrated cells have been utilized to classify tumors into subtypes to predict prognosis and response to therapy.[7,8]

To date, most scRNA-seq cancer studies only quantify expression at the gene level. However, the majority of human genes can generate multiple transcript isoforms, particularly in tumor cells where widespread alterations in transcript structure arising from 5′ and 3′ ends, and alternative splicing (AS) differences, are frequently observed.[9–11] Recently, we reported that pervasive splicing within the 3′-untranslated region (3′ UTR) is upregulated

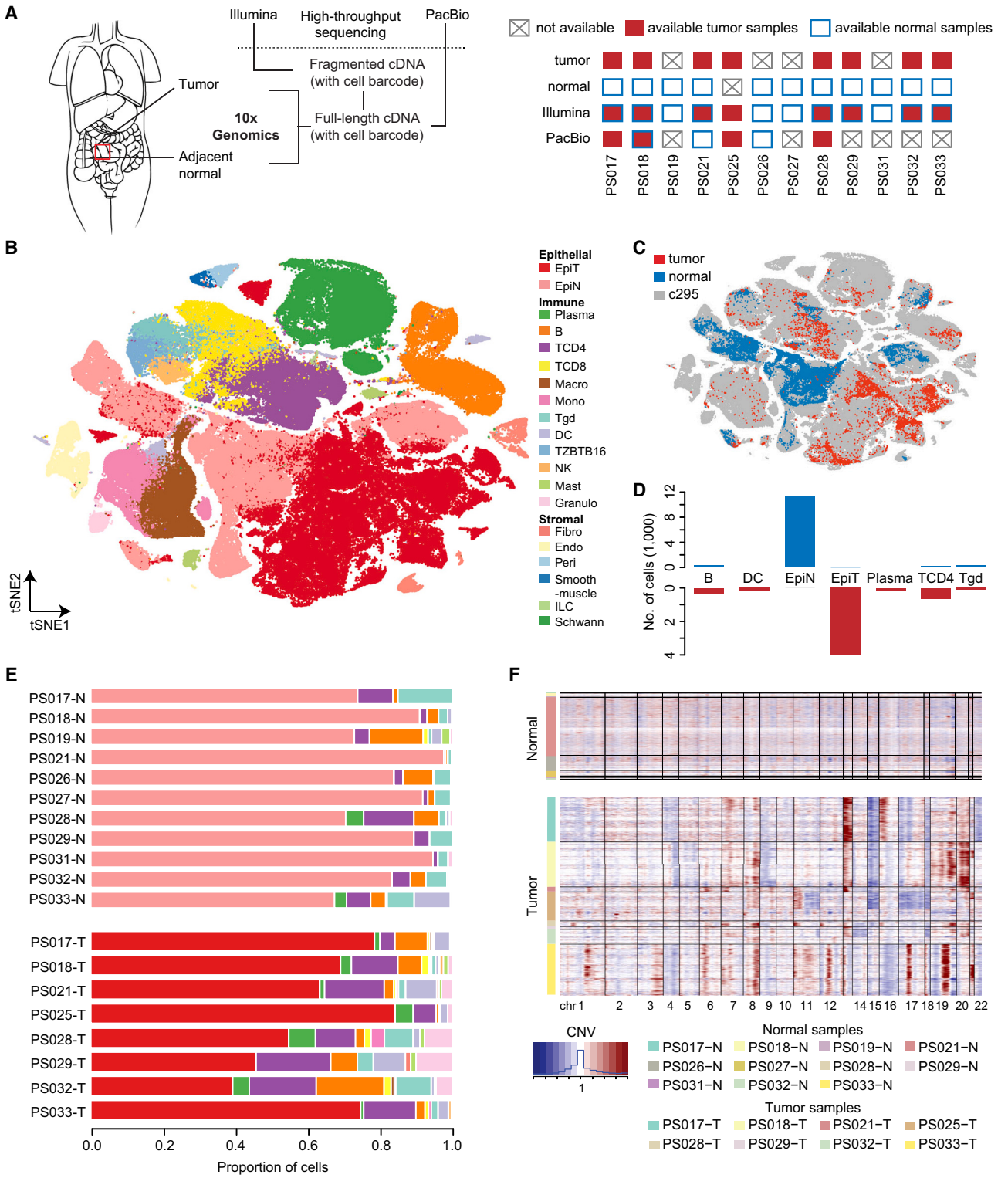


Figure 1. The short-read single-cell transcriptomic atlas of human CRC

(A) Schematic illustration of the workflow of long-read (PacBio) and short-read (Illumina) single-cell RNA sequencing (scRNA-seq) on tumor and normal samples from 12 CRC patients.

(legend continued on next page)

in cancer to promote oncogene expression and tumorigenesis.[12] These results demonstrate that in addition to gene expression, changes in transcript structures also contribute to cancer pathogenesis. Traditionally, bulk RNA sequencing (RNA-seq) detects cancer-associated aberrations in specific regions of transcripts reconstructed from short reads, whereas the third-generation long-read sequencing (LR-seq), such as PacBio Iso-Seq, is able to directly capture dysregulations of full-length transcripts. Recently, two studies employed PacBio Iso-Seq with bulk RNA from breast cancer samples and gastric cancer cell lines to characterize their full-length transcripts, revealing extensive aberrant transcript structures that may be involved in tumor-related dysfunctions.[13,14] However, only a handful of studies have combined LR-seq and scRNA-seq to dissect tumor cell heterogeneities and transcript dysregulations at the isoform level.

In addition, tumors specifically express RNA with dysregulated transcript structure (DTS) that may contain new open reading frames (ORFs) encoding unique proteins that serve as a potential source of neoantigens in tumor cells.[15] Several studies have reported that tumor-associated intron retention, neo-splice junction, exitron, and RNA editing could generate neoantigens.[9,16–18] To date, only neoantigens derived from DNA mutations with the potential to activate the immune response have been used to generate personalized neoantigen vaccines in clinical trials for melanoma and glioma treatment.[19,20] Studies of neoantigens arising from aberrant RNA processing are limited and primarily based on short-read RNA-seq data that rely heavily on annotations that could compromise the precision of full ORF predictions. Conversely, the identification of full-length tumor-specific anomalous transcripts using LR-seq may enable predictions of complete novel ORFs to more accurately and exhaustively derive neoepitopes. Nevertheless, this concept and its feasibility have not been explored.

Here, we present a study integrating short-read and PacBio Iso-Seq-based long-read scRNA-seq of matched CRC clinical samples to build a full-length transcriptomic atlas and investigate isoform-specific dysregulations in CRC. We identified 394 DTSs from 273 genes in the tumor epithelial cells, many of which were caused by the coupling of multiple splicing events. Additionally, we detected tumor-cell-associated and isoform-specific RNA-editing events. By classifying normal (EpiN) and tumor epithelial (EpiT) cells into subpopulations, we unveiled genes and transcripts associated with three EpiN lineages as well as three EpiT subpopulations comprising more than 90% of malignant cells, each linked to a distinct prognosis. Among the 31,935 isoforms with novel splice junctions, 17,990 were supported by The Cancer Genome Atlas (TCGA) RNA-seq data, and 330 were additionally supported by mass spectrometry (MS) data. We experimentally validated four novel isoforms specifically expressed in tumor cells and revealed that they can potentially generate neoepitopes with strong binding affinities to major histocompatibility complex (MHC) molecules for a wide range of CRC patients. Finally, we developed an algorithm to build a panel of 22 neoepitopes with strong binding affinities to them, which may be exploited for the development of universal neoantigen-based cancer vaccines.

RESULTS

Matched long- and short-read scRNA-seq atlas of human CRC

To study the landscape of full-length transcript isoforms in human CRC, we isolated single cells from eight tumor and 11 adjacent normal samples derived from 12 CRC patients, generated cDNA libraries using the 10× Genomics droplet-based method, and performed matched Illumina short-read sequencing and PacBio Iso-Seq (Figures 1A and S1A; Table S1). We curated a total of 18,966 high-quality cells with Illumina scRNA-seq data after applying conventional filtering (STAR Methods and Figure S1B), with the number of cells per sample being slightly lower but comparable to related studies (Figure S1C).[5,6] To accurately identify the cell types, considering the lower cell number in our dataset we did not perform clustering from scratch. Instead, we mapped these cells to the Human Colon Cancer Atlas (c295), which is currently the most comprehensive single-cell atlas of CRC.[6] By integrating these two datasets (Figure S1A and STAR Methods), the cells were grouped into 20 major clusters (Figure 1B). We found that cells from the in-house tumor and normal samples were embedded into the c295 cell clusters without obvious batch effects (Figure 1C). Sixteen out of 20 cell types from c295 were detected in the in-house dataset, of which eight consisted of more than 100 cells (Table S1C). Among these, most of the cells were epithelial cells, followed by the immune and stromal cells (Figure 1D), and the gene expression of representative cell markers were validated in each cell type (Figures S1D–S1F).

Interestingly, compared to the normal samples, the tumor samples contained a higher number and proportion of immune cells, suggesting higher immunological activity in the tumor microenvironment (Figure 1E and Table S1D). To ensure robust identification of malignant epithelial cells, we inferred somatic copy-number variations (SCNVs) of the cells from the normal and tumor samples using inferCNV.[21] In general, the normal samples did not show significant SCNVs across the chromosomes, whereas the tumor samples displayed SCNVs in multiple regions (Figure 1F). These regions were generally consistent within each sample and located within previously characterized cytobands, such as amplifications in 1q, 7p, 8q, 19q, 20p, and 20q, as well as deletions in 4q, 15q, and 22q.[22,23] Using the inferred SCNVs and gene expression profiles (STAR Methods), we obtained a collection of 3,943 and 11,388 reliable tumor (EpiT) and normal (EpiN) epithelial cells, respectively.

In addition to short-read scRNA-seq, we also performed long-read scRNA-seq on three normal and four tumor samples using PacBio Iso-Seq (Figure 1A). Among the derived 19,264,405 circular consensus long-read sequencing (CCS) reads, 11,481,904 (59.6%) full-length non-concatemer (FLNC)

(B and C) tSNE (t-distributed stochastic neighbor embedding) plots illustrating (B) the major cell types and (C) the source of cells from in-house tumor and normal samples, and the public dataset (c295).
(D) Number of detected cells for each major cell type by in-house Illumina scRNA-seq data.
(E) Proportion of detected cells from each major cell type in each sample.
(F) Copy-number variation (CNV) profiling of each cell from each sample.

reads were obtained (STAR Methods). We identified 125,205 unique transcript isoforms from 17,753 genes, of which 23,379 were detected in both the normal and tumor samples. Based on the comparison of the splice junctions (SJs) of each aligned isoform with the reference transcriptome (GENCODE v37 and NCBI RefSeq release 109), we classified each transcript into four main categories—full splice match (FSM), incomplete splice match (ISM), novel in catalog (NIC), and novel not in catalog (NNC)—as well as several other non-specific categories (e.g., antisense, genic, intergenic). FSMs and ISMs had fully or partially matched SJs against the reference, whereas the NIC and NNC transcripts contain novel SJs from a combination of known and novel splice sites, respectively (Figure 2A). In total, we identified 31,837 (25.43%) FSM, 58,570 (46.78%) ISM, 13,291 (10.62%) NIC, 18,644 (14.89%) NNC, and 2,863 (2.29%) isoforms in the other categories (Figures S2A and S2B).

With long-read scRNA-seq, we were able to perform the quantification of RNA expression at the isoform level in each cell. Overall, we matched 1,994 out of 2,197 cell barcodes by comparing those identified in long-read and short-read scRNA-seq (Figure S2C). These cells were from 12 cell types, of which EpiN and EpiT are the most abundant, while immune and stromal cells constituted only a small fraction of the cell population (Figure S3A). Across all cell types, most of the isoforms were either FSM or ISM (Figure 2B). Interestingly, we found that EpiT expressed a slightly higher proportion of novel isoforms (NIC and NNC), suggesting more transcriptomic abnormalities in tumor cells.

We further evaluated the quality of the detected isoforms by validating their structural components. The 5′ end of >60% and 3′ end of >70% of the isoforms overlapped with annotated transcription start sites (TSSs) and termination sites (TTSs) in each cell type (Figure 2C and STAR Methods). All the SJs from >99% of FSM and ISM isoforms were also detected in colon adenocarcinoma patients from TCGA (TCGA-COAD) (Figure 2D). ~5% and >30% of NIC and NNC isoforms, respectively, contained at least one SJ that has not been identified in the TCGA-COAD RNA-seq data, suggesting that some novel splice junctions could be more effectively captured by LR-seq. To evaluate the concordance between the long- and short-read scRNA-seq data, we compared and observed a high overlap of the expressed genes identified in both datasets (Figure S3B). Furthermore, we observed a significant correlation of gene expression quantified by these two methods in both EpiN and EpiT ($r = 0.77, p < 2.2\times10^{-3}$ and $r = 0.67, p < 2.2 \times 10^{-3}$, respectively) (Figure S3C).

Among the 11,138 genes detected by long-read scRNA-seq, we found that 63% expressed multiple isoforms, with 13% having more than six isoforms (Figure 2E). For example, we detected 53 isoforms for the epithelial cell surface marker gene *EPCAM*. In addition to the 16 known isoforms (FSM and ISM), we identified 12 NIC and 25 NNC isoforms with the latter arising from 11 to 17 novel 5′ and 3′ splice sites (Figure 2F). These results indicate a high transcriptomic complexity in CRC tumors, which has been overlooked in previous studies.

Dysregulation of transcript structure and RNA editing in tumor epithelial cells

To systematically investigate the dysregulation of RNA isoforms in CRC, we defined two types of dysregulations by comparing gene and isoform expression between EpiN and EpiT (Figure 3A). Dysregulated gene expression (DGE) refers to genes that have significantly different numbers of overall transcripts between the two cell types, while DTS is defined as a multi-isoform gene that expresses significantly different proportions of its isoforms between EpiN and EpiT (STAR Methods). The fold changes of the genes showing significant DGE as measured by the two sequencing methods are strongly correlated ($r = 0.77$), with only three genes (<1%) showing discordance (Figure 3B). This high concordance indicates the reliability of gene expression quantification using PacBio long-read sequencing data. We further experimentally validated the top three DGE events, demonstrating the exclusive expression of *MMP7* and *REG3A* in CRC tumor samples, while *REG1A* was significantly more highly expressed in the tumor samples compared to the adjacent normal (Figures 3C and S4A). In total, we identified 898 DGE events and 273 genes with DTS from the PacBio data (Figure 3D). We observed more DGEs that were upregulated (747) than downregulated (151), suggesting high transcriptional activity in tumor cells (Figure 3D). Interestingly, we found that a higher frequency of genes with DTS had DGE, while single-isoform genes tend to be less dysregulated (Figure 3D). Gene ontology (GO) enrichment analysis of the genes with DTSs showed enrichment of pathways related to RNA regulation, such as mRNA binding, translational initiation, and RNA catabolic process (Figure 3E). This suggests potential autoregulation of these genes, which is frequently observed for RNA-binding proteins, particularly splicing factors.[24–26] Several immunological pathways were also implicated, indicating that DTS may contribute to dysregulated immunological activity in tumor cells. For example, the proliferating cell nuclear antigen gene, *PCNA*, presented both DGE and DTS. As a cell proliferation marker,[27] its expression was upregulated in EpiT and detected in a higher proportion of cells (Figure 3F). The EpiN and EpiT cells preferentially utilized distinct alternative first exons to differentially express the three *PCNA* isoforms. Overall, isoform proportion changes accounted only for a small fraction of transcripts with differential absolute expression levels between EpiT and EpiN for all four isoform categories, suggesting that transcriptional activity may have a more profound impact on isoform expression (Figure S4B).

We further applied SUPPA2[28] to extract AS events for the identified transcript isoforms, which were categorized into alternative 3′ splice site (A3SS), alternative 5′ splice site (A5SS), alternative first exon (AF), alternative last exon (AL), retained intron (RI), skipped exon (SE), and mutually exclusive exon (MX) (Figure S4C and STAR Methods). Consistent with previous observations,[13,14] the most abundant AS events were from the AF, RI, and SE categories (Figure S4D). We inspected how each isoform with DTS could arise from the different AS categories and found that isoforms derived from AF, A3, and RI were more enriched with DTS (Figure 3G, diagonal). More importantly, our long-read scRNA-seq captured the structure of full-length isoforms, allowing us to investigate the coupling between different AS categories. We observed that the isoforms derived from coupled AF-RI, AF-A3, and A3-RI had the highest enrichment of DTS isoforms (Figure 3G, off-diagonal). For example, among the *ARGHDIA* isoforms, those utilizing the distal first exon can retain any one of the three downstream introns (Figure 3H), while the

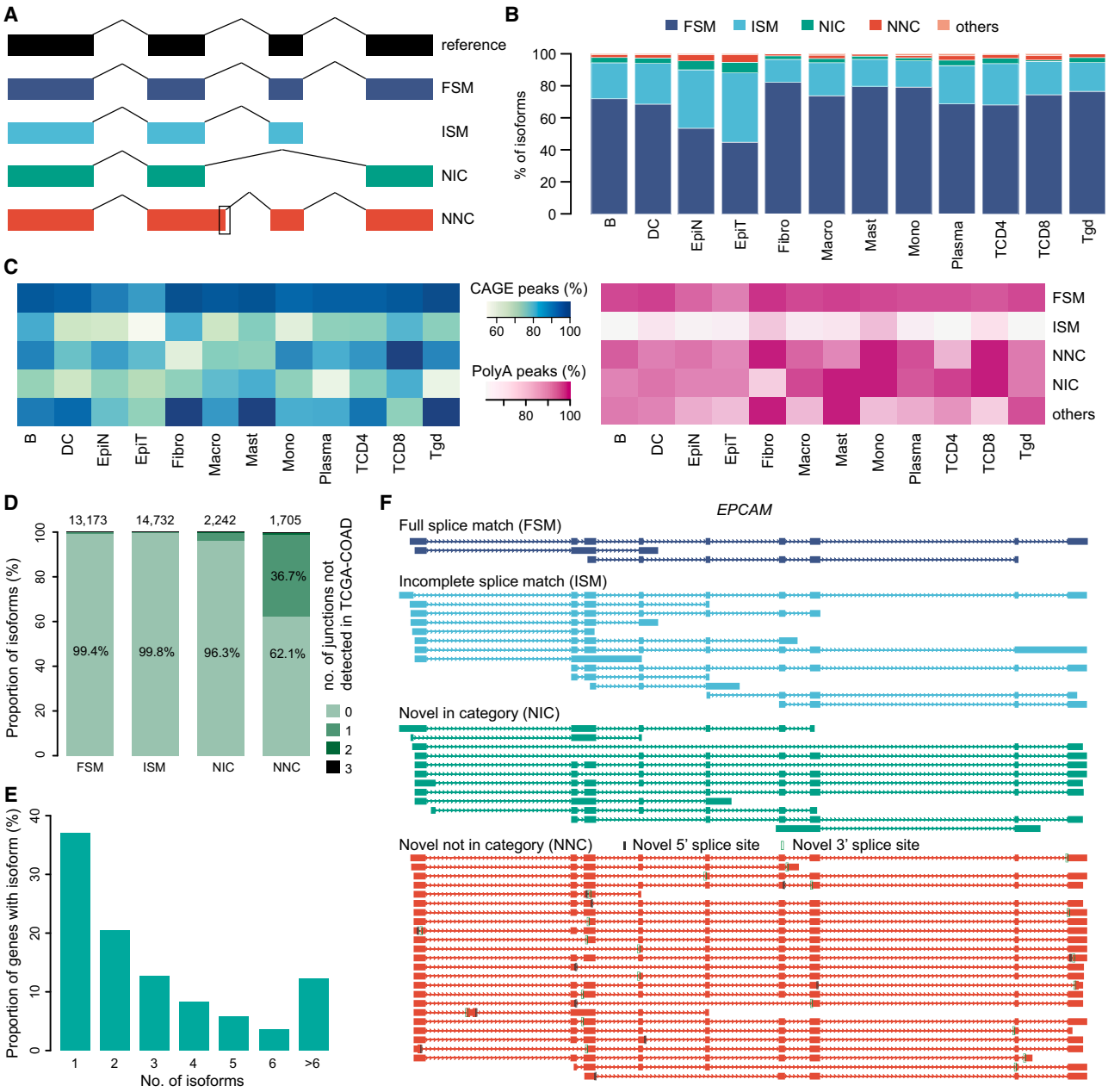


Figure 2. The long-read single-cell transcriptomic atlas of human CRC

(A) Schematic showing four structural types of isoforms identified by PacBio scRNA-seq data. FSM, full splice match; ISM, incomplete splice match; NIC, novel in catalog; NNC, novel not in catalog. The black box denotes a novel splice site.

(B) Proportion of the identified transcripts from each isoform structural type in each major cell type. Others include antisense, genic (from intronic regions), and intergenic transcripts.

(C) Percentage of the identified isoforms with supporting Cap Analysis of Gene Expression (CAGE) peaks for the 5′ ends and polyA peaks for the 3′ ends.

(D) Percentage of the identified isoforms with different numbers (0, 1, 2, 3) of splice junctions that are not detected in the TCGA-COAD bulk RNA-seq data.

(E) Percentage of genes with single and multiple detected isoforms.

(F) Structure of the four types of isoforms from *EPCAM* identified by PacBio scRNA-seq data. The identified novel splice sites are highlighted with black boxes.

TMEM259 isoforms using a proximal downstream 3′ splice site can retain an upstream intron (Figure 3I). Our findings underline the unique advantage of LR-seq in capturing isoform complexity that has thus far eluded the short-read sequencing technology.

In addition to AS, post-transcriptional RNA modifications on single nucleotides could also impact the mRNA sequence and expression without altering transcript structures. Specifically, double-stranded RNA-specific adenosine deaminase

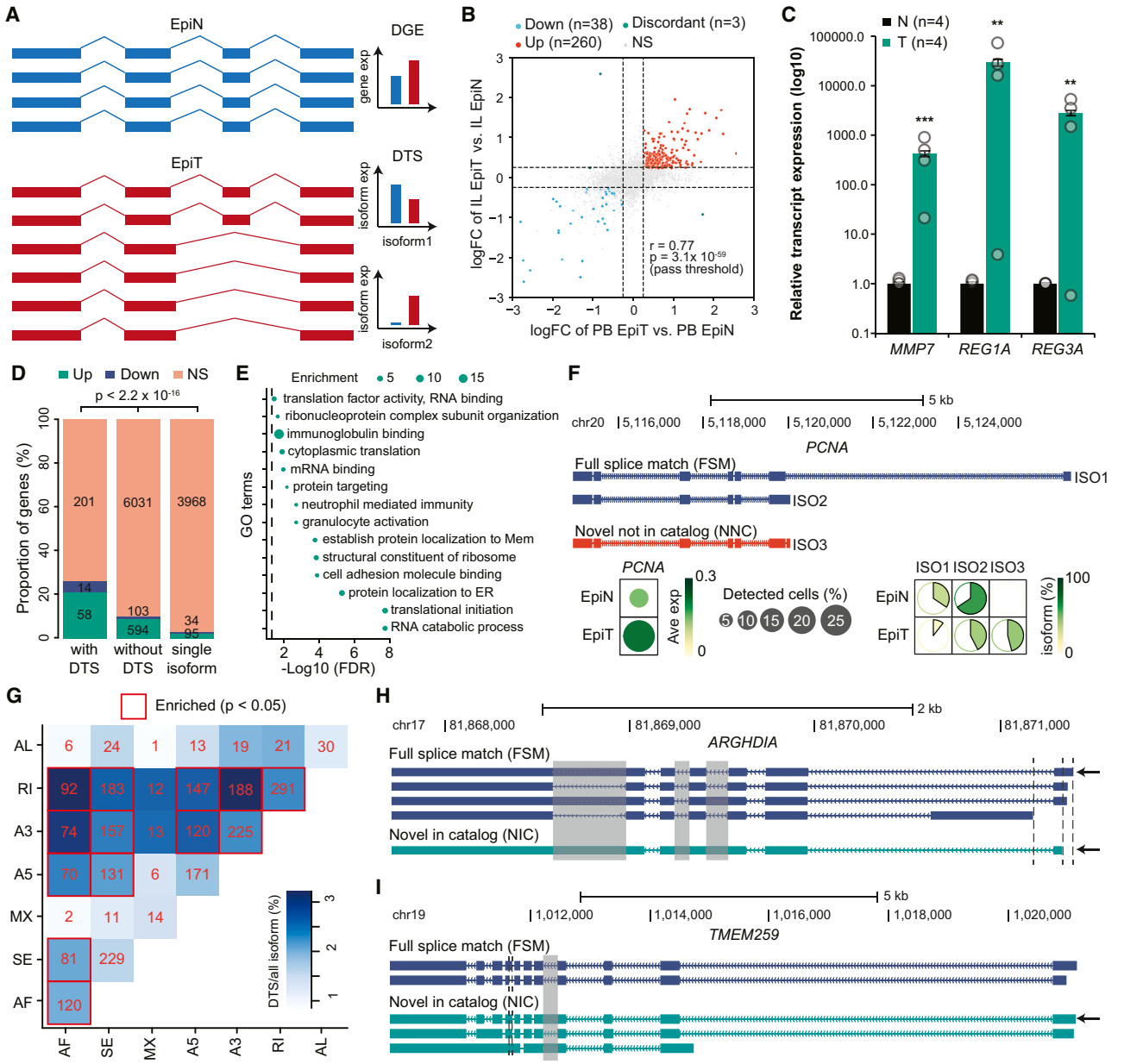


Figure 3. Dysregulated transcript structures in epithelial tumor cells

(A) Illustration of dysregulated gene expression (DGE) and dysregulated transcript structure (DTS) in epithelial tumor (EpiT) compared to normal (EpiN) cells.

(B) Scatterplot showing the correlation in fold change of genes between EpiT and EpiN quantified by long- and short-read sequencing. Up (Down), measured by both sequencing methods as significantly up- or downregulated; discordant, measured by both sequencing methods as significant but with inconsistent directions of change; NS, not significant.

(C) RT-qPCR validation of the top three DGE events, *MMP7*, *REG1A*, and *REG3A*, from (B) using four pairs of CRC tumor (T) and adjacent normal (N) patient samples (PS). *p* values from the Student's t test. $^{**}p < 0.01$, $^{***}p < 0.001$.

(D) Proportion of upregulated (Up), downregulated (Down), and not significantly changed (NS) genes for those with DTS, without DTS, and those with only one detected isoform. *p* value from chi-squared test.

(E) Enrichment of gene ontology (GO) terms for genes with DTS.

(F) Structures of the three identified isoforms from *PCNA* (upper), the gene expression pattern, and the percentage of each isoform in EpiN and EpiT.

(G) Numbers of DTS isoforms with co-occurrence of two types of AS events (or single type of events on the diagonal) and their percentage of the total corresponding isoforms. *p* value from two-tailed binomial test. AF, alternative first exon; SE, skipped exon; MX, mutually exclusive exon; A5, alternative 5′ splice site; A3, alternative 3′ splice site; RI, retained intron; AL, alternative last exon.

(H and I) Structural illustration of isoforms (arrows indicate DTS isoforms) from (H) *ARGHDIA* and (I) *TMEM259* with co-occurrence of two different types of splicing events. *ARGHDIA* DTS contains coupled AF and RI, while DTS from *TMEM259* contains coupled RI and A3.

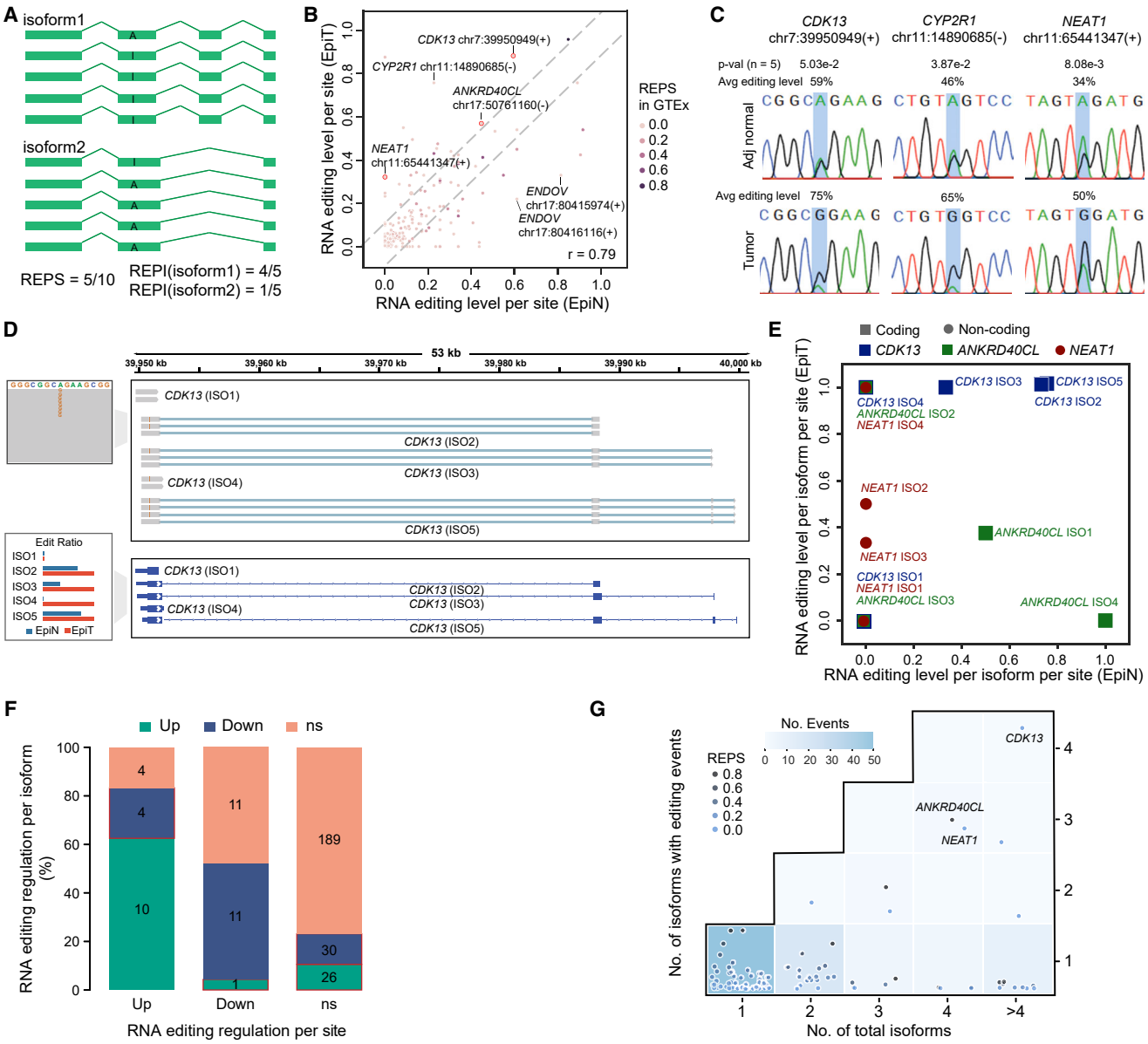


Figure 4. Dysregulated RNA editing in epithelial tumor cells

(A) Illustration of two approaches to calculate RNA-editing levels: RNA-editing level per site (REPS) and RNA-editing level per isoform (REPI).

(B) Scatterplot showing REPS for each detected event in epithelial tumor (EpiT) and normal (EpiN) cells. REPS based on RNA-seq data from GTEx are illustrated by scaled colors.

(C) Sanger sequencing validation of the RNA-editing sites and levels as shown in (B). Representative chromatograms are shown.

(D) Illustration of the detected RNA-editing events on each isoform from *CDK13* (upper) and the corresponding REPI in EpiN and EpiT.

(E) REPI for each event on each isoform from the sites highlighted in (B) in *CDK13*, *ANKRD40CL*, and *NEAT1*.

(F) Consistency between REPS and REPI in EpiT compared to EpiN. Red boxes indicate inconsistencies between REPS and REPI.

(G) Heatmap showing the number of isoforms with detected RNA editing versus the total number of detected isoforms of the gene. Each dot represents an editing event at a gene locus. Dots are colored according to their REPS, and those for the three RNA-editing events in (E) are highlighted.

(ADAR)-mediated A-to-I RNA editing can modify both coding and non-coding regions of mRNAs, as well as long non-coding RNAs, and has been implicated in several autoimmune disorders[29] and multiple cancers.[30]

We systematically identified the RNA-editing events from isoforms sequenced using long-read scRNA-seq by counting the number of edited reads (STAR Methods). To characterize the RNA-editing level of a site, we computed RNA-editing level per site (REPS) and per isoform (REPI), which are the ratio of the number of all edited reads against all reads for all isoforms and against all reads for a specific isoform, respectively (Figure 4A and Table S2). Of note, REPI calculation is only possible with

LR-seq, as it allows unambiguous association of each edited mRNA read with the isoform of origin.

A total of 196 RNA-editing sites with sufficient sequencing read support were identified in EpiN and EpiT cells. Most of these sites were on non-coding transcripts or the UTRs of coding transcripts, with more events in the 3′ UTRs than in 5′ UTRs (Figure S5A), which is consistent with previous studies.[31–33] Furthermore, ten sites were found in the coding regions, which could potentially introduce codon alterations. For example, the tumor-associated event at chr7:39950949(+) that results in the amino acid change Q103R in CDK13 (Figure 4B) was also observed in hepatocellular carcinoma.[34] Based on the REPS, RNA-editing events in genes such as *CDK13*, *CYP2R1*, and *NEAT1* were significantly upregulated in EpiT compared to EpiN (Figure 4B). We validated these events using normal colon CCD 841 CoN cells, and CRC cell lines DLD-1 and HCT116, as well as patient samples. The RNA-editing levels for *CDK13*, *CYP2R1*, and *NEAT1* were variable and did not follow any trend in the cell lines (Figure S5B). However, these were consistently higher in the patient tumor samples of at least four out of five matched pairs tested (Figures 4C and S5C), highlighting the importance of performing transcriptomic and validation studies using clinical samples.

To investigate isoform-specific RNA editing, we computed and compared REPI between EpiN and EpiT. For example, we first extracted the PacBio long reads belonging to each *CDK13* isoform and calculated the editing level of the site on each isoform separately using the corresponding reads (Figure 4D). We observed heterogeneity of RNA-editing levels of the same site among different isoforms (Figures 4E and S5D). Four of the *CDK13* isoforms showed different editing levels in EpiN cells but consistently high editing levels in EpiT, whereas all four isoforms of *NEAT1* showed low editing levels in EpiN with varying editing levels in EpiT. Overall, the direction of RNA-editing changes between EpiT and EpiN was largely concordant using REPS and REPI (Figure 4F). However, due to lower read coverage, we only observed a maximum of four edited isoforms per gene for a specific editing site while the per-isoform RNA-editing events were only observed on one isoform for most genes (Figure 4G). Taken together, our results suggest isoform specificity of the RNA-editing mechanism, which is still poorly understood.

Genes and isoforms associated with normal epithelial cells from different lineages

As epithelial cells constituted the majority of the identified cells in the scRNA-seq profiles, we performed cell subtyping for them using the c295 dataset as a reference to investigate their subpopulations[35] (STAR Methods). We identified 11 cell subtypes containing three differentiation lineages from the EpiN cells, including the newly characterized *BEST4*$^+$ lineage. The projection of the EpiN cells displayed a strong semantic structure in the two-dimensional t-distributed stochastic neighbor embedding (t-SNE) space, suggesting diverse differentiation lineages of the subtypes (Figure 5A). An increased level of differentiation was observed from the center to the edge in the tSNE space. For example, along the goblet lineage, cE02, cE06, and cE08 successively showed increased goblet maturity. Different combinations of subtypes were detected in each normal sample, and the proportions were similar across multiple samples (Figure 5B).

Using monocle 3,[36] we inferred diffusion pseudotime along the differentiation trajectory of the three epithelial lineages (STAR Methods) and identified 46, 31, and 38 lineage-specific genes (Moran's $I > 0.5$, adjusted p value $<1.0 \times 10^{-5}$) for the enterocyte, goblet, and *BEST4* lineages, respectively, including both well-known and novel markers (Figures 5C–5E). These showed a steady increase in expression along the lineage with respect to pseudotime (Figures 5F–5H). We next performed differential gene and isoform expression analysis between the stem/transient amplifying (TA) and differentiated subtypes along the three lineages. Overall, we observed a similar number of up- and downregulated genes in the stem/TA subtypes but mostly upregulation of lineage-specific genes in the three differentiated lineages (Figures S6A and S6B; Table S3). The genes with differential regulation were almost mutually exclusive between the three differentiated subtypes (Figure S6C), indicating distinct transcriptomic regulatory mechanisms to commit to a specific differentiation lineage.

By categorizing the genes in each differential gene and isoform analysis with DGE and DTS (defined in Figure 3A), we observed a higher percentage of genes with DTSs subjected to DGE compared to those without DTSs or with only one isoform (Figure 5I), similar to the trend observed between EpiN and EpiT (Figure 3D). The gene *LMNA* encoding two protein isoforms, Prelamin-A and Prelamin-C, through AS and alternative polyadenylation (APA), serves as a representative of differential isoform usage between cell subtypes (Figure 5J). We observed a trend of Prelamin-C to Prelamin-A switching when stem cells differentiated into enterocytes, goblet cells, and *BEST4*$^+$ cells (Figure 5K), consistent with the findings in basal cell carcinomas, where Prelamin-A was negatively correlated with cell proliferation and appeared in the later stage of differentiation.[37] Overall, our analysis provides the first comprehensive characterization of isoforms associated with three colon epithelial lineages.

Tumor epithelial subtypes display different levels of stemness and differentiation

We subtyped the EpiT cells with the same classifier described above to classify the EpiT cells to their closest EpiN subtypes (Figure 6A). Consistent with previous reports,[6,7] more than 93% of tumor epithelial cells were categorized as the three stem/TA subtypes (Figure 6B), suggesting that malignant EpiT cells may possess stem/TA characteristics to maintain their proliferation and regeneration capabilities. Based on the integrative analysis of multiple CRC single-cell sequencing datasets, a recent study classified colon EpiT cells into two intrinsic consensus molecular subtypes (iCMSs), iCMS2 and iCMS3,[8] and revealed that over 90% of EpiT cells are from either subtype. This holds true in ours and the c295 datasets, with mostly one type of iCMS cells in each tumor (Figures S7A and S7B). However, the distribution of the three subtypes from our categorization was more heterogeneous in each tumor (Figures 6B and S7B). These distinct results suggest potentially diverse and independent molecular properties of the two classification methods.

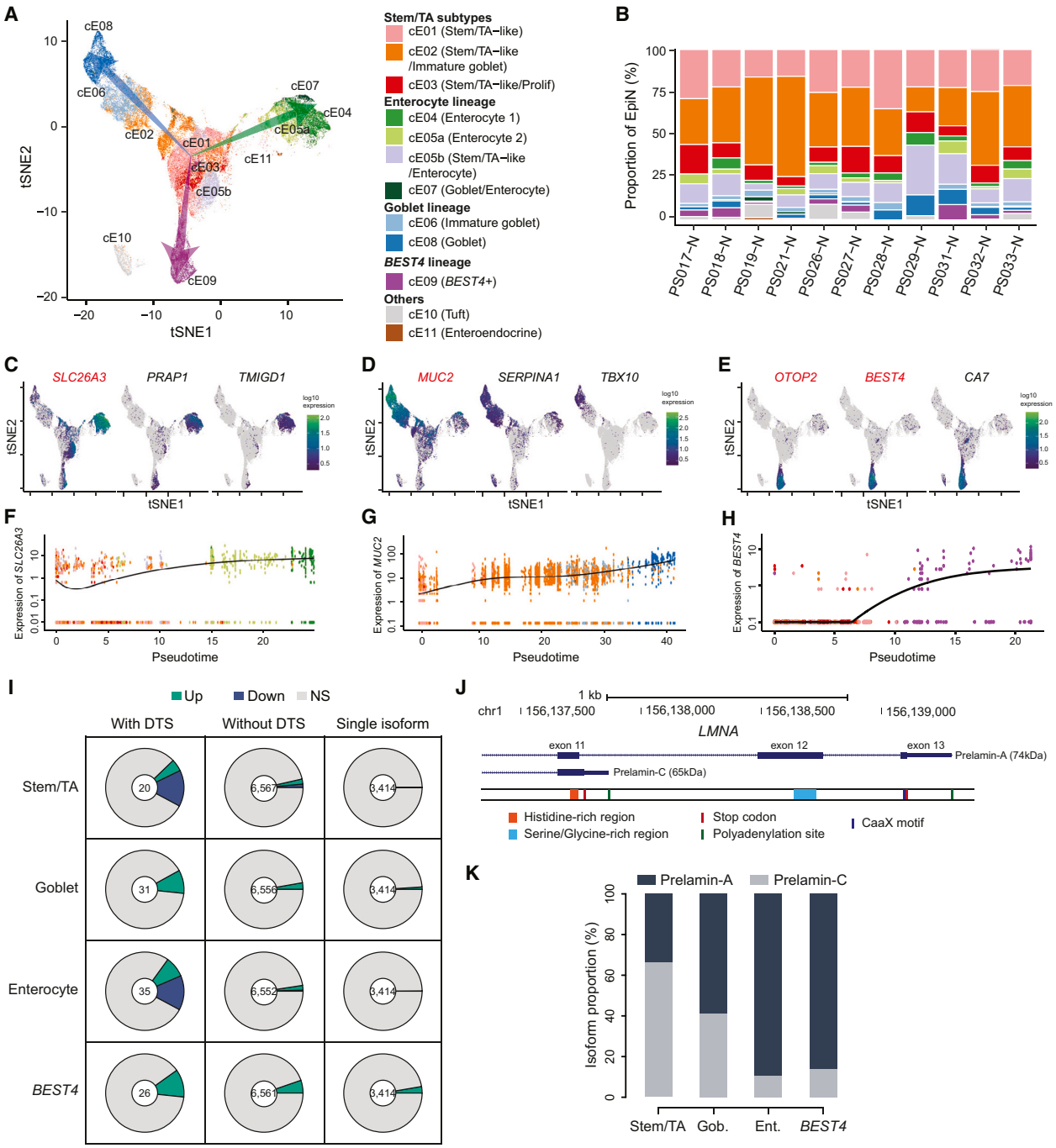


Figure 5. Transcriptome profiling of normal epithelial cell subtypes from multiple lineages

(A) tSNE plot illustrating subtypes of epithelial normal cells (EpiN) and the three main lineages of differentiation (indicated by arrows), including enterocyte (green), goblet (blue), and *BEST4* (purple).

(B) Proportion of EpiN subtypes in each sample.

(C–E) The top identified markers with lineage-specific expression for (C) enterocyte, (D) goblet, and (E) *BEST4*. Known markers are in red.

(F–H) RNA expression of lineage-specific marker genes over the pseudotime was estimated by trajectory analysis along the (F) enterocyte, (G) goblet, and (H) *BEST4* lineages.

(I) Proportion of upregulated (Up), downregulated (Down), and not significantly changed (NS) genes for genes with DTS, without DTS, and those with only one detected isoform in the stem/TA subtypes and each differentiated lineage.

(J) Structure of the *LMNA* transcripts encoding two Prelamin protein isoforms, Prelamin-A and Prelamin-C.

(K) Proportion of the *LMNA* transcripts encoding the two Prelamin protein isoforms in the stem/TA subtypes and the three differentiation lineages.

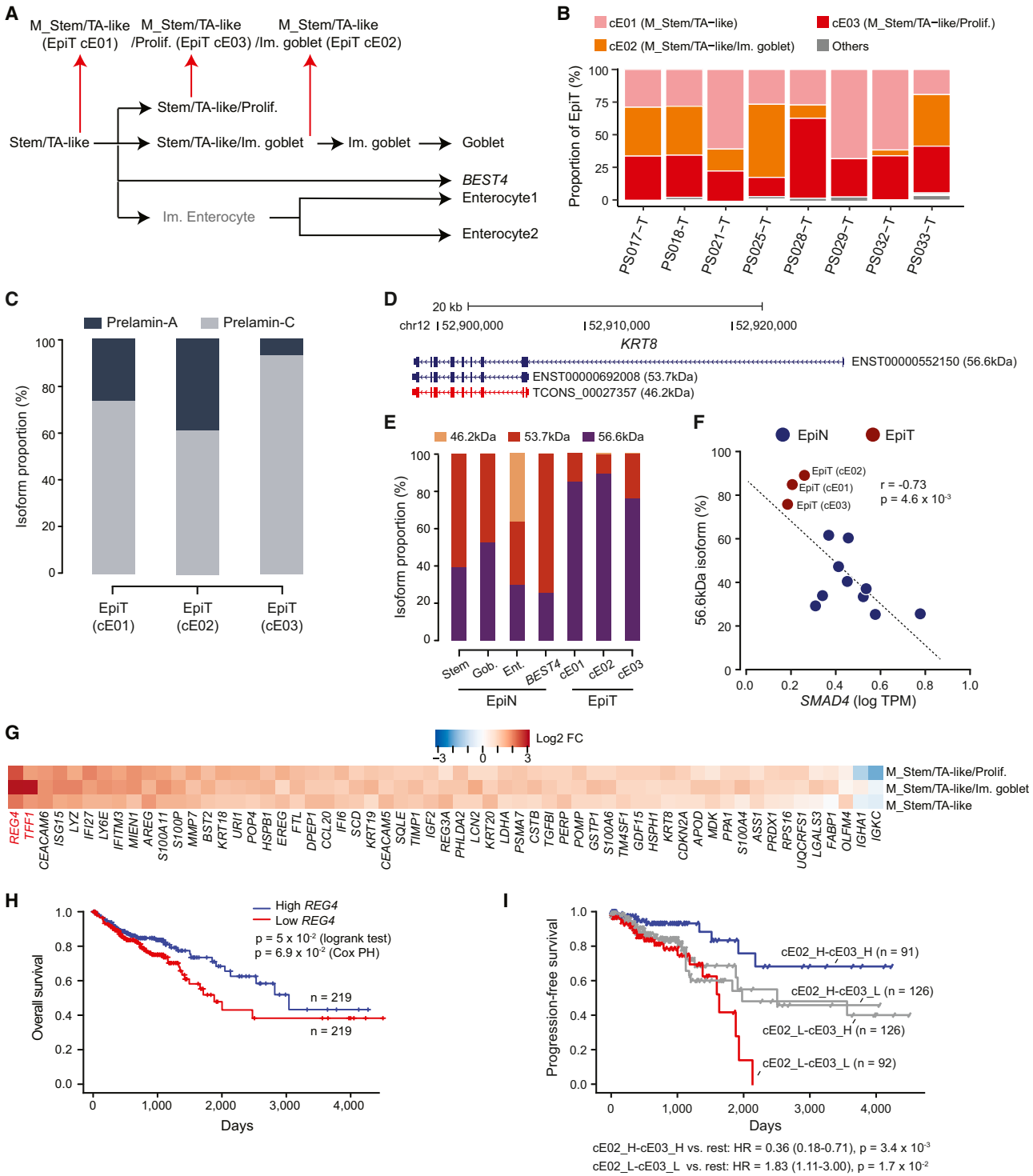


Figure 6. Dysregulation of genes and isoforms in the epithelial tumor cell subtypes

(A) Schematic showing the comparisons between each EpiT subtype and their corresponding EpiN cell subtype.

(B) Proportion of EpiT subtypes in each sample.

(C) Proportion of the *LMNA* transcripts encoding the two Prelamin protein isoforms in each EpiT subtype.

(D) Structure of the *KRT8* transcripts encoding three CK8 protein isoforms with molecular weights of 56.6, 53.7, and 46.2 kDa.

(legend continued on next page)

We further identified dysregulated genes in the EpiT subtypes by comparing them to the corresponding EpiN subtypes (Figure S7C and Table S4). Most of these were shared among the three EpiT subtypes (Figure S7D), indicating a common underlying dysregulated mechanism of tumorigenesis. In addition to DGE, we performed differential isoform analysis for the EpiT subpopulations (Table S5). Differential usage of the *LMNA* isoforms was also observed in the EpiT subtypes, with the highest usage of the Prelamin-C isoform in the most proliferative subtype, cE03, and Prelamin-A isoform in cE02 (Figure 6C). Furthermore, we noticed that several cytokeratin genes, including *KRT8* and *KRT19*, were among the most significant DTS in all three EpiT subtypes (Table S5). The overexpression of *KRT8* has been linked to increased tumor progression and invasiveness in several epithelial tumors.[38–40] We detected three major *KRT8* isoforms, including two annotated and one novel isoform (Figure 6D). The tumor subtypes mainly expressed the full-length isoform encoding a 56.6-kDa protein, while the normal subtypes primarily utilized the short isoforms (Figure 6E). Consistent with a previous study showing that the tumor suppressor *SMAD4* regulated *KRT8* splicing, [41] we observed a negative correlation between *SMAD4* expression and the 56.6-kDa isoform usage among the epithelial subtypes (Figure 6F), further confirming the reliability of our isoform analysis. Overall, around 20 genes with DTS were identified in each EpiT subtype and, consistent with those in the EpiN cells, genes with DTS were more likely to be DGE (Figure S8A). Contrary to the DGEs, the genes with DTS barely overlapped among the three subtypes (Figure S8B), suggesting potential dysregulation of isoform switching in the EpiT subpopulations.

Among the top dysregulated genes in the EpiT subtypes, *REG4* and *TFF1* showed the most striking upregulation in cE02 (Figure 6G). Interestingly, we found that *REG4* and *TFF1* expression was highly correlated across the TCGA-COAD patients ($r = 0.75$, $p = 3.81 \times 10^{-71}$), and both were associated with better overall survival (OS) (Figures 6H and S8C). We speculate that the tumor samples with high *REG4* and *TFF1* expression may consist of a high proportion of cE02 cells, leading to better patient survival. To this end, we identified the marker genes in each subtype to define a signature score for the three subtypes (STAR Methods and Table S5). By classifying TCGA tumors with cE02 and cE03 scores (Figure S8D), we found that the patient groups with both high cE02 and cE03 scores and low cE02 and cE03 scores had the best and worst OS and progression-free survival (PFS), respectively (Figures 6I and S8E), suggesting a synergistic effect of cE02 and cE03 scores on patient prognosis. Similar trends were observed using an independent microarray cohort curated by the CRC Subtyping Consortium (CRCSC)[42] (Figures S8F and S8G). As cE02 and cE03 have relatively lower stemness features and higher proliferative activity compared to cE01, these results indicate that CRC tumors enriched with high stemness and low replication of cells are associated with more frequent progression and shorter OS time.

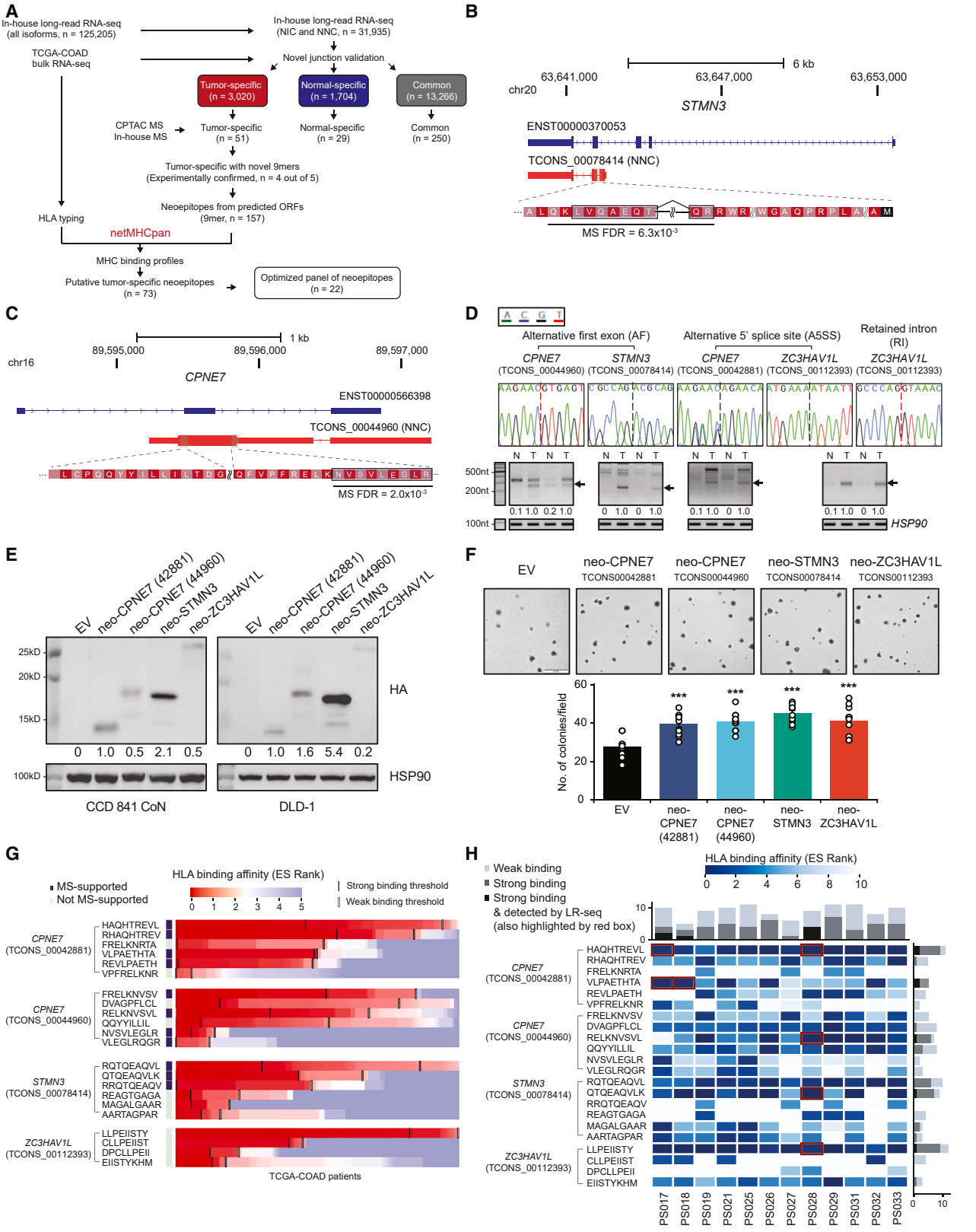
Identification of neoepitopes from recurrent tumor-specific isoforms for cancer vaccine development

Among the 125,205 identified isoforms, 31,935 are NIC and NNC isoforms that contain novel SJs (Figure 7A). Considering the small sample size and expression sparsity of our single-cell sequencing data, we further verified and requantified the expression of these isoforms in tumor and normal samples in a large CRC cohort, utilizing the reads supporting novel junctions from TCGA RNA-seq data (Figure 7A). 3,020 and 1,704 isoforms exhibit high expression specificity in tumor and normal samples, respectively, while 13,226 are expressed in both sample types (Figure 7A). We then predicted the ORFs from these isoforms. By integrating publicly available MS data[43] from TCGA-COAD patients and in-house MS data from matched tumor and normal CRC patient samples and HCT116 cell line, we found peptides supporting 194 and 136 novel proteins from the NIC and NNC isoforms, respectively, among which 51 isoforms are tumor specific (Figure 7A). Examples of the two novel isoforms from the genes *STMN3* and *CNPE7* that are supported by MS data are illustrated in Figures 7B and 7C.

We experimentally validated the tumor-specific expression and novel SJs of four recurrent isoforms in colon epithelial and CRC cell lines, as well as at least four of the five matched patient samples tested (Table S6 and Figures 7D, S9A, and S9B). Additionally, to assess the potential for stable protein synthesis from these RNA isoforms, we overexpressed the ORFs fused to a hemagglutinin (HA) tag in both colon epithelial and CRC cell lines, demonstrating that all four are capable of generating proteins *in vitro* (Figure 7E). Given their tumor-specific expression pattern, we speculate that these tumor-specific isoforms may possess oncogenic functions. In line with this, we found that the overexpression of their ORFs promotes cell growth *in vitro* (Figures 7F and S9C).

As abnormal proteins produced in cancer cells are critical for spontaneous anti-tumor immune response, they serve as a potential source for neoantigen-based cancer vaccine development to leverage the immune system for cancer treatment.[15] Therefore, we systematically investigated putative neoepitopes from these four tumor recurrent isoforms. Among the 157 novel 9-mers from these novel isoforms, 73 exhibit strong binding affinity for patient MHC, indicating their potential as neoepitopes (Table S7 and STAR Methods). To develop cancer vaccines for a broad range of patients, we considered the diversity of HLA alleles across different patients and predicted the MHC binding profiles of 458 TCGA-COAD patients to the neoepitopes from the 12 recurrent tumor-specific isoforms. We formulated a greedy algorithm and selected a total of 22 neoepitopes from the four isoforms to maximize their coverage to the HLA alleles of TCGA patients while prioritizing those that overlapped with

(E) Proportion of the *KRT8* transcripts encoding each CK8 protein isoform in each EpiN and EpiT subtype.
(F) Correlation between the usage of the 56.6-kDa transcript isoform and *SMAD4* expression in each EpiN and EpiT subtype.
(G) Fold changes ($\log_2$ transformed) of top significant genes with dysregulated expression in each EpiT subtype compared to the corresponding EpiN subtype.
(H) Overall survival of TCGA-COAD patients with different expression levels of *REG4*.
(I) Progression-free survival of TCGA-COAD patients with different scores of cE02 and cE03 signature genes. H, high; L, low.



(legend on next page)

MS-supported peptides (STAR Methods and Algorithm 1). Interestingly, unlike the neoepitopes derived from genomic mutations that are highly patient specific,[44] we found many of these predicted neoepitopes are shared across CRC patients. For example, 13 of the top 22 predicted neoepitopes have binding affinities with HLA molecules from a wide range (>50%) of CRC patients (Figure 7G and Table S7). Furthermore, we observed strong HLA binding potential for this panel of neoepitopes in the 12 in-house patients, with all patients having at least six strong-binding neoepitopes (Figure 7H and Table S8). This aligns with our previous findings that mutation-derived neoepitopes are usually patient specific, whereas neoepitopes from alternative transcripts generated from intronic polyadenylation and intron retention could be recurrent across patients.[45] Collectively, these findings demonstrate a proof of concept for the discovery of neoepitopes based on the tumor-specific transcriptome as well as the invaluable application potential of long-read scRNA-seq in the development of neoepitope-based cancer vaccines.

DISCUSSION

The intestinal epithelium is the fastest self-renewing tissue in mammals and contains epithelial cells with diverse differentiation statuses, proliferation activities, and other properties and functions. Despite various markers and gene expression profiles that have been characterized for each epithelial cell subtype, the diversity of RNA isoforms in these subtypes remains underappreciated. Our study provides a full-length transcriptome across different epithelial subtypes and identifies isoforms associated with enterocyte, goblet, and *BEST4* lineages, respectively. Although alternative RNA isoforms have been shown to regulate cell-fate determination of stem cells,[46] the prevalence and potential role of alternative isoforms in intestinal epithelial differentiation are poorly understood; therefore, our data may provide a valuable resource to facilitate further investigations.

CRC is the third most common cancer type and ranks as the second leading cause of cancer death around the world.[47] Consistent with the first high-throughput scRNA-seq study on CRC and other single-cell analyses that traced the transformation of polyps to CRC,[3,4] we find that the vast majority of CRC tumor cells have stem-like cell characteristics. However, studies for EpiT subtypes are limited. Based on the integrative analysis of multiple CRC single-cell sequencing datasets, a recent study classified colon EpiT cells into two subtypes and showed that 90% of EpiT in each tumor were from either subtypes.[8] They further found that these two subtypes were associated with different cancer driver mutations, suggesting that their classification method mainly captured heterogeneities across different tumors. However, our EpiT subtype classification demonstrated that each tumor comprised a substantial proportion of three cell subtypes with varying differentiation status and proliferation activities, thus capturing more intratumor heterogeneities. As our classification indicates that CRC tumors are differentially associated with PFS and OS depending on their subtype composition, the markers we propose for each subtype may serve as useful prognostic indicators. In addition to epithelial cells, other stromal and immune cells may also impact patient survival and response to treatment.[5,7] However, due to limited cell numbers, our study could not effectively capture the full-length transcriptome in these cells, highlighting the need for more advanced and robust cell-isolation technologies.

Despite recent reports of the full transcriptome by bulk RNA LR-seq in breast and gastric cancers,[13,14] single-cell LR-seq studies are rare and limited to cancer cell lines.[48,49] Moreover, previous LR-seq studies in cancer focused only on alterations in single AS events. Here, we identify hundreds of DTSs arising from combined AS events and calculate isoform- and cell-type-specific RNA-editing levels, all of which originate from primary CRC tissues and are only achievable by single-cell LR-seq. Nonetheless, our analyses are limited by technological constraints leading to the relatively low coverage of single-cell PacBio sequencing. Thus, the development of methods to increase the capture efficiency of single-cell isoform sequencing will be beneficial for future transcriptomic studies.[50]

Therapeutic neoantigen-based cancer vaccines aim to leverage the immune system for the treatment of cancer.[15] Due to somatic mutations and transcriptomic dysregulation, tumor cells produce abnormal proteins that are absent in normal cells. In the event that fragments of these proteins are transported to the cell surface via the MHC I pathway, they may potentially be recognized by the immune system as an antigen. We show that the full-length transcriptome in tumor cells with MS data from

Figure 7. Identification of common neoantigens for cancer vaccine from recurrent tumor-specific transcripts

(A) Workflow for the identification of neoepitopes from novel tumor-specific recurrent transcript isoforms for cancer vaccine development.

(B and C) Selection of neoepitopes for a novel isoform of (B) *STMN3* (TCONS_0078414) and (C) *CPNE7* (TCONS_00044960). The selected MS-supported neoepitopes are encircled by solid boxes, and the non-MS-supported ones by dashed-line boxes.

(D) PCR and Sanger sequencing validation of the unique splice junctions for the recurrent tumor-specific isoforms from which the neoepitope panel is derived. Novel splice junctions are depicted by black dotted lines and 5′ junction of sequences unique to selected isoforms by red dotted lines. Representative chromatograms and images are shown.

(E) Western blot validation of overexpressing the HA-tagged open reading frames (ORFs) derived from the validated neoepitopes in colon epithelial cell line, CCD 841 CoN, and CRC cell line, DLD-1.

(F) Effect of ORF overexpression from (E) on anchorage-independent growth in DLD-1 cells. p values from the Student's t test, $^{***}p < 0.001$.

(G) HLA binding profile of the panel of 22 neoepitopes against the HLA alleles of the TCGA-COAD patients (values of highest binding affinity of the patients' alleles are shown). Each row is sorted from the individual with the highest binding affinity to the lowest. The thresholds of strong binding affinity (ES rank <0.5) and weak binding affinity (ES rank <2) are marked accordingly.

(H) Binding affinity of the 22 neoepitopes for the 12 in-house patients. "Weak/strong binding" denotes the binding affinity of the neoepitopes to at least one patient HLA allele (each row). "Strong binding & detected by LR-seq" neoepitopes are indicated by red boxes in the corresponding heatmap. The top panel summarizes the total number of neoepitopes that satisfies each criterion above per patient. The right panel summarizes the total number of patients for whom the epitope satisfies the criteria above.

matched patient samples can provide a comprehensive source to detect aberrant mRNA-derived neoantigens. Moreover, we propose an algorithm that provides an optimized combination of recurrent tumor-specific neoepitopes with high HLA binding affinities for a wide spectrum of CRC patients. Recently, the phase II clinical trial for a neoantigen-based mRNA vaccine by Moderna has met the endpoint.[51] In contrast to the customized vaccine design based on a neoepitope identified by exome sequencing of a patient tumor, the panel we propose can potentially target larger CRC patient cohorts without sequencing individual tumors. This could facilitate the development of neoantigen-based cancer vaccines with less economic and time constraints to benefit a larger population.

Limitations of the study

This study presents several limitations that should be acknowledged. First, cell viability of the clinical colorectal tumor samples was not optimal (Figure S1B), which may have impacted the quality and inclusiveness of the transcriptomic data obtained. Second, the yield of PacBio long-read sequencing was suboptimal, especially in light of the number of cells detected (Figure S1C), which constrained the depth of analysis for each sample. Additionally, the number of samples analyzed using long-read sequencing was limited to three normal and four tumor samples (Figure 1A), a restriction primarily due to budget constraints and the low success rate of library construction. Future studies could potentially overcome these limitations by utilizing new technologies such as PacBio MAS-ISO-seq,[52] which offer better yield and are more cost effective, thereby enabling more comprehensive and reliable transcriptomic analyses.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yvonne Tay (yvonnetay@nus.edu.sg).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Raw sequencing data used in this study are deposited in European Nucleotide Archive under session number PRJEB68074. Processed long-read and short-read sequencing data are deposited at Zenodo (https://doi.org/10.5281/zenodo.12750017). The identified RNA transcript isoforms from LR-seq data are visualized and available for downloading from UCSC genome browser tracks with the link: https://genome.ucsc.edu/s/binzhang/COAD_Colored_PacBio. Code for data analysis and neoepitope selection is available at https://github.com/lzx325/CRC-atlas.git.

ACKNOWLEDGMENTS

We thank all past and present Y.T. and X.G. lab members for their constructive feedback on this project, Ng Desi for assisting with the single-cell isolation procedure, and Luke Esau and other members of the bioscience core laboratory at KAUST for performing the single-cell Illumina and PacBio sequencing. We additionally thank personnel from Iain Bee Huat Tan's lab for providing the iCMS classification on the c295 dataset. The results in the study are in part based upon data generated by the TCGA Research Network, https://www.cancer.gov/tcga. This study is supported by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiative. Y.T. is funded by NMRC OF-IRGs (NMRC/OFIRG/MOH-000380, MOH-000923). X.G. is supported by the King Abdullah University of Science and Technology (KAUST) Office of Research Administration (ORA) under Award No REI/1/5234-01-01, REI/1/5414-01-01, REI/1/5289-01-01, REI/1/5404-01-01, REI/1/5992-01-01, and URF/1/4663-01-01; Center of Excellence for Smart Health (KCSH) under award number 5932; and Center of Excellence on Generative AI under award number 5940.

AUTHOR CONTRIBUTIONS

B.Z., X.G., and Y.T. conceived the study. S.W., B.E.S., I.J.-W.T., K.-Y.L., B.L., W.-K.C., and K.-K.T. provided the CRC clinical samples. J.J.C., H.T., Q.Y.T., X.H.C., X.F., C.C., F.C., and D.K. constructed the Illumina short-read and PacBio long-read libraries. P.D. performed the Illumina and PacBio sequencing at the KAUST Bioscience Core Lab. Z.L. and B.Z. conducted all of the computational analyses and drafted the majority of the manuscript. J.J.C. did the experimental validation of the DEGs, DTSs, RNA-editing events, and neoantigen candidate isoforms and ORFs. Z.L., B.Z., and J.J.C. drafted the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - CRC clinical samples
- METHOD DETAILS
  - Tissue processing, library construction, and single-cell sequencing
  - Plasmids
  - Cell culture and transfection
  - Soft agar assay
  - RNA extraction, RT-qPCR, and PCR
  - Protein extraction and western blot analysis
  - Mass spectrometry
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Summary of bioinformatic analysis
  - Illumina short-read scRNA-seq data pre-processing
  - Cell type identification
  - Dimensionality reduction and visualization
  - PacBio long-read scRNA-seq data pre-processing
  - Isoform quality control, filtering, and classification
  - Splice event extraction and analysis
  - Identification of dysregulated gene expression and transcript structures
  - GO enrichment analysis
  - Calling RNA-editing events
  - Lineage and trajectory analysis
  - Estimation of the signature scores of EpiT subtypes in bulk samples
  - Survival analysis
  - Proteomic analysis of isoforms
  - Mass spectrometry data analysis
  - Proposing putative neoepitopes for cancer vaccine development
  - Command lines of software tools used in this study

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100641.

Received: June 21, 2023
Revised: June 6, 2024

Accepted: August 7, 2024
Published: August 30, 2024

REFERENCES

1. Burrell, R.A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. Nature *501*, 338–345. https://doi.org/10.1038/nature12625.
2. Meacham, C.E., and Morrison, S.J. (2013). Tumour heterogeneity and cancer cell plasticity. Nature *501*, 328–337. https://doi.org/10.1038/nature12624.
3. Li, H., Courtois, E.T., Sengupta, D., Tan, Y., Chen, K.H., Goh, J.J.L., Kong, S.L., Chua, C., Hon, L.K., Tan, W.S., et al. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. *49*, 708–718. https://doi.org/10.1038/ng.3818.
4. Becker, W.R., Nevins, S.A., Chen, D.C., Chiu, R., Horning, A.M., Guha, T.K., Laquindanum, R., Mills, M., Chaib, H., Ladabaum, U., et al. (2022). Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. Nat. Genet. *54*, 985–995. https://doi.org/10.1038/s41588-022-01088-x.
5. Zhang, L., Li, Z., Skrzypczynska, K.M., Fang, Q., Zhang, W., O'Brien, S.A., He, Y., Wang, L., Zhang, Q., Kim, A., et al. (2020). Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell *181*, 442–459.e29. https://doi.org/10.1016/j.cell.2020.03.048.
6. Pelka, K., Hofree, M., Chen, J.H., Sarkizova, S., Pirl, J.D., Jorgji, V., Bejnood, A., Dionne, D., Ge, W.H., Xu, K.H., et al. (2021). Spatially organized multicellular immune hubs in human colorectal cancer. Cell *184*, 4734–4752.e20. https://doi.org/10.1016/j.cell.2021.08.003.
7. Lee, H.-O., Hong, Y., Etlioglu, H.E., Cho, Y.B., Pomella, V., Van den Bosch, B., Vanhecke, J., Verbandt, S., Hong, H., Min, J.-W., et al. (2020). Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat. Genet. *52*, 594–603. https://doi.org/10.1038/s41588-020-0636-z.
8. Joanito, I., Wirapati, P., Zhao, N., Nawaz, Z., Yeo, G., Lee, F., Eng, C.L.P., Macalinao, D.C., Kahraman, M., Srinivasan, H., et al. (2022). Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. Nat. Genet. *54*, 963–975. https://doi.org/10.1038/s41588-022-01100-4.
9. Kahles, A., Lehmann, K.-V., Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., and Cancer Genome Atlas Research Network; and Rätsch, G. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell *34*, 211–224.e6. https://doi.org/10.1016/j.ccell.2018.07.001.
10. Xia, Z., Donehower, L.A., Cooper, T.A., Neilson, J.R., Wheeler, D.A., Wagner, E.J., and Li, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat. Commun. *5*, 5274. https://doi.org/10.1038/ncomms6274.
11. Demircioğlu, D., Cukuroglu, E., Kindermans, M., Nandi, T., Calabrese, C., Fonseca, N.A., Kahles, A., Lehmann, K.-V., Stegle, O., Brazma, A., et al. (2019). A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. Cell *178*, 1465–1477.e1417. https://doi.org/10.1016/j.cell.2019.08.018.
12. Chan, J.J., Zhang, B., Chew, X.H., Salhi, A., Kwok, Z.H., Lim, C.Y., Desi, N., Subramaniam, N., Siemens, A., Kinanti, T., et al. (2022). Pan-cancer pervasive upregulation of 3′ UTR splicing drives tumourigenesis. Nat. Cell Biol. *24*, 928–939. https://doi.org/10.1038/s41556-022-00913-z.
13. Veiga, D.F.T., Nesta, A., Zhao, Y., Deslattes Mays, A., Huynh, R., Rossi, R., Wu, T.-C., Palucka, K., Anczukow, O., Beck, C.R., and Banchereau, J. (2022). A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. Sci. Adv. *8*, eabg6711. https://doi.org/10.1126/sciadv.abg6711.
14. Huang, K.K., Huang, J., Wu, J.K.L., Lee, M., Tay, S.T., Kumar, V., Ramnarayanan, K., Padmanabhan, N., Xu, C., Tan, A.L.K., et al. (2021). Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. Genome Biol. *22*, 44.
15. Sellars, M.C., Wu, C.J., and Fritsch, E.F. (2022). Cancer vaccines: Building a bridge over troubled waters. Cell *185*, 2770–2788. https://doi.org/10.1016/j.cell.2022.06.035.
16. Smart, A.C., Margolis, C.A., Pimentel, H., He, M.X., Miao, D., Adeegbe, D., Fugmann, T., Wong, K.-K., and Van Allen, E.M. (2018). Intron retention is a source of neoepitopes in cancer. Nat. Biotechnol. *36*, 1056–1058. https://doi.org/10.1038/nbt.4239.
17. Wang, T.-Y., Liu, Q., Ren, Y., Alam, S.K., Wang, L., Zhu, Z., Hoeppner, L.H., Dehm, S.M., Cao, Q., and Yang, R. (2021). A pan-cancer transcriptome analysis of exitron splicing identifies novel cancer driver genes and neoepitopes. Mol. Cell *81*, 2246–2260.e12. https://doi.org/10.1016/j.molcel.2021.03.028.
18. Zhang, M., Fritsche, J., Roszik, J., Williams, L.J., Peng, X., Chiu, Y., Tsou, C.-C., Hoffgaard, F., Goldfinger, V., Schoor, O., et al. (2018). RNA editing derived epitopes function as cancer antigens to elicit immune responses. Nat. Commun. *9*, 3919. https://doi.org/10.1038/s41467-018-06405-9.
19. Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. Nature *547*, 217–221. https://doi.org/10.1038/nature22991.
20. Keskin, D.B., Anandappa, A.J., Sun, J., Tirosh, I., Mathewson, N.D., Li, S., Oliveira, G., Giobbie-Hurder, A., Felt, K., Gjini, E., et al. (2019). Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature *565*, 234–239. https://doi.org/10.1038/s41586-018-0792-9.
21. (2019). The Trinity CTAT Project. InferCNV of the Trinity CTAT. Project. https://github.com/broadinstitute/inferCNV.
22. Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. (2007). The Genomic Landscapes of Human Breast and Colorectal Cancers. Science *318*, 1108–1113. https://doi.org/10.1126/science.1145720.
23. Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., Kovar, C.L., Lewis, L.R., Morgan, M.B., Newsham, I.F., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337. https://doi.org/10.1038/nature11252.
24. Dredge, B.K., Stefani, G., Engelhard, C.C., and Darnell, R.B. (2005). Nova autoregulation reveals dual functions in neuronal splicing. The EMBO journal *24*, 1608–1620.
25. Guo, J., Jia, J., and Jia, R. (2015). PTBP1 and PTBP2 impaired autoregulation of SRSF3 in cancer cells. Sci. Rep. *5*, 14548. https://doi.org/10.1038/srep14548.
26. Ding, F., Su, C.J., Edmonds, K.K., Liang, G., and Elowitz, M.B. (2022). Dynamics and functional roles of splicing factor autoregulation. Cell Rep. *39*, 110985. https://doi.org/10.1016/j.celrep.2022.110985.
27. Dietrich, D.R. (1993). Toxicological and pathological applications of proliferating cell nuclear antigen (PCNA), a novel endogenous marker for cell proliferation. Crit. Rev. Toxicol. *23*, 77–109.
28. Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., and Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. *19*, 40. https://doi.org/10.1186/s13059-018-1417-1.
29. Rice, G.I., Kasher, P.R., Forte, G.M.A., Mannion, N.M., Greenwood, S.M., Szynkiewicz, M., Dickerson, J.E., Bhaskar, S.S., Zampini, M., Briggs, T.A., et al. (2012). Mutations in ADAR1 cause Aicardi-Goutieres syndrome associated with a type I interferon signature. Nat. Genet. *44*, 1243–1248.
30. Galeano, F., Tomaselli, S., Locatelli, F., and Gallo, A. (2012). A-to-I RNA editing: the "ADAR" side of human cancer. Semin. Cell Dev. Biol. *23*, 244–250.
31. Zhang, R., Deng, P., Jacobson, D., and Li, J.B. (2017). Evolutionary analysis reveals regulatory and functional landscape of coding and non-coding

RNA editing. PLoS Genet. *13*, e1006563. https://doi.org/10.1371/journal.pgen.1006563.

32. Nishikura, K. (2010). Functions and Regulation of RNA Editing by ADAR Deaminases. Annu. Rev. Biochem. *79*, 321–349. https://doi.org/10.1146/annurev-biochem-060208-105251.
33. Gu, T., Buaas, F.W., Simons, A.K., Ackert-Bicknell, C.L., Braun, R.E., and Hibbs, M.A. (2012). Canonical A-to-I and C-to-U RNA Editing Is Enriched at 3′UTRs and microRNA Target Sites in Multiple Mouse Tissues. PLoS One *7*, e33720. https://doi.org/10.1371/journal.pone.0033720.
34. Dong, X., Chen, G., Cai, Z., Li, Z., Qiu, L., Xu, H., Yuan, Y., Liu, X.L., and Liu, J. (2018). CDK13 RNA Over-Editing Mediated by ADAR1 Associates with Poor Prognosis of Hepatocellular Carcinoma Patients. Cell. Physiol. Biochem. *47*, 2602–2612. https://doi.org/10.1159/000491656.
35. Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
36. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982.
37. Venables, R.S., McLean, S., Luny, D., Moteleb, E., Morley, S., Quinlan, R.A., Lane, E.B., and Hutchison, C.J. (2001). Expression of individual lamins in basal cell carcinomas of the skin. Br. J. Cancer *84*, 512–519. https://doi.org/10.1054/bjoc.2000.1632.
38. Fang, J., Wang, H., Liu, Y., Ding, F., Ni, Y., and Shao, S. (2017). High KRT 8 expression promotes tumor progression and metastasis of gastric cancer. Cancer Sci. *108*, 178–186.
39. Golob-Schwarzl, N., Bettermann, K., Mehta, A.K., Kessler, S.M., Unterluggauer, J., Krassnig, S., Kojima, K., Chen, X., Hoshida, Y., Bardeesy, N.M., et al. (2019). High Keratin 8/18 Ratio Predicts Aggressive Hepatocellular Cancer Phenotype. Transl. Oncol. *12*, 256–268. https://doi.org/10.1016/j.tranon.2018.10.010.
40. Hendrix, M.J., Seftor, E.A., Seftor, R.E., and Trevor, K.T. (1997). Experimental co-expression of vimentin and keratin intermediate filaments in human breast cancer cells results in phenotypic interconversion and increased invasive behavior. Am. J. Pathol. *150*, 483–495.
41. Stühler, K., Köper, K., Pfeiffer, K., Tagariello, A., Souquet, M., Schwarte-Waldhoff, I., Hahn, S.A., Schmiegel, W., and Meyer, H.E. (2006). Differential proteome analysis of colon carcinoma cell line SW480 after reconstitution of the tumour suppressor Smad4. Anal. Bioanal. Chem. *386*, 1603–1612. https://doi.org/10.1007/s00216-006-0803-9.
42. Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*, 1350–1356. https://doi.org/10.1038/nm.3967.
43. Ellis, M.J., Gillette, M., Carr, S.A., Paulovich, A.G., Smith, R.D., Rodland, K.K., Townsend, R.R., Kinsinger, C., Mesri, M., Rodriguez, H., et al. (2013). Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. Cancer Discov. *3*, 1108–1112.
44. Rojas, L.A., Sethna, Z., Soares, K.C., Olcese, C., Pang, N., Patterson, E., Lihm, J., Ceglia, N., Guasp, P., Chu, A., et al. (2023). Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. Nature *618*, 144–150. https://doi.org/10.1038/s41586-023-06063-y.
45. Ren, X., Zhang, B., Li, J., Manoharan, T., Liu, B., Song, Y., Tian, S., Tan, K.-T., Ding, L., and Li, Y. (2022). Pervasive Intronic Polyadenylation Serves as a Potential Source of Cancer Neoantigens. Preprint at Research Square. https://doi.org/10.21203/rs.3.rs-1537870/v1.
46. Han, H., Irimia, M., Ross, P.J., Sung, H.-K., Alipanahi, B., David, L., Golipour, A., Gabut, M., Michael, I.P., Nachman, E.N., et al. (2013). MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. Nature *498*, 241–245. https://doi.org/10.1038/nature12270.
47. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA. Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.
48. Philpott, M., Watson, J., Thakurta, A., Brown, T., Brown, T., Oppermann, U., and Cribbs, A.P. (2021). Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. Nat. Biotechnol. *39*, 1517–1520. https://doi.org/10.1038/s41587-021-00965-w.
49. Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., Shirley, L.M., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat. Methods *12*, 519–522. https://doi.org/10.1038/nmeth.3370.
50. Al'Khafaji, A.M., Smith, J.T., Garimella, K.V., Babadi, M., Sade-Feldman, M., Gatzen, M., Sarkizova, S., Schwartz, M.A., Popic, V., Blaum, E.M., et al. (2021). High-throughput RNA isoform sequencing using programmable cDNA concatenation. bioRxiv. https://doi.org/10.1101/2021.10.01.462818.
51. ModernaTX, I., Sharp, M., and LLC, D. (2019). An Efficacy Study of Adjuvant Treatment With the Personalized Cancer Vaccine mRNA-4157 and Pembrolizumab in Participants With High-Risk Melanoma (KEYNOTE-942). https://ClinicalTrials.gov/show/NCT03897881.
52. Al'Khafaji, A.M., Smith, J.T., Garimella, K.V., Babadi, M., Popic, V., Sade-Feldman, M., Gatzen, M., Sarkizova, S., Schwartz, M.A., Blaum, E.M., et al. (2024). High-throughput RNA isoform sequencing using programmed cDNA concatenation. Nat. Biotechnol *42*, 582–586. https://doi.org/10.1038/s41587-023-01815-7.
53. Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., and Smith, N.J. (2020). Array programming with NumPy. Nature *585*, 357–362.
54. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., and Bright, J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.
55. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., III, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
56. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
57. Pardo-Palacios, F.J., Arzalluz-Luque, A., Kondratova, L., Salguero, P., Mestre-Tomás, J., Amorín, R., Estevan-Morió, E., Liu, T., Nanni, A., McIntyre, L., et al. (2024). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nat. Methods *21*, 793–797. https://doi.org/10.1038/s41592-024-02229-2.
58. Orenbuch, R., Filip, I., Comito, D., Shaman, J., Pe'er, I., and Rabadan, R. (2020). arcasHLA: high-resolution HLA typing from RNAseq. Bioinformatics *36*, 33–40. https://doi.org/10.1093/bioinformatics/btz474.
59. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, W449–W454. https://doi.org/10.1093/nar/gkaa379.
60. Eng, J.K., Jahan, T.A., and Hoopmann, M.R. (2013). Comet: An open-source MS/MS sequence database search tool. Proteomics *13*, 22–24. https://doi.org/10.1002/pmic.201200439.
61. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods *4*, 923–925. https://doi.org/10.1038/nmeth1113.
62. The, M., MacCoss, M.J., Noble, W.S., and Käll, L. (2016). Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J. Am. Soc. Mass Spectrom. *27*, 1719–1727. https://doi.org/10.1007/s13361-016-1460-7.

63. Maaten, L.v.d., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

64. (2018). Pacific Biosciences. IsoSeq v3. https://github.com/PacificBiosciences/IsoSeq.

65. Tseng, E. (2017). https://github.com/Magdoll/cDNA_Cupcake.

66. Pertea, G., and Pertea, M. (2020). GFF utilities: GffRead and GffCompare. F1000Research *9*, ISCB Comm J-304. https://doi.org/10.12688/f1000research.23297.2.

67. Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. *28*, 396–411. https://doi.org/10.1101/gr.222976.117.

68. Abugessaisa, I., Noguchi, S., Hasegawa, A., Kondo, A., Kawaji, H., Carninci, P., and Kasukawa, T. (2019). refTSS: a reference data set for human and mouse transcription start sites. J. Mol. Biol. *431*, 2407–2422.

69. Herrmann, C.J., Schmidt, R., Kanitz, A., Artimo, P., Gruber, A.J., and Zavolan, M. (2020). PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. Nucleic Acids Res. *48*, D174–D179. https://doi.org/10.1093/nar/gkz918.

70. Wang, R., Zheng, D., Yehia, G., and Tian, B. (2018). A compendium of conserved cleavage and polyadenylation events in mammalian genes. Genome Res. *28*, 1427–1441.

71. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. *47*, W199–W205.

72. Ramaswami, G., and Li, J.B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. *42*, D109–D113. https://doi.org/10.1093/nar/gkt996.

73. Kiran, A.M., O'Mahony, J.J., Sanjeev, K., and Baranov, P.V. (2012). Darned in 2013: inclusion of model organisms and linking with Wikipedia. Nucleic Acids Research *41*, D258–D261.

74. Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R., Ramaswami, G., Ariyoshi, K., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. Nature *550*, 249–254.

75. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008.

76. Moran, P.A. (1950). Notes on continuous stochastic phenomena. Biometrika *37*, 17–23.

77. Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. *43*, e78.

78. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489.

79. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421.

80. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *32*, D138–D141.

81. Almagro Armenteros, J.J., Sønderby, C.K., Sønderby, S.K., Nielsen, H., and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics *33*, 3387–3395.

82. Bulik-Sullivan, B., Busby, J., Palmer, C.D., Davis, M.J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., et al. (2018). Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. Nat. Biotechnol. *37*, 55–63. https://doi.org/10.1038/nbt.4313.

83. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J. Immunol. *199*, 3360–3368.

84. Nemhauser, G.L., Wolsey, L.A., and Fisher, M.L. (1978). An analysis of approximations for maximizing submodular set functions—I. Math. Program. *14*, 265–294.

85. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.-H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The Immune Landscape of Cancer. Immunity *48*, 812–830.e14. https://doi.org/10.1016/j.immuni.2018.03.023.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit monoclonal HA-Tag (C29F4)	Cell Signaling Technology	#3724; AB_1549585
HSP90 alpha/beta Antibody (F-8)	Santa Cruz Biotechnology	sc-13119; AB_675659
Biological samples		
Human colorectal cancer surgical resections	National University Hospital (NUH) Singapore	Table S1A
Human colorectal cancer adjacent normal surgical resections	National University Hospital (NUH) Singapore	Table S1A
Deposited data		
The Cancer Genome Atlas (TCGA)	National Cancer Institute	https://portal.gdc.cancer.gov/
10x Illumina short-read single cell RNA-seq data	European Nucleotide Archive (ENA) and Zenodo	Raw data deposited to ENA (PRJEB68074). Processed data deposited to Zenodo (https://doi.org/10.5281/zenodo.12750017)
10x PacBio long-read single cell RNA-seq data	European Nucleotide Archive (ENA) and Zenodo	Raw data deposited to ENA (PRJEB68074). Processed data deposited to Zenodo (https://doi.org/10.5281/zenodo.12750017)
UCSC Genome Broswer track for the identified isoforms	UCSC Genome Browser	https://genome.ucsc.edu/s/binzhang/COAD_Colored_PacBio
Experimental models: Cell lines		
Human normal colon cell line, CCD 841 CoN	ATCC	CRL-1790
Human CRC cell lines DLD-1	Horizon Discovery	PAR-086
Human CRC cell lines HCT116	ATCC	CCL-247
Oligonucleotides		
Sequence targeting CPNE7 isoform TCONS_00042881 Forward: GACATGCAGGTCCTGGAC Reverse: CTCATGTGTGACATGTGTGTC	This paper	N/A
Sequence targeting CPNE7 isoform TCONS_00044960 Forward: CATGGCAGGTGTGGTGTTG Reverse: GCAGTGACATGAACAGGGAC	This paper	N/A
Sequence targeting STMN3 isoform TCONS_00078414 Forward: GACCACAGGCCGGATG Reverse: GTGCCTCGCGGATCTC	This paper	N/A
Sequence targeting ZC3HAV1L isoform TCONS_00112393 Forward: GGTTTTGCGATGCTTCACTG Reverse: CATCTTCTTTACTTCTCGCAAG	This paper	N/A
for the complete list PCR primers, see Table S9		
Recombinant DNA		
pcDNA3.1 (+)	ThermoFisher Scientific	V79020
Software and algorithms		
Code for data analysis and neoepitope selection	This paper	https://doi.org/10.5281/zenodo.13338121
Python (3.8.13)	Anaconda	https://www.anaconda.com/
R (4.0.5)	Anaconda	https://www.anaconda.com/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
xgboost (1.5.0.2)	Chen and Guestrin[35]	https://xgboost.readthedocs.io/en/latest/
Numpy (1.22.4)	Harris et al., 2020[53]	https://numpy.org/
Scipy (1.10.1)	Virtanen et al., 2020[54]	https://scipy.org/
CellRanger (4.0.0)	10X Genomics	https://www.10xgenomics.com/
Seurat (3.2.3)	Stuart et al.[55]	https://satijalab.org/seurat/
cDNA cupcake (27.0.0)	Github	https://github.com/Magdoll/cDNA_Cupcake
minimap2 (2.17)	Li[56]	https://github.com/lh3/minimap2
SQANTI3 (3.3)	Pardo-Palacios et al., 2024[57]	https://github.com/ConesaLab/SQANTI3
arcasHLA (0.5.0)	Orenbuch et al.[58]	https://github.com/RabadanLab/arcasHLA
netMHCpan (4.1)	Reynisson et al., 2020[59]	https://services.healthtech.dtu.dk/services/NetMHCpan-4.1/
Comet (2022.01)	Eng et al.[60]	https://comet-ms.sourceforge.net/
Percolator (2021-10-13)	Käll et al., 2007[61] and The et al., 2016[62]	http://percolator.ms/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

CRC clinical samples

The protocols for the human studies comply with all relevant ethical regulations and are approved by the National Healthcare Group Domain Specific Review Boards (NHG DSRB ref. 2018/01032). All patients gave informed written consent. The CRC tissue samples were derived from 12 patients from National University Hospital (details on the patients and samples are summarized in Table S1).

METHOD DETAILS

Tissue processing, library construction, and single-cell sequencing

Colon tissues from patients were cut into approximately 5 mm pieces and washed three times with cold PBS until the supernatant became clear. The supernatant was discarded and the tissue pieces were further cut into 1–2 mm pieces. 5 mL of PBS containing 5 mM EDTA (EDTA-PBS) was added to the tissue fragments and incubated at 4°C for 30 min with gentle agitation, followed by centrifugation at 200 *x*g, 4°C for 5 min. The supernatant was discarded and the tissue fragments were vigorously resuspended in 5 mL of cold EDTA-PBS. The fragments were allowed to settle by gravity and the supernatant was collected in a fresh 15 mL tube. This resuspension/sedimentation step was repeated five times. The supernatant from each round was collected in a separate 15 mL tube and checked under an inverted microscope for single cells. All fractions containing single cells were combined and the cells were pelleted by centrifugation at 200 *x*g, 4°C for 5 min, and washed with cold EDTA-PBS. The single cells were resuspended in serum-free DMEM containing 0.05% trypsin and 5 U/mL DNaseI and incubated at 37°C with shaking for 15 min. The dissociated cells were centrifuged at 200 *x*g, 4°C for 5 min, the supernatant was removed and the pellet was resuspended in 2 mL of serum-free DMEM supplemented with 2 U/mL DNaseI, and filtered using a 35 mm mesh. The cells were pelleted and washed once with cold PBS and resuspended in PBS supplemented with 0.4% BSA, followed by cell quantification.

For GEM generation, 10,000 cells from each tissue sample were used to load the Chromium Next GEM Chip G (10x Genomics) and Chromium controller following the manufacturer's protocol (Chromium Next GEM Single Cell 3′ v3.1). Post GEM-RT cleanup, cDNA amplification and library construction were performed according to the manufacturer's instructions with some modifications. Prior to the fragmentation step, 20 μL of cDNA from each sample were separately aliquoted for PacBio long-read sequencing. To generate the Single Cell 3′ Gene Expression library for Illumina sequencing, 10 μL of cDNA were processed as per the 10x Genomics protocol, with an extension time of 3 min in the sample index PCR step. The final Illumina libraries were sequenced on the NovaSeq 6000 platform with S1 or S4 flowcells using the recommended sequencing protocol.

The remaining full-length cDNA from the 10x Single Cell workflow was processed for PacBio long-read sequencing. The standard PacBio protocol for single cell Iso-Seq libraries was used with the minimum number of PCR cycles (typically 12 cycles) for amplification of cDNA to increase the mass of the input DNA prior to starting library preparation. Following purification, the library was processed with the normal DNA damage and end-repair steps, and SMRT Bell adapter ligation. The final library was then bound, loaded onto 8M SMRT cells and sequenced on the PacBio Sequel II instrument with a 24-h movie time.

Plasmids

The open reading frames (ORFs) of validated neoepitopes followed by an HA tag were cloned into pcDNA3.1+ using the primers and restriction sites listed in Table S9. All constructs were verified by Sanger sequencing.

Cell culture and transfection

Human normal colon cell line, CCD 841 CoN (ATCC: CRL-1790), was cultured in Dulbecco's Modified Eagle Medium (DMEM). Human CRC cell lines, DLD-1 (Horizon Discovery: HD PAR-086) and HCT116 (ATCC: CCL-247) were cultured in Roswell Park Memorial Institute (RPMI) 1640 Medium and DMEM, respectively. All culture media were supplemented with 10% FBS, glutamine and penicillin/streptomycin. The cells were maintained at 37°C and 5% CO_2 in a humidified atmosphere. For overexpression experiments, cells were seeded at 100,000 cells per well in 12-well plates 24 h prior to transfecting 1 μg of each plasmid using Lipofectamine 3000 (Thermo Fisher) following the manufacturer's instructions.

Soft agar assay

Cells were transfected 18–24 h prior to seeding as described above. A 0.6% base agarose was prepared in 12-well plates on the day of seeding. Transfected cells were trypsinized, harvested, and counted. Seeding densities of 5,000 and 7,000 cells per well were used for CCD 841 CoN and DLD-1, respectively. The cells were resuspended in their respective growth media, mixed with agarose to a final agarose concentration of 0.3% and seeded on the prepared base. After agarose solidification, 0.5 mL of growth medium was added to each well. The cells were maintained in conditions described above and the growth medium was changed every 2–3 days. The colonies were imaged after 7–14 days under 4× magnification using the Olympus IX71 microscope and quantified using ImageJ (v1 51j8).

RNA extraction, RT-qPCR, and PCR

Trizol and the PureLink RNA Mini Kit (Thermo Fisher) were used to extract total RNA from DLD-1 and HCT116 cell lines. The ISOLATE II RNA Mini Kit (Bioline) was used to extract RNA from CRC patient samples following the manufacturer's protocol. cDNA was generated using the High Capacity cDNA Reverse Transcription Kit (Thermo Fisher). qPCR experiments were performed using PowerUp SYBR Green Master Mix for qPCR according to the manufacturer's instructions (Thermo Fisher). PCR experiments were performed using Platinum Taq DNA Polymerase (Thermo Fisher). Subsequently, PCR products were subjected to agarose gel electrophoresis, gel extraction, and purification using the QIAquick Gel Extraction Kit (Qiagen) and Sanger sequencing. Chromatograms were visualized and multiple sequence alignments were performed using SnapGene (v6.2.1). RNA-editing levels were quantified using ImageJ (v1 51j8). PCR primers used are listed in Table S9.

Protein extraction and western blot analysis

Cells were harvested 48 h post-transfection, lysed and 10 μg of lysates were fractionated using 12% SDS-PAGE followed by transfer to PVDF membranes as described previously.[12] Specific primary and secondary antibodies in 5% BSA-TBST were incubated with the membranes to probe for genes of interest. Protein expression levels were quantified using ImageJ (v1 51j8).

Mass spectrometry

To generate an in-house mass spectrometry dataset of matched tumor-normal sample pairs, <10mg tissue for each tissue were processed with the iST sample preparation kit (Preomics). Samples were taken up in 100 μL LYSE buffer and mixed with 50 mg of protein extraction beads (Diagenode). Samples were sonicated with 10–20 cycles at 30s ON/OFF in a Diagenode Bioruptor Plus until complete tissue disruption was observed. Protein concentrations were determined using the Pierce BCA Protein Assay Kit (Thermo Scientific). 20 μg of protein per sample were digested with the iST kit according to the manufacturer's instructions and samples were eluted with the fractionation add-on into 3 fractions. Digested peptides were quantified with the Quantitative Fluorometric Peptide Assay kit (Thermo Scientific), normalized to 0.1 μg/μL and stored at −20°C prior to analysis on an EASY-nLC 1200 Liquid Chromatograph (Thermo Scientific) coupled to a timsTOF fleX (Bruker) mass spectrometer. For each sample, 5 μL digested peptides (500 ng) for each fraction were injected and separated on an Aurora series column (25 cm length, 75 μm inner diameter, C-18 1.7 μm; IonOpticks) with an integrated captive spray emitter. The column was mounted on a captive spray ionization source and temperature controlled by a column oven (Sonation) at 50°C. A 105-min gradient from 2 to 40% (v/v) acetonitrile in 0.1% (v/v) formic acid at a flow of 400 nL/min was used. Spray voltage was set to 1.65 kV. The timsTOF fleX was operated with data-dependent acquisition (DDA) in PASEF mode with 10 PASEF ramps per topN acquisition cycle (cycle time 1.17s) and a target intensity of 10,000. Singly charged precursor ions were excluded based on their position in the m/z-ion mobility plane and precursor ions that reached the target intensity were dynamically excluded for 24 s.

QUANTIFICATION AND STATISTICAL ANALYSIS

Summary of bioinformatic analysis

The bioinformatic analysis workflow is summarized in Figure S10. Additionally, the command lines and parameters used in the analysis are also provided.

Illumina short-read scRNA-seq data pre-processing

CellRanger v4.0.0 (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was used for the pre-processing of 10x droplet-based scRNA-seq raw data. Specifically, 'cellranger count' with default parameters was used to align

sequencing reads to the human reference genome (GRCh38). The gene expression count matrix was then obtained based on the genomic coordinates defined by the GENCODE annotation (release v37 for GRCh38). Subsequently, 'cellranger aggr' was performed to aggregate the count matrices of samples in multiple 'cellranger count' runs. The aggregated count matrix of all samples was then loaded into Seurat (v3.2.3)[55] for subsequent analysis. A cell was filtered out if it satisfied any of the following criteria: (1) expression of fewer than 250 genes, (2) detected fewer than 300 UMIs, (3) more than 70% of the UMIs mapped to mitochondrial genes. All mitochondrial genes were discarded in subsequent analysis. This analysis resulted in a total of 18,966 high-quality cells from 22 tissue samples.

The gene expression count matrix was log-normalized using Seurat's 'NormalizeData' function. The top 2000 genes with the highest variations were selected using the 'SelectIntegrationFeatures' subroutine of Seurat, and the 'IntegrateData' subroutine was used to remove batch effects across the samples. Integration anchors were chosen based on the first thirty dimensions of the canonical correlation analysis (CCA).[55] We performed the analysis with above procedures separately on the normal and tumor samples to prevent the batch effect correction procedure from eliminating differences between the normal and tumor cells (Figure S1).

Cell type identification

We integrated the in-house data to a reference dataset, Human Colon Cancer Atlas (c295)[6] from the Broad Institute Single Cell Portal (https://singlecell.broadinstitute.org/), for more robust and reliable identification of the cell types. The c295 atlas contains gene expression profiles of 371,223 cells. Due to memory constraint, a subset of the atlas containing 50% of the cells in the c295 atlas was used as the reference based on a stratified sampling procedure across different cell types. Transfer anchors were chosen by the 'FindTransferAnchors' subroutine of Seurat using the first thirty dimensions of the principal component analysis (PCA). The 'TransferData' subroutine was used to transfer the c295's 'ClusterMidway' annotation to the in-house data via the selected anchors. The 'ClusterMidway' labels of the c295 atlas annotate cells into 20 different types, including (with boldface indicating the availability of the cell type in the in-house data).

(1) 2 epithelial cell types: normal epithelial cells (EpiN) and tumor epithelial cells (EpiT).
(2) 13 immune cell types: B cells, plasma cells, $CD4^+$ T cells, $CD8^+$ T cells, gamma-delta T cells, PLZF+ T cells, natural killer cells, innate lymphoid cells, dendritic cells, granulocytes, macrophages, mast cells, monocytes,
(3) 5 stromal cell types: **fibroblasts**, endothelial cells, pericytes, smooth muscle cells, Schwann cells

Several post-processing steps were performed for the result from the above classification: (1) We corrected all the EpiN misclassified as EpiT cells in the normal samples. (2) To improve the purity of the identified epithelial tumor cells, we trained a gradient boosting classifier (xgboost v1.5.0[35]) using the combined EpiN and EpiT cells from the c295 atlas and the in-house dataset. Following a similar 5-fold cross-validation scheme as in Pelka et al.[6] we divided the combined dataset into five independent splits and used 4-folds for training and the rest 1-fold for testing. The classifier was trained to perform the binary classification of whether a cell is a EpiT or a EpiN cell. Only when a cell came from a tumor sample and with >75% prediction probability as EpiT cells in the test split did we regard it as a candidate for the third step. (3) For all candidates from step (2), we inferred somatic copy number variations (SCNV) of the cells based on their gene expression profiles using inferCNV (v1.11.1, https://github.com/broadinstitute/infercnv).[21] The final set of tumor epithelial cells was kept as those having >10% genomic SCNV regions. The statistics of epithelial cells that passed the xgboost and the SCNV criteria are reported in Table S1E. The epithelial cells from tumor samples that did not pass either the gradient boosting criteria or the SCNV criteria were deemed as undetermined epithelial cells.

Dimensionality reduction and visualization

PCA was performed on the log-normalized and centered gene expression matrix for dimensionality reduction. t-distributed stochastic neighbor embedding (t-SNE)[63] was then performed to generate 2D embeddings for visualization. tSNE was supplied with the first 30 components of PCA and run with a perplexity value of 30.

PacBio long-read scRNA-seq data pre-processing

SMRT Link (v9.0.0, https://www.pacb.com/support/software-downloads/) was utilized to pre-process PacBio Iso-Seq long-read scRNA-seq data. We used the Isoseq-deduplication pipeline[64] recommended by the official repository of PacBio (https://github.com/PacificBiosciences/IsoSeq/) for the processing of raw data from the PacBio Sequel II system. The 'ccs' module of SMRT Link was applied to generate consensus reads (CCS Reads) from the BAM file that contained the PacBio subreads. The reads with less than 90% accuracy ('–min-rq 0.9') were discarded. The 'lima' module was utilized in orientation determination and primer removal of the reads. The cell barcodes and UMIs of each read were extracted by 'isoseq3 tag' using the library design 'T-12U-16B'. The processed reads were trimmed of the Poly(A) tail and potential concatemers were removed by the command 'isoseq3 refine'. The resulting full-length non-concatemer (FLNC) reads were clustered together according to their cell barcode and UMI with 'isoseq3 dedup' to deduplicate and generate a consensus sequence of each molecule. Those consensus sequences were then mapped to the human reference genome (GRCh38) using minimap2.[56] According to the genomic coordinates of the mapped molecules, the structures of each isoform represented in the gtf format were obtained by utilizing the script 'collapse_isoforms_by_sam.py' from the cDNA cupcake toolkit[65] (v27.0.0, https://github.com/Magdoll/cDNA_Cupcake).

Isoform quality control, filtering, and classification

We merged the isoforms identified across different samples with gffcompare.[66] We then utilized SQANTI3 (v3.3)[67] to compare the identified isoforms against the reference transcriptome annotation, which produced quality control and classification information for each isoform. To provide a comprehensive reference for SQANTI3, we merged databases from multiple sources, including GENCODE (Release v37 for GRCh38) and RefSeq (NCBI Homo sapiens Annotation Release 109 for GRCh38), for the annotation of isoform structures. With this assembled reference, SQANTI3 classified each isoform into FSM, ISM, NIC, NNC and several other structural types.[67] The rule-based filter functionality of SQANTI3 was used to filter out the isoforms that were likely to be sequencing artifacts such as intra-priming, reverse transcriptase template switching (RTS), and non-canonical splice sites with low read coverage. The remaining isoforms and their associated barcodes are assembled into an isoform-level expression matrix $\mathbf{X}^{(\mathrm{iso})} \in \mathbb{R}^{N_{\mathrm{iso}} \times N_{\mathrm{cells}}}$, where N_{iso} is the number of isoforms, and N_{cells} is the number of all cells associated with those isoforms. Finally, we searched the barcodes from PacBio data against the cell barcodes detected in the same sample in the 10x Illumina sequencing data to calculate the expression (UMI counts) of each isoform in each cell. To evaluate the quality of the identified isoforms, we examined whether the 5′ and 3′ ends of each isoform are supported by the TSS and PAS reference. The TSS reference was a union of TSSs from GENCODE, RefSeq and refTSS (v3.1),[68] while the PAS reference contains all transcript 3′ ends in GENCODE, RefSeq and PAS from PolyASite (v2.0)[69] and PolyADB (v3.2).[70] Furthermore, splice junctions from each identified isoform were overlapped with junction reads extracted from the RNA-seq data in The Cancer Genome Atlas colon adenocarcinoma cohort (TCGA-COAD).

Splice event extraction and analysis

For all isoforms that had passed the SQANTI3's rule-based filter, we utilized SUPPA (v2.3)[28] for the extraction of seven types of AS events including alternative 5′ splice sites (A5), alternative 3′ splice sites (A3), alternative first exon (AF), alternative last exon (AL), skipped exon (SE), mutually exclusive exon (MX) and retained intron (RI). SUPPA analyzed the gtf file obtained from the isoform calling and quality control pipeline to produce seven event files (the 'ioe' files) each containing one of the above event types based on the pairwise comparison of the input isoform structures. Taking into consideration the accuracy of splice sites and the inaccuracy of TSS and TTS, we required SUPPA2 to use stringent boundaries for the splice sites of the seven event types, but flexible boundaries (with a variability of 48nt) for the TSS and TTS coordinates of AF and AL.

Identification of dysregulated gene expression and transcript structures

Dysregulated gene expression (DGE) was identified by comparing the expression of genes (log-normalized UMI) in cells from one group (cell type or subtype) to that from another group. In brief, for each gene, we first counted the number of cells (n_1) expressing this gene with the expression threshold (0) and the total number of cells (n_2) in each group. Next, we applied Fisher's exact test on the 2 × 2 contingency table formed by n_1 and n_2 from two groups. In addition, we also applied the Mann-Whitney U test (Wilcoxon Rank-Sum test) on the gene expression (log-normalized UMI) in cells from different groups. A significantly upregulated gene was identified by requiring: (1) Benjamini-Hochberg (BH) - adjusted p values (FDR) from Fisher's exact test <0.01, (2) odds ratio >2 and (3) log2 fold change >0. Significantly downregulated genes were identified using the identical p-value threshold, but with the opposite directional changes (log2 fold change <0 and odds ratio <1/2). Dysregulation in isoform expression was identified similarly using the expression of each isoform. For evaluating concordance of DGE quantified by the two sequencing methods, we used genes showing significant changes in both measurements.

To identify dysregulated transcript structures (DTSs), we counted the number of PacBio UMIs supporting a specific isoform versus all alternative isoforms of the same gene for cells from the two groups. The significance of DTSs was computed using Fisher's exact test on the 2 × 2 contingency table formed by four numbers from the two groups of cells. The significant DTSs were identified by requiring: (1) BH-adjusted p-value (FDR) < 0.01, (2) the percentage of isoform changes >0 and (3) odds ratio >2, or (1) BH-adjusted p-value (FDR) < 0.01, (2) the percentage of isoform changes <0 and (3) odds ratio <1/2.

GO enrichment analysis

The list of genes with DTS from comparing EpiT to EpiN was uploaded to WebGestalt[71] as input and all the genes with detected expression in these two cell types were used as the background to identify the enriched GO terms (FDR <0.05, enrichment score >2).

Calling RNA-editing events

We assembled a comprehensive reference of potential RNA-editing sites from three different sources: (1) the RADAR database,[72] (2) the DARNED database[73] and (3) Tan et al., 2017.[74] For each putative RNA-editing site in the reference, we systematically examined the evidence of reads supporting an A>G change at that locus (or a T>C change on the negative strand) using the 'mpileup' subcommand of bcftools.[75] We then associated each event with the isoforms and cell barcodes from which the event was discovered. To remove interference from genomic SNVs, we searched the candidate RNA-editing sites against the dbSNP database and discarded the sites with an allele frequency greater than 0.01 in any genomic studies it had been concerned. We retained RNA-editing sites with more than 10 read coverage and computed an edit ratio for each site, defined as the number of edited reads divided by the number of total read coverage. For these RNA-editing sites, we additionally calculated their per isoform edit ratio using the reads that can be unambiguously identified to a specific isoform and if the per isoform read coverage is greater than five. The per isoform edit ratio is defined as the number of edited reads divided by the number of total read coverage from the isoform (Figure 4A).

Lineage and trajectory analysis

We employed monocle 3[36] for the single-cell lineage and trajectory analysis on the epithelial component of the short-read scRNA-seq dataset. The principal graph was constructed using the 'learn_graph' subroutine of monocle 3 with default parameters using the tSNE coordinates and the nodes corresponding to the four stem cells subclusters were designated as the root for diffusion pseudotime inference. The epithelial cells that were also detected in the PacBio long-read scRNA-seq were associated with their isoform level expression values. The expression of genes and isoforms, as well as the percentage of isoforms that were associated with a specific lineage, were identified based on Moran's I statistic.[76]

Estimation of the signature scores of EpiT subtypes in bulk samples

We identified the top differentially expressed genes in the EpiT subtypes cE02 and cE03 (Table S6). We then computed the z-scores of those up- and downregulated genes in each bulk sample. The signature score of an EpiT subtype is defined as the averaged z-scores of the upregulated genes minus the averaged z-scores of the downregulated genes.

Survival analysis

We correlated the gene expression or activity scores with the PFS and OS of the CRC patients based on the clinical data from TCGA-COAD and CRCSC using the Cox proportional hazards regression model. The Kaplan-Meier method was used to estimate the survival function of the patients under each condition.

Proteomic analysis of isoforms

To comprehensively characterize the proteome in CRC, GeneMarkS-T[77] was used to predict ORFs in the PacBio-identified isoforms. For comparison and annotation of the predicted ORFs, we collected the human reference proteome from UniProt (release 2022-02),[78] including all sequences from the reviewed (SwissProt), unreviewed (TrEmbl) and spliced isoform (Varsplice) collections at protein existence (PE) level PE = 1 and PE = 2. Using blastp,[79] we searched each predicted ORF sequence against the UniProt reference proteome, and the reference sequence with the highest alignment score was assigned to it as its nearest homolog in the database. The ORFs with a similarity of less than 99% to their nearest homolog in the reference were considered novel protein sequences. Protein domains of each ORF sequence, as defined by the Pfam-A hidden Markov model (HMM)-based multiple sequence alignments,[80] were annotated by hmmscan (http://hmmer.org/). We regarded a novel ORF sequence as having a domain gain if it contained a proper superset of the Pfam domains from its nearest UniProt homolog, or a domain loss if it contained only a proper subset of them. We also inferred the sensitivity of the PacBio-identified isoforms to nonsense-mediated decay (NMD) based on the relative position of the predicted stop codon from the last splice junction. Lastly, the subcellular localization of the ORF-derived proteins was predicted using DeepLoc.[81]

Mass spectrometry data analysis

To validate the predicted ORFs, particularly the novel ORFs, we collected mass spectrometry (MS) data from three independent sources: (1) a CRC cohort from the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC), (2) an in-house MS dataset consisting of 22 samples from 11 CRC patients (11 tumor and 11 paired adjacent normal), and (3) an in-house MS data from a colon cancer cell line, HCT116. With the same pipeline as a previous study,[82] we searched MS2 spectra from the MS data against a reference by integrating annotated protein sequence from UniProt (release 2022-02) and novel ORFs predicted based on isoforms derived from LR-seq using Comet.[60] Five peptides were reported for each spectrum query and these peptides were further ranked and scored for final identification using Percolator.[61] The validated novel ORFs were defined by its inclusion of at least one novel peptide (FDR <0.01) in the MS data that does not match with any protein sequence from the UniProt database.

Proposing putative neoepitopes for cancer vaccine development

To obtain a shortlist of tumor-specific novel transcript isoforms, we extracted the NIC and NNC isoforms identified in our long-read scRNA-seq profiles and quantified their expression levels in the TCGA-COAD bulk samples along with the known isoforms in GENCODE and RefSeq. Tumor-specificity was ensured by filtering for isoforms discovered only from the EpiT cells and not expressed (TPM <0.5) in the TCGA-COAD normal samples and additionally by requiring each novel isoform to possess at least one novel junction which has supporting reads in <5% of TCGA normal samples and >5% TCGA tumor samples. To ensure active translation of the isoforms, we prioritized the isoforms from the shortlist whose ORF subsequences were supported at least once by the MS data from tumor samples but without any peptides detected in MS data from normal samples. Out of five isoforms in this shortlist selected for experimental confirmation, we validated the unique splice junction sequences and RT-PCR products for three isoforms arising from AF and A5SS AS events, and one isoform with retained introns (RI) in CRC cell lines, as well as at least two of the five matched patient samples tested.

To derive the set of tumor-specific neoepitopes from the tumor-specific novel transcript isoforms, we assembled a collection of all the *k*-mers ($k = 9$) that are present in the predicted ORF sequences and further curated a subset containing tumor-specific novel *k*-mers by requiring: (1) the *k*-mers are not in the UniProt reference, (2) the *k*-mers are only present in ORFs in transcript isoforms from EpiT but not from EpiN or other cell types, (3) the *k*-mers are from the transcript isoforms that are expressed (TPM >0.5) in more than five TCGA-COAD tumor samples and not expressed (TPM <0.5) in the normal samples (4) the *k*-mers are from the

ORFs in the transcript isoforms classified as NIC and NNC. In this way, we obtained a set $\mathcal{S}$ of candidate neoepitope *k*-mers that are from recurrent tumor-specific isoforms across tumor samples.

For a given population, we predicted the binding affinity of each neoepitope *k*-mer against the frequent HLA alleles in the population (represented as the set $\mathcal{H}$) using netMHCpan (v4.1).[83] We used the rank of elution score, EL_rank < 2 as the threshold of binding/non-binding for a *k*-mer to a specific HLA allele. In this way, we obtained a mapping $\mathcal{M}$ that represented the HLA alleles that a neoepitope *k*-mer can bind to. For example, $\mathcal{M}[s]$ ($\mathcal{M}[s] \subseteq \mathcal{H}$) contains the HLA alleles to which neoepitope s ($s \in \mathcal{S}$) binds.

To optimize a list of neoepitope *k*-mers with a high probability of containing at least one epitope with binding potential to a patient's HLA allele for a given population, we searched the combination of neoepitope *k*-mers in the shortlist to cover as many frequent HLA alleles in a population as possible. Such an optimization task was formulated as a maximum coverage problem:

Input: The size of the shortlist n and the sets of HLA alleles covered by each epitope $\{M[s_1], M[s_2], \ldots, M[s_{|\mathcal{S}|}]\}$.

Output: The subset $\mathcal{S}' \subseteq \mathcal{S}$ with $|\mathcal{S}'| \leq n$ such that $\left| \bigcup_{\{s \in \mathcal{S}'\}} M[s] \right|$ is maximized.

Additionally, we took into account the HLA's allele frequency by assigning it as the weight in a weighted maximum coverage problem:

Input: The size of the shortlist n, the sets of HLA alleles covered by each epitope $\{M[s_1], M[s_2], \ldots, M[s_{|\mathcal{S}|}]\}$, and the weight of each HLA allele *a* ($a \in \mathcal{H}$), $\mathcal{W}[a]$.

Output: The subset $\mathcal{S}' \subseteq \mathcal{S}$ with $|\mathcal{S}'| \leq n$ such that $\sum_{a \in \bigcup_{\{s \in \mathcal{S}'\}} M[s]} \mathcal{W}[a]$ is maximized.

As both the weighted and unweighted maximum coverage problems are NP-hard,[84] we used the following greedy algorithm for the selection of an epitope shortlist with size *n* (Algorithm 1, SELECT-KMERS-FOR-POPULATION). We omitted the algorithm for the unweighted version in the presentation here as it is a special case of the weighted version when $\mathcal{W}[a] = 1$ ($\forall a \in \mathcal{H}$). On each iteration, we selected an epitope *k*-mer not yet in the shortlist that would result in the maximum gain in the weight of covered HLA alleles (Algorithm 1, SELECT-BEST-KMER). If the HLA alleles were already completely covered before the size of the shortlist reaches n, we reset the set of uncovered HLA alleles and continued the iteration. In this way, the remaining epitopes in the shortlist were optimized again for the maximum coverage of HLA alleles in the population.

To optimize the neoepitope shortlist for the TCGA-COAD patients, we first obtained the HLA subtyping of these patients from Thorsson et al.[85] and computed the binding affinity of the neoepitopes in set $\mathcal{S}$ to all the HLA alleles found in them. We defined the coverage of a neoepitope for a patient as at least one HLA allele of the patient covered by the neoepitope. We then optimized the shortlist of neoepitopes for the maximum coverage of TCGA patients by solving a similarly defined unweighted maximum coverage problem for the population HLA alleles. This resulted in a similar procedure (Algorithm 1, SELECT-KMERS-FOR-PATIENT-GROUP) as above, by redefining $M[s]$ to be the patients covered by neoepitope *s*.

Algorithm 1. Select splicing-derived neoepitopes for cancer vaccine development

function SELECT-KMERS-FOR-POPULATION ($\mathcal{S}, \mathcal{M}, \mathcal{H}, \mathcal{W}, n$)
inputs:
  $\mathcal{S}$, the set of neoepitope *k*-mers from recurrent tumor-specific isoforms
$\mathcal{M}$, the mapping from *k*-mers to the HLA alleles they bind
$\mathcal{H}$, the set of HLA alleles in the population
$\mathcal{W}$, the mapping from HLA alleles to their allele frequency
n, the number of putative *k*-mers for cancer vaccine development
outputs:
S', the putative set of *k*-mers for cancer vaccine development
  c', coverage of allele frequency by S'
    $S' \leftarrow \varnothing$
    $c' \leftarrow 0$
    $\mathcal{U} \leftarrow \mathcal{H}$//the set of uncovered HLA alleles
    for i = 1 to n do
      $s_i \leftarrow$ SELECT-BEST-KMER ($\mathcal{S}, \mathcal{M}, \mathcal{U}, \mathcal{W}$)
      $S' \leftarrow S' \bigcup \{s_i\}, \mathcal{S} \leftarrow \mathcal{S} \setminus \{s_i\}$
      $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{M}(s_i)$ if $\mathcal{U} \setminus \mathcal{M}(s_i) \neq \varnothing$ else $\mathcal{U} \leftarrow \mathcal{H}$//reset uncovered HLA alleles if all covered
    end
    $c' \leftarrow$ SUM($\{\mathcal{W}[a]$ foreach a in $S'\}$)
    return S', c'

```
function SELECT-KMERS-FOR-PATIENT-GROUP (S, M, P, n)
inputs:
  S, the set of neoepitope k-mers from recurrent tumor-specific isoforms
  M, the mapping from k-mers to the patients whose HLA they bind to
  P, the set of patients
  n, number of putative k-mers for cancer vaccine development
outputs:
  S', the putative set of k-mers for cancer vaccine development
  c', the number of patients covered by S'
    S' ← ∅
    c' ← 0
    U ← P//the set of uncovered patients
    for i = 1 to n do
    s_i ← SELECT-BEST-KMER (S, M, U, {p : 1 foreach p in P})
      S' ← S' ⋃ {s_i}, S ← S {s_i}
      U ← U M(s_i) if U M(s_i) ≠ ∅ else U ← H//reset uncovered patients if all covered
    end.
    c' = SIZE(S')
    return S', c'
```

```
function SELECT-BEST-KMER (S, M, U, W)
inputs:
  S, the set of k-mers from recurrent tumor-specific isoforms
  M, the mapping from k-mers to the identities (HLA alleles or patients) they bind to
  U, the set of uncovered identities
  W, the weight of the identities
outputs:
  s_best, the selected best k-mer in set S
    s_best = POP(S) , c_best = SUM({W[a] foreach a in M[s_best]})
    foreach s in S
      c = SUM( { W[a] foreach x in M[a] ⋂ U } )
      if c > c_best
        c_best ← c, s_best ← s
      end
      end
      return s_best
```

Command lines of software tools used in this study

Mapping in-house short-read scRNA-seq profiles to bc295

```
anchors <- FindTransferAnchors(
      reference = [reference Seurat object],
      query = [query Seurat object],
      dims = 1:30.
  )
  prediction_cluster_midway = TransferData(
      anchorset = anchors, refdata = [reference Seurat object 'ClusterMidway' labels],
      dims = 1:30.
  )
```

Alignment of the long-read molecules to genome using minimap2

```
minimap2 -t 30 -ax splice -uf –secondary = no -C5 [path to genome fasta file] [PacBio molecule fasta file] > [alignment output file]
```

Identification of long-read isoforms using cDNA cupcake

```
collapse_isoforms_by_sam.py –input [PacBio molecule fasta file] \
      –bam [alignment output file] -c 0.99 -i 0.95 \
      –gen_mol_count \
```

```
    -o [output file prefix] \
    –cpus 20.
```

Classification of isoforms using SQANTI3

```
  sqanti3_qc.py \
      –gtf [path to cDNA cupcake's gff output file] \
      [path to the reference transcript isoform gtf file] [path to the genome fasta file] \
      –fl_count [path to cDNA cupcake's isoform abundance file] \
      –cage_peak [reference CAGE peaks file] \
      –polyA_motif_list [reference polyA motifs] \
      –polyA_peak [reference polyA peaks file] \
      –dir [output directory]
```

Subtyping of the normal and tumor epithelial cells using xgboost

Training the xgboost model:

```
  library(xgboost)
  bst <- xgb.train(
    data = dtrain, max.depth = 4, eta = 0.5, nthread = 50, nrounds = 200, objective = "multi:softmax",
    eval_metric = "merror",num_class = n_class,
    verbose = 2)
  )
```

Inference with the trained xgboost model

```
Prediction <- predict(bst,dtest)
```

Genotyping of patients' HLA alleles using arcasHLA[58]

```
arcasHLA extract –single [path to the bam file of long-read or short-read sequencing] -o [output directory]
  arcasHLA genotype –single –min_count 3 [arcasHLA extracted fastq file] -g A,B,C -o [output directory] -t 8.
```

Prediction of the HLA binding affinity of neoepitopes using netMHCpan

```
netMHCpan -a [MHC allele] -l 9 -f [neoepitope fasta file]
```

Peptides identification from MS data with Comet and Percolator

```
comet.linux.exe -Pcomet_param_file [path to the mzML file]
  crux percolator –overwrite T –output-dir sample –test-fdr 0.1 –train-fdr 0.1 [path to the pin format file outputted by comet]
```

Some parameters in comet_param_file was shown below

```
decoy_search = 2
  peptide_mass_tolerance = 3
  peptide_mass_units = 2
  mass_type_parent = 1
  mass_type_fragment = 1
  precursor_tolerance_type = 0
  isotope_error = 0
  search_enzyme_number = 1
  num_enzyme_termini = 2
  allowed_missed_cleavage = 2
  variable_mod01 = 15.9949 M 0 3 -1 0 0 0.0
  max_variable_mods_in_peptide = 5
  require_variable_mod = 0
  fragment_bin_tol = 1.0005
  fragment_bin_offset = 0.4
  theoretical_fragment_ions = 1
  output_percolatorfile = 1
  ms_level = 2
  sample_enzyme_number = 1
  digest_mass_range = 600.0 5000.0
  peptide_length_range = 5 63
  max_duplicate_proteins = 20
  max_fragment_charge = 3
  max_precursor_charge = 6
  minimum_peaks = 10
  remove_precursor_tolerance = 1.5.
```