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

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

[1]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

[2]Computational Bioscience Research Center, King Abdullah University of Science and Technology

[3]Cancer Science Institute of Singapore, National University of Singapore, Singapore

[4]Core Labs, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

[5]Department of Pathology, National University Health System, Singapore

[6]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[7]Division of Colorectal Surgery, University Surgical Cluster, National University Health System, Singapore

[8]NUS Centre for Cancer Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[9]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[10]Present address: Peter MacCallum Cancer Centre, Melbourne, Australia

[11]These authors contributed equally to this work

*Correspondence should be addressed to Yvonne Tay, Xin Gao and Bin Zhang

Email: Yvonne Tay, yvonnetay@nus.edu.sg; Xin Gao, xin.gao@kaust.edu.sa; Bin Zhang, bin.zhang@kaust.edu.sa

Summary

Colorectal cancer (CRC) is the second leading cause of cancer death worldwide. In recent years, short-read single-cell RNA sequencing (scRNA-seq) has been instrumental in deciphering tumor cell heterogeneities. However, these studies only enable gene-level expression quantification but neglect alterations in transcript structures, which arise from alternative end processing or splicing, and are frequently observed in cancer. In this study, we integrated short- and long-read scRNA-seq of CRC patient samples to build the first isoform-resolution CRC transcriptomic atlas. We identified 394 dysregulated transcript structures in tumor epithelial cells, including 299 resulting from various combinations of multiple splicing events. Secondly, we characterized genes and isoforms associated with epithelial lineages and subpopulations that exhibit distinct prognoses. Finally, we built an algorithm that integrated novel peptides derived from predicted ORFs of recurrent tumor-specific transcripts with mass spectrometry data and identified a panel of recurring neoepitopes that may aid the development of neoantigen-based cancer vaccines.

Keywords

colorectal cancer, long-read RNA-seq, scRNA-seq, neoantigen

Introduction

The heterogeneity of cells and their composition is critical for cancer progression, patient survival and response to therapy (Burrell et al., 2013; Meacham and Morrison, 2013). Single-cell RNA sequencing (scRNA-seq) enables RNA expression profiling for thousands of genes in hundreds and thousands of cells in parallel to facilitate 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 (Becker et al., 2022; Li et al., 2017; Pelka et al., 2021; Zhang et al., 2020). 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 (Joanito et al., 2022; Lee et al., 2020).

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 in which widespread alterations in transcript structure arising from 5' and 3' ends, and alternative splicing (AS) differences are frequently observed (Demircioğlu et al., 2019; Kahles et al., 2018; Xia et al., 2014). Recently, we reported that pervasive splicing within the 3'-untranslated region (3' UTR) is upregulated in cancer to promote oncogene expression and tumorigenesis (Chan et al., 2022). 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 (Huang et al., 2021; Veiga Diogo et al., 2022). 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 read frames (ORF) encoding unique proteins that serve as a potential source of neoantigen in tumor cells (Sellars et al., 2022). Several studies have reported that tumor-associated intron retention, neo-splice junction, exitron and RNA editing could generate neoantigens (Kahles et al., 2018; Smart et al., 2018; Wang et al., 2021; Zhang et al., 2018). To date, only neoantigens

34 derived from DNA mutations with the potential to activate the immune response have been used to
35 generate personalized neoantigen vaccines in clinical trials for melanoma and glioma treatment
36 (Keskin et al., 2019; Ott et al., 2017). Studies for neoantigens arising from aberrant RNA processing
37 are limited and primarily based on short-read RNA-seq data that rely heavily on annotations that
38 could compromise the precision of full ORF predictions. Conversely, the identification of full-
39 length tumor-specific anomalous transcripts using LR-seq may enable predictions of complete
40 novel ORFs to more accurately and exhaustively derive neoepitopes. Nevertheless, this concept
41 and its feasibility have not been explored.

42 Here, we present a study integrating short-read and PacBio Iso-Seq-based long-read scRNA-seq of
43 matched CRC clinical samples to build a full-length transcriptomic atlas and investigate isoform-
44 specific dysregulations in CRC. We identified 394 DTSs from 273 genes in the tumor epithelial
45 cells, many of which were caused by the coupling of multiple splicing events. Additionally, we
46 detected tumor cell-associated and isoform-specific RNA editing events. By classifying normal
47 (EpiN) and tumor epithelial (EpiT) cells into subpopulations, we unveiled genes and transcripts
48 associated with three EpiN lineages, as well as three EpiT subpopulations comprising more than
49 90% of malignant cells, each linked to a distinct prognosis. Finally, we predicted the proteome of
50 CRC based on the full-length transcriptome. By integrating the novel ORFs from recurrent tumor-
51 specific transcripts and mass spectrometry (MS) data, we developed an algorithm to build a panel
52 of 24 neoepitopes with strong binding affinities to MHC molecules for a wide range of CRC
53 patients, which may be exploited for the development of universal neoantigen-based cancer
54 vaccines.

55 ## Results

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

57 To study the landscape of full-length transcript isoforms in human CRC, we isolated single cells
58 from eight tumor and eleven adjacent normal samples derived from 12 CRC patients, generated
59 cDNA libraries using the 10x Genomics droplet-based method and performed matched Illumina
60 short-read sequencing and PacBio Iso-Seq (**Figure 1A, Table S1**). We obtained a total of 19,015
61 high-quality cells with Illumina scRNA-seq data (**Methods**). To accurately identify the cell types,
62 we mapped these cells to the Human Colon Cancer Atlas (c295), which is currently the most
63 comprehensive single-cell atlas of CRC (Pelka et al., 2021). By integrating these two datasets
64 (**Figure S1**, **Methods**), the cells were grouped into 20 major clusters (**Figure 1B**). We found that
65 cells from the in-house tumor and normal samples were embedded into the c295 cell clusters
66 without obvious batch effects (**Figure 1C**). Sixteen out of 20 cell types from c295 were detected in

67 the in-house dataset, of which seven consisted of more than 100 cells. Among these, most of the
68 cells were epithelial cells, followed by immune and stromal cells (**Figure 1D**).

69 Interestingly, compared to the normal samples, the tumor samples contained a higher number and
70 proportion of immune cells, suggesting higher immunological activity in the tumor
71 microenvironment (**Figure 1E**). To ensure robust identification of malignant epithelial cells, we
72 inferred somatic copy number variations (SCNVs) of the cells from the normal and tumor samples
73 using inferCNV (The Trinity CTAT Project, 2019). In general, the normal samples did not show
74 significant SCNVs across the chromosomes, whereas the tumor samples displayed SCNVs in
75 multiple regions (**Figure 1F**). These regions were generally consistent within each sample and
76 located within previously characterized cytobands, such as amplifications in 1q, 7p, 8q, 19q, 20p
77 and 20q, as well as deletions in 4q, 15q and 22q (Muzny et al., 2012; Wood et al., 2007). Using the
78 inferred SCNVs and gene expression profiles (**Methods**), we obtained a collection of 3,943 and
79 11,388 reliable tumor (EpiT) and normal (EpiN) epithelial cells, respectively.

80 In addition to short-read scRNA-seq, we also performed long-read scRNA-seq on three normal and
81 four tumor samples using PacBio Iso-Seq (**Figure 1A**). Among the derived 19,264,405 circular
82 consensus long-read sequencing (CCS) reads, 11,481,904 (59.6%) full-length non-concatemer
83 (FLNC) reads were obtained (**Methods**). We identified 125,205 unique transcript isoforms from
84 17,753 genes, of which 23,379 were detected in both the normal and tumor samples. Based on the
85 comparison of the splice junctions (SJs) of each aligned isoform with the reference transcriptome
86 (GENCODE v37 and NCBI RefSeq release 109), we classified each transcript into four main
87 categories: full splice match (FSM), incomplete splice match (ISM), novel in catalog (NIC) and
88 novel not in catalog (NNC), as well as several other non-specific categories (antisense, genic,
89 intergenic, fusion, etc.). FSMs and ISMs had fully or partially matched SJs against the reference,
90 whereas the NIC and NNC transcripts contain novel SJs from a combination of known and novel
91 splice sites, respectively (**Figure 2A**). In total, we identified 31,837 (25.43%) FSM, 58,570
92 (46.78%) ISM, 13,291 (10.62%) NIC, 18,644 (14.89%) NNC and 2,863 (2.29%) isoforms in the
93 other categories (**Figures S2A** and **S2B**).

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

100 slightly higher proportion of novel isoforms (NIC and NNC), suggesting more transcriptomic
101 abnormalities in tumor cells.

102 We further evaluated the quality of the detected isoforms by validating their structural components.
103 The 5'-end of >60% and 3'-end of >70% of the isoforms overlapped with annotated transcription
104 start sites (TSSs) and termination sites (TTSs) in each cell type (**Figure 2C, Methods**). All the
105 splice junctions from >99% of FSM and ISM isoforms were also detected in colon adenocarcinoma
106 patients from the cancer genome atlas (TCGA-COAD) (**Figure 2D**). ~5% and >30% of NIC and
107 NNC isoforms, respectively, contained at least one splice junction that has not been identified in
108 the TCGA-COAD RNA-seq data, suggesting that some novel splice junctions could be more
109 effectively captured by LR-seq. To evaluate the concordance between the long- and short-read
110 scRNA-seq data, we compared and observed a high overlap of the expressed genes identified in
111 both datasets (**Figure S3B**). Furthermore, we correlated the gene expression quantified by these
112 two methods for the same cell types and across different cell types. The correlations for both EpiN
113 and EpiT were relatively high ($r = 0.77, p < 2.2 \times 10^{-308}$ and $r = 0.67, p < 2.2 \times 10^{-308}$)
114 compared to other cell types with fewer cells (**Figures S3C** and **S3D**).

115 Among the 11,138 genes detected by long-read scRNA-seq, we found that 63% expressed multiple
116 isoforms, with 13% having more than 6 isoforms (**Figure 2E**). For example, we detected 53
117 isoforms for the epithelial cell surface marker gene *EPCAM*. In addition to the 16 known isoforms
118 (FSM and ISM), we identified 12 NIC and 25 NNC isoforms with the latter arising from 11 and 17
119 novel 5' and 3' splice sites (**Figure 2F**). These results indicate a high transcriptomic complexity in
120 CRC tumors, which has been overlooked in previous studies.

121 **Dysregulation of transcript structure and RNA editing in tumor epithelial cells**

122 To systematically investigate the dysregulation of RNA isoforms in CRC, we defined two types of
123 dysregulations by comparing gene and isoform expression between EpiN and EpiT (**Figure 3A**).
124 Dysregulated gene expression (DGE) refers to genes that have significantly different numbers of
125 overall transcripts between the two cell types, while dysregulated transcript structure (DTS) is
126 defined as a multi-isoform gene that expresses significantly different proportions of its isoforms
127 between EpiN and EpiT (**Methods**). We identified 1,154 DGE events and 273 genes with DTS
128 (**Figure 3B**). We observed more DGEs that were upregulated (1,000) than downregulated (154),
129 suggesting high transcriptional activity in tumor cells (**Figure 3B**). Interestingly, we found that a
130 higher frequency of genes with DTS had DGE, while single-isoform genes tend to be less
131 dysregulated (**Figure 3B**). Gene ontology (GO) enrichment analysis of the genes with DTSs
132 showed enrichment of pathways related to RNA regulation, such as mRNA binding, translational

133 initiation and RNA catabolic process (**Figure 3C**). This suggests potential autoregulation of these
134 genes, which is frequently observed for RNA-binding proteins, particularly splicing factors (Ding
135 et al., 2022; Dredge et al., 2005; Guo et al., 2015). Several immunological pathways were also
136 implicated, indicating that DTS may contribute to dysregulated immunological activity in tumor
137 cells. For example, the proliferating cell nuclear antigen gene, *PCNA*, presented both DGE and
138 DTS. As a cell proliferation marker (Dietrich, 1993), its expression was upregulated in EpiT and
139 detected in a higher proportion of cells (**Figure 3D**). The EpiN and EpiT cells preferentially utilized
140 distinct alternative first exons to differentially express the three *PCNA* isoforms. Overall, isoform
141 proportion changes accounted only for a small fraction of transcripts with differential absolute
142 expression levels between EpiT and EpiN for all four isoform categories, suggesting that
143 transcriptional activity may have a more profound impact on isoform expression (**Figure 3E**).

144 We further applied SUPPA2 (Trincado et al., 2018) to extract AS events for the identified transcript
145 isoforms, which were categorized into alternative 3' splice site (A3SS), alternative 5' splice site
146 (A5SS), alternative first exon (AF), alternative last exon (AL), retained intron (RI), skipped exon
147 (SE) and mutually exclusive exon (MX) (**Figure S4A**, **Methods**). Consistent with previous
148 observations (Huang et al., 2021; Veiga Diogo et al., 2022), the most abundant AS events were
149 from the AF, RI and SE categories (**Figure S4B**). We inspected how each isoform with DTS could
150 arise from the different AS categories and found that isoforms derived from AF, A3 and RI were
151 more enriched with DTS (**Figure 3F**, diagonal). More importantly, our long-read scRNA-seq
152 captured the structure of full-length isoforms, allowing us to investigate the coupling between
153 different AS categories. We observed that the isoforms derived from coupled AF-RI, AF-A3 and
154 A3-RI had the highest enrichment of DTS isoforms (**Figure 3F**, off-diagonal). For example, among
155 the *ARGHDIA* isoforms, those utilizing the distal first exon can retain any one of the three
156 downstream introns (**Figure 3G**), while the *TMEM259* isoforms using a proximal downstream 3'
157 splice site can retain an upstream intron (**Figure 3H**). Our findings underline the unique advantage
158 of LR-seq in capturing isoform complexity that have thus far eluded the short-read sequencing
159 technology.

160 In addition to alternative splicing, post-transcriptional RNA modifications on single nucleotides
161 could also impact the mRNA sequence and expression without altering transcript structures.
162 Specifically, ADAR-mediated A-to-I RNA editing can modify both coding and non-coding regions
163 of mRNAs, as well as long non-coding RNAs, and has been implicated in several autoimmune
164 disorders (Rice et al., 2012) and multiple cancers (Galeano et al., 2012).

165 We systematically identified the RNA editing events from isoforms sequenced using long-read
166 scRNA-seq by counting the number of edited reads (**Methods**). To characterize the RNA editing
167 level of a site, we computed RNA editing level per site (REPS) and per isoform (REPI), which are
168 the ratio of the number of all edited reads against all reads for all isoforms and against all reads for
169 a specific isoform, respectively (**Figure 4A, Table S2**). Of note, REPI calculation is only possible
170 with LR-seq as it allows unambiguous association of each edited mRNA read with the isoform of
171 origin.

172 A total of 196 RNA editing sites with sufficient sequencing read support were identified in EpiN
173 and EpiT cells. Most of these sites were on non-coding transcripts or the untranslated regions of
174 coding transcripts, with more events in the 3'UTRs than 5'UTRs (**Figure S5A**), which is consistent
175 with previous studies (Gu et al., 2012; Nishikura, 2010; Zhang et al., 2017). Furthermore, 10 sites
176 were found in the coding regions, which could potentially introduce codon alterations. For example,
177 the tumor-associated event at chr7:39950949(+) that results in the amino acid change Q103R in
178 CDK13 (**Figure 4B**), was also observed in hepatocellular carcinoma (Dong et al., 2018). Based on
179 the REPS, RNA editing events in genes such as *CDK13*, *CYP2R1* and *NEAT1* were significantly
180 upregulated in EpiT compared to EpiN (**Figure 4B**). We validated these events using normal colon,
181 CCD 841 CoN, and CRC cell lines, DLD-1 and HCT116, as well as patient samples. The RNA
182 editing levels for *CDK13, CYP2R1* and *NEAT1* were variable and did not follow any trend in the
183 cell lines (**Figure S5B**). However, these were consistently higher in the patient tumor samples of
184 at least four out of five matched pairs tested (**Figures 4C** and **S5C**), highlighting the importance of
185 performing transcriptomic and validation studies using clinical samples.

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

198 **Genes and isoforms associated with normal epithelial cells from different lineages**
199 As epithelial cells constituted the majority of the identified cells in the scRNA-seq profiles, we
200 performed cell subtyping for them using the c295 dataset as a reference to investigate their
201 subpopulations (Chen and Guestrin, 2016) (**Methods**). We identified 11 cell subtypes containing
202 three differentiation lineages from the EpiN cells, including the newly characterized *BEST4+*
203 lineage. The projection of the EpiN cells displayed a strong semantic structure in the two-
204 dimensional tSNE space, suggesting diverse differentiation lineages of the subtypes (**Figure 5A**).
205 An increased level of differentiation was observed from the center to the edge in the tSNE space.
206 For example, along the goblet lineage, cE02, cE06 and cE08 successively showed increased goblet
207 maturity. Different combinations of subtypes were detected in each normal sample and the
208 proportions were similar across multiple samples (**Figure 5B**).

209 Using monocle 3 (Qiu et al., 2017), we inferred diffusion pseudotime along the differentiation
210 trajectory of the three epithelial lineages (**Methods**) and identified 46, 31 and 38 lineage-specific
211 genes (moran's I > 0.5, adjusted p-value < 1.0×10^{-5}) for the enterocyte, goblet and *BEST4*
212 lineages, respectively, including both well-known and novel markers (**Figures 5C-E**). These
213 showed a steady increase in expression along the lineage with respect to pseudotime (**Figures 5F-**
214 **H**). We next performed differential gene and isoform expression analysis between the
215 stem/transient amplifying (TA) and differentiated subtypes along the three lineages. Overall, we
216 observed a similar number of up- and downregulated genes in the stem/TA subtypes, but mostly
217 upregulation of lineage-specific genes in the three differentiated lineages (**Figures S6A** and **S6B,**
218 **Table S3**). The genes with differential regulation were almost mutually exclusive between the three
219 differentiated subtypes (**Figure S6C**), indicating distinct transcriptomic regulatory mechanisms to
220 commit to a specific differentiation lineage.

221 By categorizing the genes in each differential gene and isoform analysis with DGE and DTS
222 (defined in **Figure 3A**), we observed a higher percentage of genes with DTSs subjected to DGE
223 compared to those without DTSs or with only one isoform (**Figure 5I**), similar to the trend observed
224 between EpiN and EpiT (**Figure 3B**). The gene *LMNA* encoding two protein isoforms, Prelamin-
225 A and Prelamin-C, through AS and alternative polyadenylation (APA), serves as a representative
226 of differential isoform usage between cell subtypes (**Figure 5J**). We observed a trend of Prelamin-
227 C to Prelamin-A switching when stem cells differentiated into enterocytes, goblet and *BEST4+* cells
228 (**Figure 5K**), consistent with the findings in basal cell carcinomas, where Prelamin-A was
229 negatively correlated to cell proliferation and appeared in the later stage of differentiation

230 (Venables et al., 2001). Overall, our analysis provides the first comprehensive characterization of
231 isoforms associated with three colon epithelial lineages.

232 **Tumor epithelial subtypes display different levels of stemness and differentiation**

233 We subtyped the EpiT cells with the same classifier above to classify the EpiT cells to their closest
234 EpiN subtypes (**Figure 6A**). Consistent with previous reports (Lee et al., 2020; Pelka et al., 2021),
235 more than 93% of tumor epithelial cells were categorized as the three stem/TA subtypes (**Figure**
236 **6B**), suggesting that malignant EpiT cells may possess stem/TA characteristics to maintain their
237 proliferation and regeneration capabilities. Based on the integrative analysis of multiple CRC
238 single-cell sequencing datasets, a recent study classified colon EpiT cells into two intrinsic
239 consensus molecular subtypes (iCMS), iCMS2 and iCMS3 (Joanito et al., 2022), and revealed that
240 over 90% of EpiT cells are from either subtype. This holds true in ours and the c295 datasets with
241 mostly one type of iCMS cells in each tumor (**Figures S7A** and **S7B**). However, the distribution of
242 the three subtypes from our categorization was more heterogenous in each tumor (**Figure 6B** and
243 **S7B**). These distinct results suggest potentially diverse and independent molecular properties of
244 the two classification methods.

245 We further identified dysregulated genes in the EpiT subtypes by comparing them to the
246 corresponding EpiN subtypes (**Figure S7C, Table S4**). Most of these were shared among the three
247 EpiT subtypes (**Figure S7D**), indicating a common underlying dysregulated mechanism of
248 tumorigenesis. In addition to DGE, we performed differential isoform analysis for the EpiT
249 subpopulations (**Table S5**). Differential usage of the *LMNA* isoforms was also observed in the EpiT
250 subtypes, with the highest usage of the Prelamin-C isoform in the most proliferative subtype, cE03,
251 and Prelamin-A isoform in cE02 (**Figure 6C**). Furthermore, we noticed that several cytokeratin
252 genes, including *KRT8* and *KRT19*, were among the most significant DTS in all three EpiT subtypes
253 (**Table S5**). The overexpression of *KRT8* has been linked to increased tumor progression and
254 invasiveness in several epithelial tumors (Fang et al., 2017; Golob-Schwarzl et al., 2019; Hendrix
255 et al., 1997). We detected three major *KRT8* isoforms, including two annotated and one novel
256 isoform (**Figure 6D**). The tumor subtypes mainly expressed the full-length isoform encoding a
257 56.6kDa protein, while the normal subtypes primarily utilized the short isoforms (**Figure 6E**).
258 Consistent with a previous study showing that the tumor suppressor *SMAD4* regulated *KRT8*
259 splicing (Stühler et al., 2006), we observed a negative correlation between *SMAD4* expression and
260 the 56.6kDa isoform usage among the epithelial subtypes (**Figure 6F**), further confirming the
261 reliability of our isoform analysis. Overall, around 20 genes with DTS were identified in each EpiT
262 subtype and consistent with those in the EpiN cells, genes with DTS were more likely to be DGE

263 (**Figure S8A**). Contrary to the DGEs, the genes with DTS barely overlapped among the three
264 subtypes (**Figure S8B**), suggesting potential dysregulation of isoform switching in the EpiT
265 subpopulations.

266 Among the top dysregulated genes in the EpiT subtypes, *REG4* and *TFF1* showed the most striking
267 upregulation in cE02 (**Figure 6G**). Interestingly, we found that *REG4* and *TFF1* expression was
268 highly correlated across the TCGA-COAD patients ($r = 0.75$, $p = 3.81 \times 10^{-71}$) and both were
269 associated with better overall survival (OS) (**Figures 6H** and **S8C**). We speculate that the tumor
270 samples with high *REG4* and *TFF1* expression may consist of a high proportion of cE02 cells,
271 leading to better patient survival. To this end, we identified the marker genes in each subtype to
272 define a signature score for the three subtypes (**Methods, Table S5**). By classifying the TCGA
273 tumors with cE02 and cE03 scores (**Figure S8D**), we found that the patient groups with both high
274 cE02 and cE03 scores, and low cE02 and cE03 scores, had the best and worst OS and progression-
275 free survival (PFS), respectively (**Figures 6I** and **S8E**), suggesting a synergistic effect of cE02 and
276 cE03 scores on patient prognosis. Similar trends were observed using an independent microarray
277 cohort curated by the CRC Subtyping Consortium (CRCSC) (Guinney et al., 2015) (**Figures S8F**
278 and **S8G**). As cE02 and cE03 have relatively lower stemness features and higher proliferative
279 activity compared to cE01, these results indicate that CRC tumors enriched with high stemness and
280 low replicative cells are associated with more frequent progression and shorter OS time.

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

283 To investigate the effect of transcript structure alterations on the proteome, we predicted the open
284 reading frames (ORFs) of all the identified isoforms. We performed a series of analyses on the
285 predicted proteome and identified 8,984 novel ORFs from NIC and NNC isoforms (**Figures S9A-**
286 **E**, Supplemental Notes Section S1). After integrating publicly available mass spectrometry (MS)
287 data (Ellis et al., 2013) from TCGA-COAD patients and in-house MS data from matched tumor
288 and normal CRC patient samples and cell line, HCT116, 26 and 44 novel proteins from the NIC
289 and NNC isoforms were supported (**Figure S9F**). There were also more MS-supported novel
290 isoforms from EpiT and tumor samples compared to EpiN and normal samples, suggesting that
291 transcriptomic dysregulation in tumor cells could result in increased abnormalities of the proteome
292 (**Figure S9G**).

293 As abnormal proteins produced in cancer cells are critical for spontaneous anti-tumor immune
294 response, they serve as a potential source for neoantigen-based cancer vaccine development to
295 leverage the immune system for cancer treatment (Sellars et al., 2022). Therefore, we

296 systematically investigated putative neoepitopes in CRC derived from tumor-specific transcripts.
297 Using the identified novel isoforms together with annotated isoforms as reference, we quantified
298 the expression of each isoform in the TCGA-COAD samples (**Figure 7A**, **Methods**). We retained
299 12 recurrent (>5 samples) novel isoforms with significantly higher expression in the TCGA-COAD
300 tumor samples and whose ORF-derived peptides were MS-supported (**Figure 7B**, **Methods**). All
301 novel 9-mers derived from the predicted ORFs of the 12 isoforms were regarded as candidate
302 neoepitopes in CRC.

303 As a limited number of epitopes can be loaded into a cancer vaccine, it is imperative to maximize
304 vaccine efficacy by selecting an optimized subset of neoepitopes with a high binding affinity to the
305 class I MHC molecules that facilitate antigen presentation. To develop cancer vaccines for a broad
306 range of patients, we considered the diversity of HLA alleles across different patients and predicted
307 the MHC binding profiles of 458 TCGA-COAD patients to the neoepitopes from the 12 recurrent
308 tumor-specific isoforms. We formulated a greedy algorithm to select two neoepitopes from each of
309 the 12 isoforms to maximize their coverage to the HLA alleles or patients, while prioritizing those
310 that overlapped with MS-supported peptides (**Table S6, Methods**, **Algorithm 1**).

311 Among the 24 neoepitopes, there was always one or more showing strong binding affinities (ES
312 rank < 0.5) to 85% of HLA-A and almost all HLA-B and HLA-C MHC molecules, with only one
313 HLA-C and eight HLA-A alleles showing weak binding affinities (0.5 < ES rank < 2.0) (**Figure**
314 **7C**). For each patient, the highest binding affinities with the three types of MHC molecules were
315 always above the strong binding threshold. All patients carry at least four strong-binding
316 neoepitopes and 39% of them have at least 12 strong-binding neoepitopes from the panel (**Figure**
317 **7D**). To better represent the global population, we generated another panel of neoepitopes (**Figure**
318 **S10A, Table S6**) from the same 12 isoforms optimized for the top 50 high-frequency alleles in
319 each HLA type (Gonzalez-Galarza et al., 2020), which similarly showed strong HLA binding
320 affinities (**Figure S10B** and **S10C**).

321 To illustrate an example from our panel, the neoepitopes selected for the tumor-specific novel
322 isoform of *STMN3* (TCONS_0078414) that utilizes a coding alternative first exon, were
323 'RQTQEAQVL' that spans the novel splice junction and overlaps with MS-detected peptides, and
324 'RAAGLAGAM' that has a strong HLA binding affinity (**Figure 7E**). The neoepitopes for *CPNE7*
325 resulting from an alternative 5' splice site were similarly selected (**Figure 7F**). The HLA binding
326 profile of the neoepitope panel showed heterogeneous affinities to the three types of HLA (**Figure**
327 **7G**). Overall, all patients have at least one strong-binding neoepitope and 65% of them with
328 corresponding isoform detection in the tumor samples. Considering the shallow coverage of MS

329 data, additional analysis using only MS-supported neoepitopes or a panel generated irrespective of
330 MS data support showed that both still exhibited high tumor-specific isoform expression and were
331 able to achieve high HLA binding affinities (**Figures S11A-G**). Furthermore, we observed strong
332 HLA binding potential in the 12 in-house patients, with all patients having at least six strong-
333 binding neoepitopes (**Figure 7H, Table S7**). For the four patients with available long-read scRNA-
334 seq data (PS017, PS018, PS025 and PS028), at least one strong-binding epitope in this panel
335 originated from the isoforms directly discovered in the LR-seq profiles. We validated the unique
336 splice junctions for four isoforms arising from AF, AL, A5SS and A3SS AS events, and the retained
337 introns in three isoforms in CRC cell lines, as well as at least two of the five matched patient
338 samples tested (**Figures 7I, S12A** and **S12B**). In line with our scRNA-seq analysis, we also
339 observed tumor-specific expression of five of these isoforms in the patient samples. Collectively,
340 these findings demonstrate a proof-of-concept for the discovery of neoepitopes based on the tumor-
341 specific transcriptome, as well as the invaluable application potential of long-read scRNA-seq in
342 the development of neoepitope-based cancer vaccines.

343 **Discussion**

344 The intestinal epithelium is the fastest self-renewing tissue in mammals, which contains epithelial
345 cells with diverse differentiation statuses, proliferation activities, and other properties and
346 functions. Despite various markers and gene expression profiles that have been characterized for
347 each epithelial cell subtype, the diversity of RNA isoforms in these subtype remains
348 underappreciated. Our study provides a full-length transcriptome across different epithelial
349 subtypes and identifies isoforms associated with enterocyte, goblet and *BEST4* lineages,
350 respectively. Although alternative RNA isoforms have been shown to regulate cell fate
351 determination of stem cells (Han et al., 2013), the prevalence and potential role of alternative
352 isoforms in intestinal epithelial differentiation are poorly understood, therefore our data may
353 provide a valuable resource to facilitate further investigations.

354 CRC is the third most common cancer type and ranks as the second leading cause of cancer death
355 around the world (Sung et al., 2021). Consistent with the first high-throughput scRNA-seq study
356 on CRC and other single-cell analyses that traced the transformation of polyps to CRC (Becker et
357 al., 2022; Li et al., 2017), we find that the vast majority of CRC tumor cells have stem-like cell
358 characteristics. However, studies for EpiT subtypes are limited. Based on the integrative analysis
359 of multiple CRC single-cell sequencing datasets, a recent study classified colon EpiT cells into two
360 subtypes and showed that 90% of EpiT in each tumor were from either subtypes (Joanito et al.,
361 2022). They further found that these two subtypes were associated with different cancer driver

362 mutations, suggesting their classification method mainly captured heterogeneities across different
363 tumors. However, our EpiT subtype classification demonstrated that each tumor comprised a
364 substantial proportion of three cell subtypes with varying differentiation status and proliferation
365 activities, thus capturing more intratumor heterogeneities. As our classification indicates that CRC
366 tumors are differentially associated with PFS and OS depending on their subtype composition, the
367 markers we propose for each subtype may serve as useful prognostic indicators. In addition to
368 epithelial cells, other stromal and immune cells may also impact patient survival and response to
369 treatment (Lee et al., 2020; Zhang et al., 2020). However, due to limited cell numbers, our study
370 could not effectively capture the full-length transcriptome in these cells, highlighting the need for
371 more advanced and robust cell isolation technologies.

372 Despite recent reports of the full transcriptome by bulk RNA LR-seq in breast and gastric cancers
373 (Huang et al., 2021; Veiga Diogo et al., 2022), single-cell LR-seq studies are rare and limited to
374 cancer cell lines (Macaulay et al., 2015; Philpott et al., 2021). Moreover, previous LR-seq studies
375 in cancer focused only on alterations in single AS events. Here, we identify hundreds of DTS arising
376 from combined AS events and calculate isoform- and cell-type-specific RNA editing levels, all of
377 which originate from primary CRC tissues and are only achievable by single-cell LR-seq.
378 Nonetheless, our analyses are limited by technological constraints leading to the relatively low
379 coverage of single-cell PacBio sequencing. Thus, the development of methods to increase the
380 capture efficiency of single-cell isoform sequencing will be beneficial for future transcriptomic
381 studies (Al'Khafaji et al., 2021).

382 Therapeutic neoantigen-based cancer vaccines aim to leverage the immune system for the treatment
383 of cancer (Sellars et al., 2022). Due to somatic mutations and transcriptomic dysregulation, tumor
384 cells produce abnormal proteins that are absent in normal cells. In the event that fragments of these
385 proteins are transported to the cell surface via the MHC I pathway, they may potentially be
386 recognized by the immune system as an antigen. We show that the full-length transcriptome in
387 tumor cells with MS data from matched patient samples can provide a comprehensive source to
388 detect aberrant mRNA-derived neoantigens. Moreover, we propose an algorithm that provides an
389 optimized combination of recurrent tumor-specific neoepitopes with high HLA binding affinities
390 for a wide spectrum of CRC patients. Recently, the phase II clinical trial for a neoantigen-based
391 mRNA vaccine by Moderna has met the endpoint (ModernaTX et al., 2019). In contrast to the
392 customized vaccine designed based on a neoepitope identified by exome sequencing of a patient
393 tumor, the panel we propose can potentially target larger CRC patient cohorts without sequencing

394 individual tumors. This could facilitate the development of neoantigen-based cancer vaccines with
395 less economic and time constraints to benefit a larger population.

396 ## Methods

397 **CRC clinical samples**

398 The protocols for the human studies comply with all relevant ethical regulations and are approved
399 by the National Healthcare Group Domain Specific Review Boards (NHG DSRB Ref:
400 2018/01032). All patients gave informed written consent. The CRC tissue samples were derived
401 from 12 patients from National University Hospital (details on the patients and samples are
402 summarized in **Table S1**).

403 **Tissue processing, library construction and single-cell sequencing**

404 Colon tissues from patients were cut into approximately 5 mm pieces and washed three times with
405 cold PBS until the supernatant became clear. The supernatant was discarded and the tissue pieces
406 were further cut into 1-2 mm pieces. 5 ml of PBS containing 5 mM EDTA (EDTA-PBS) was added
407 to the tissue fragments and incubated at 4°C for 30 minutes with gentle agitation, followed by
408 centrifugation at 200 *x*g, 4°C for 5 minutes. The supernatant was discarded and the tissue fragments
409 were vigorously resuspended in 5 ml of cold EDTA-PBS. The fragments were allowed to settle by
410 gravity and the supernatant was collected in a fresh 15 ml tube. This resuspension/ sedimentation
411 step was repeated five times. The supernatant from each round was collected in a separate 15 ml
412 tube and checked under an inverted microscope for single cells. All fractions containing single cells
413 were combined and the cells were pelleted by centrifugation at 200 *x*g, 4°C for 5 minutes, and
414 washed with cold EDTA-PBS. The single cells were resuspended in serum-free DMEM containing
415 0.05% trypsin and 5 U/ml DNaseI and incubated at 37°C with shaking for 15 minutes. The
416 dissociated cells were centrifuged at 200 *x*g, 4°C for 5 minutes, the supernatant was removed and
417 the pellet was resuspended in 2 ml of serum-free DMEM supplemented with 2 U/ml DNaseI, and
418 filtered using a 35 mm mesh. The cells were pelleted and washed once with cold PBS and
419 resuspended in PBS supplemented with 0.4% BSA, followed by cell quantification.

420 For GEM generation, 10,000 cells from each tissue sample were used to load the Chromium Next
421 GEM Chip G (10x Genomics) and Chromium controller following the manufacturer's protocol
422 (Chromium Next GEM Single Cell 3' v3.1). Post GEM-RT cleanup, cDNA amplification and
423 library construction were performed according to the manufacturer's instructions with some
424 modifications. Prior to the fragmentation step, 20 µl of cDNA from each sample were separately
425 aliquoted for PacBio long-read sequencing. To generate the Single Cell 3' Gene Expression library
426 for Illumina sequencing, 10 µl of cDNA were processed as per the 10x Genomics protocol, with an

427 extension time of 3 minutes in the sample index PCR step. The final Illumina libraries were
428 sequenced on the NovaSeq 6000 platform with S1 or S4 flowcells using the recommended
429 sequencing protocol.

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

437 **CRC cell lines**

438 Human normal colon cell line, CCD 841 CoN (ATCC: CRL-1790), was cultured in Dulbecco's
439 Modified Eagle Medium (DMEM). Human CRC cell lines, DLD-1 (Horizon Discovery: HD PAR-
440 086) and HCT116 (ATCC: CCL-247) were cultured in Roswell Park Memorial Institute (RPMI)
441 1640 Medium and DMEM, respectively. All culture media were supplemented with 10% FBS,
442 glutamine and penicillin/streptomycin. The cells were maintained at 37°C and 5% CO_2 in a
443 humidified atmosphere.

444 **RNA extraction and PCR**

445 Trizol and the PureLink RNA Mini Kit (Thermo Fisher) were used to extract total RNA from DLD-
446 1 and HCT116 cell lines. The ISOLATE II RNA Mini Kit (Bioline) was used to extract RNA from
447 CRC patient samples following the manufacturer's protocol. cDNA was generated using the High
448 Capacity cDNA Reverse Transcription Kit (Thermo Fisher). PCR experiments were performed
449 using Platinum™ Taq DNA Polymerase (Thermo Fisher). Subsequently, PCR products were
450 subjected to agarose gel electrophoresis, gel extraction and purification using the QIAquick Gel
451 Extraction Kit (Qiagen) and Sanger sequencing. Chromatograms were visualized and multiple
452 sequence alignments were performed using SnapGene (v6.2.1). RNA editing levels were quantified
453 using ImageJ (v1.50). PCR primers used are listed in **Table S8**.

454 **Mass Spectrometry**

455 To generate an in-house mass spectrometry dataset of matched tumor-normal sample pairs, <10mg
456 tissue for each tissue were processed with the iST sample preparation kit (Preomics). Samples were
457 taken up in 100 µL LYSE buffer and mixed with 50 mg of protein extraction beads (Diagenode).
458 Samples were sonicated with 10-20 cycles at 30sec ON/OFF in a Diagenode Bioruptor Plus until
459 complete tissue disruption was observed. Protein concentrations were determined using the Pierce

460 BCA Protein Assay Kit (Thermo Scientific). 20 µg of protein per sample were digested with the
461 iST kit according to the manufacturer's instructions and samples were eluted with the fractionation
462 add-on into 3 fractions. Digested peptides were quantified with the Quantitative Fluorometric
463 Peptide Assay kit (Thermo Scientific), normalized to 0.1 µg/µL and stored at -20°C prior to analysis
464 on an EASY-nLC 1200 Liquid Chromatograph (Thermo Scientific) coupled to a timsTOF fleX
465 (Bruker) mass spectrometer. For each sample, 5 µL digested peptides (500 ng) for each fraction
466 were injected and separated on an Aurora series column (25 cm length, 75 µm inner diameter, C-
467 18 1.7 µm; IonOpticks) with an integrated captive spray emitter. The column was mounted on a
468 captive spray ionization source and temperature controlled by a column oven (Sonation) at 50°C.
469 A 105-min gradient from 2 to 40 % (v/v) acetonitrile in 0.1 % (v/v) formic acid at a flow of 400
470 nL/min was used. Spray voltage was set to 1.65 kV. The timsTOF fleX was operated with data-
471 dependent acquisition (DDA) in PASEF mode with 10 PASEF ramps per topN acquisition cycle
472 (cycle time 1.17s) and a target intensity of 10,000. Singly charged precursor ions were excluded
473 based on their position in the m/z-ion mobility plane and precursor ions that reached the target
474 intensity were dynamically excluded for 24 seconds.

475 **Summary of bioinformatic analysis**

476 The bioinformatic analysis workflow is summarized in **Figure S13**. Additionally, the command
477 lines and parameters used in the analysis are also provided (Supplemental Notes Section S2).

478 **Illumina short-read scRNA-seq data pre-processing**

479 CellRanger v4.0.0 (https://support.10xgenomics.com/single-cell-gene-
480 expression/software/downloads/latest) was used for the pre-processing of 10x droplet-based
481 scRNA-seq raw data. Specifically, 'cellranger count' with default parameters was used to align
482 sequencing reads to the human reference genome (GRCh38). The gene expression count matrix
483 was then obtained based on the genomic coordinates defined by the GENCODE annotation (release
484 v37 for GRCh38). Subsequently, 'cellranger aggr' was performed to aggregate the count matrices
485 of samples in multiple 'cellranger count' runs. The aggregated count matrix of all samples was then
486 loaded into Seurat (v3.2.3) (Stuart et al., 2019) for subsequent analysis. A cell was filtered out if it
487 satisfied any of the following criteria: (1) expression of fewer than 250 genes, (2) detected fewer
488 than 300 UMIs, (3) more than 70% of the UMIs mapped to mitochondrial genes. All mitochondrial
489 genes were discarded in subsequent analysis. This analysis resulted in a total of 19,015 high-quality
490 cells from 22 tissue samples.

491 The gene expression count matrix was log-normalized using Seurat's 'NormalizeData' function.
492 The top 2000 genes with the highest variations were selected using the 'SelectIntegrationFeatures'

493 subroutine of Seurat, and the 'IntegrateData' subroutine was used to remove batch effects across
494 the samples. Integration anchors were chosen based on the first thirty dimensions of the canonical
495 correlation analysis (CCA) (Stuart et al., 2019). We performed the analysis with above procedures
496 separately on the normal and tumor samples to prevent the batch effect correction procedure from
497 eliminating differences between the normal and tumor cells (**Figure S1**).

498 **Cell type identification**

499 We integrated the in-house data to a reference dataset, Human Colon Cancer Atlas (c295) (Pelka
500 et al., 2021)from the Broad Institute Single Cell Portal (https://singlecell.broadinstitute.org/), for
501 more robust and reliable identification of the cell types. The c295 atlas contains gene expression
502 profiles of 371,223 cells. Due to memory constraint, a subset of the atlas containing 50% of the
503 cells in the c295 atlas was used as the reference based on a stratified sampling procedure across
504 different cell types. Transfer anchors were chosen by the 'FindTransferAnchors' subroutine of
505 Seurat using the first thirty dimensions of the principal component analysis (PCA). The
506 'TransferData' subroutine was used to transfer the c295's 'ClusterMidway' annotation to the in-
507 house data via the selected anchors. The 'ClusterMidway' labels of the c295 atlas annotate cells
508 into 20 different types, including (with boldface indicating the availability of the cell type in the in-
509 house data):

510 - 2 epithelial cell types: **normal epithelial cells (EpiN)** and **tumor epithelial cells (EpiT)**.
511 - 13 immune cell types: **B cells, plasma cells, CD4+ T cells, CD8+ T cells, gamma-delta**
512 **T cells, PLZF+ T cells, natural killer cells, innate lymphoid cells, dendritic cells,**
513 **granulocytes, macrophages, mast cells, monocytes**,
514 - 5 stromal cell types: **fibroblasts**, endothelial cells, pericytes, smooth muscle cells,
515 Schwann cells

516 Several post-processing steps were performed for the result from the above classification: (1) We
517 corrected all the EpiN misclassified as EpiT cells in the normal samples. (2) To improve the purity
518 of the identified epithelial tumor cells, we trained a gradient boosting classifier (xgboost v1.5.0
519 (Chen and Guestrin, 2016)) using the combined EpiN and EpiT cells from the c295 atlas and the
520 in-house dataset. Following a similar 5-fold cross-validation scheme as in (Pelka et al., 2021), we
521 divided the combined dataset into five independent splits and used four folds for training and the
522 rest one fold for testing. The classifier was trained to perform the binary classification of whether
523 a cell is a EpiT or a EpiN cell. Only when a cell came from a tumor sample and with >75%
524 prediction probability as EpiT cells in the test split did we regard it as a candidate for the third step.
525 (3) For all candidates from step (2), we inferred somatic copy number variations (SCNV) of the

526 cells based on their gene expression profiles using inferCNV (v1.11.1,
527 https://github.com/broadinstitute/infercnv) (The Trinity CTAT Project, 2019). The final set of
528 tumor epithelial cells was kept as those having >10% genomic SCNV regions. The epithelial cells
529 from tumor samples that did not pass either the gradient boosting criteria or the SCNV criteria were
530 deemed as undetermined epithelial cells.

531 **Dimensionality reduction and visualization**

532 PCA was performed on the log-normalized and centered gene expression matrix for dimensionality
533 reduction. t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008) was
534 then performed to generate 2D embeddings for visualization. tSNE was supplied with the first 30
535 components of PCA and run with a perplexity value of 30.

536 **PacBio long-read scRNA-seq data pre-processing**

537 SMRT Link (v9.0.0, https://www.pacb.com/support/software-downloads/) was utilized to pre-
538 process PacBio Iso-seq long-read scRNA-seq data. We used the Isoseq-deduplication pipeline
539 (Pacific Biosciences, 2018) recommended by the official repository of PacBio
540 (https://github.com/PacificBiosciences/IsoSeq/) for the processing of raw data from the PacBio
541 Sequel II system. The 'ccs' module of SMRT Link was applied to generate consensus reads (CCS
542 Reads) from the BAM file that contained the PacBio subreads. The reads with less than 90%
543 accuracy ('--min-rq 0.9') were discarded. The 'lima' module was utilized in orientation
544 determination and primer removal of the reads. The cell barcodes and UMIs of each read were
545 extracted by 'isoseq3 tag' using the library design 'T-12U-16B'. The processed reads were trimmed
546 of the Poly(A) tail and potential concatemers were removed by the command 'isoseq3 refine'. The
547 resulting full-length non-concatemer (FLNC) reads were clustered together according to their cell
548 barcode and UMI with 'isoseq3 dedup' to deduplicate and generate a consensus sequence of each
549 molecule. Those consensus sequences were then mapped to the human reference genome
550 (GRCh38) using minimap2 (Li, 2018). According to the genomic coordinates of the mapped
551 molecules, the structures of each isoform represented in the gtf format were obtained by utilizing
552 the script 'collapse_isoforms_by_sam.py' from the cDNA cupcake toolkit (Elizabeth Tseng, 2017)
553 (v27.0.0, https://github.com/Magdoll/cDNA_Cupcake).

554 **Isoform quality control, filtering and classification**

555 We merged the isoforms identified across different samples with gffcompare (Pertea and Pertea,
556 2020). We then utilized SQANTI3 (v3.3) (Tardaguila et al., 2018) to compare the identified
557 isoforms against the reference transcriptome annotation, which produced quality control and
558 classification information for each isoform. To provide a comprehensive reference for SQANTI3,

559 we merged databases from multiple sources, including GENCODE (Release v37 for GRCh38) and
560 RefSeq (NCBI Homo sapiens Annotation Release 109 for GRCh38), for the annotation of isoform
561 structures. With this assembled reference, SQANTI3 classified each isoform into FSM, ISM, NIC,
562 NNC and several other structural types (Tardaguila et al., 2018). The rule-based filter functionality
563 of SQANTI3 was used to filter out the isoforms that were likely to be sequencing artifacts such as
564 intra-priming, reverse transcriptase template switching (RTS), and non-canonical splice sites with
565 low read coverage. The remaining isoforms and their associated barcodes are assembled into an
566 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
567 N_{cells} is the number of all cells associated with those isoforms. Finally, we searched the barcodes
568 from PacBio data against the cell barcodes detected in the same sample in the 10x Illumina
569 sequencing data to calculate the expression (UMI counts) of each isoform in each cell. To evaluate
570 the quality of the identified isoforms, we examined whether the 5' and 3' ends of each isoform are
571 supported by the TSS and PAS reference. The TSS reference was a union of TSSs from
572 GENCODE, RefSeq and refTSS (v3.1) (Abugessaisa et al., 2019), while the PAS reference
573 contains all transcript 3' ends in GENCODE, RefSeq and PAS from PolyASite (v2.0) (Herrmann
574 et al., 2020) and PolyADB (v3.2) (Wang et al., 2018). Furthermore, splice junctions from each
575 identified isoform were overlapped with junction reads extracted from the RNA-seq data in The
576 Cancer Genome Atlas colon adenocarcinoma cohort (TCGA-COAD).

577 **Splice event extraction and analysis**

578 For all isoforms that had passed the SQANTI3's rule-based filter, we utilized SUPPA (v2.3)
579 (Trincado et al., 2018) for the extraction of seven types of AS events including alternative 5' splice
580 sites (A5), alternative 3' splice sites (A3), alternative first exon (AF), alternative last exon (AL),
581 skipped exon (SE), mutually exclusive exon (MX) and retained intron (RI). SUPPA analyzed the
582 gtf file obtained from the isoform calling and quality control pipeline to produce seven event files
583 (the 'ioe' files) each containing one of the above event types based on the pairwise comparison of
584 the input isoform structures. Taking into consideration the accuracy of splice sites and the
585 inaccuracy of TSS and TTS, we required SUPPA2 to use stringent boundaries for the splice sites
586 of the seven event types, but flexible boundaries (with a variability of 48nt) for the TSS and TTS
587 coordinates of AF and AL.

588 **Identification of dysregulated gene expression and transcript structures**

589 Dysregulated gene expression (DGE) was identified by comparing the expression of genes (log-
590 normalized UMI) in cells from one group (cell type or subtype) to that from another group. In brief,
591 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.

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 (Liao et al., 2019) 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 (Ramaswami and Li, 2014), (2) the DARNED database (Kiran et al., 2012) and (3) Tan et al., 2017. (Tan et al., 2017). 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 (Danecek et al., 2021). 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

625 identified to a specific isoform and if the per isoform read coverage is greater than five. The per
626 isoform edit ratio is defined as the number of edited reads divided by the number of total read
627 coverage from the isoform (**Figure 4A**).

628 **Lineage and trajectory analysis**

629 We employed monocle 3 (Qiu et al., 2017) for the single-cell lineage and trajectory analysis on the
630 epithelial component of the short-read scRNA-seq dataset. The principal graph was constructed
631 using the 'learn_graph' subroutine of monocle 3 with default parameters using the tSNE
632 coordinates and the nodes corresponding to the four stem cells subclusters were designated as the
633 root for diffusion pseudotime inference. The epithelial cells that were also detected in the PacBio
634 long-read scRNA-seq were associated with their isoform level expression values. The expression
635 of genes and isoforms, as well as the percentage of isoforms that were associated with a specific
636 lineage, were identified based on Moran's I statistic (Moran, 1950).

637 **Quantification of isoforms in TCGA samples**

638 Isoform-level expression of the TCGA-COAD bulk samples was quantified using stringtie (v2.2.1)
639 (Pertea et al., 2016) using the merged transcriptomic reference of novel identified isoforms,
640 GENCODE (version 37) and RefSeq (NCBI Homo sapiens Annotation Release 109 for GRCh38).

641 **Estimation of the signature scores of EpiT subtypes in bulk samples**

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

646 **Survival analysis**

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

651 **Proteomic analysis of isoforms**

652 To comprehensively characterize the proteome in CRC, GeneMarkS-T (Tang et al., 2015) was used
653 to predict ORFs in the PacBio-identified isoforms. For comparison and annotation of the predicted
654 ORFs, we collected the human reference proteome from Uniprot (release 2022-02) (2021),
655 including all sequences from the reviewed (SwissProt), unreviewed (TrEmbl) and spliced isoform
656 (Varsplice) collections at protein existence (PE) level PE=1 and PE=2. Using blastp (Camacho et
657 al., 2009), we searched each predicted ORF sequence against the UniProt reference proteome, and

658 the reference sequence with the highest alignment score was assigned to it as its nearest homolog
659 in the database. The ORFs with a similarity of less than 99% to their nearest homolog in the
660 reference were considered novel protein sequences. Protein domains of each ORF sequence, as
661 defined by the Pfam-A hidden Markov model (HMM)-based multiple sequence alignments
662 (Bateman et al., 2004), were annotated by hmmscan (http://hmmer.org/). We regarded a novel ORF
663 sequence as having a domain gain if it contained a proper superset of the Pfam domains from its
664 nearest Uniprot homolog, or a domain loss if it contained only a proper subset of them. We also
665 inferred the sensitivity of the PacBio-identified isoforms to nonsense-mediated decay (NMD) based
666 on the relative position of the predicted stop codon from the last splice junction. Lastly, the
667 subcellular localization of the ORF-derived proteins was predicted using DeepLoc (Almagro
668 Armenteros et al., 2017).

669 **Mass spectrometry data analysis**

670 To validate the predicted ORFs, particularly the novel ORFs, we collected mass spectrometry (MS)
671 data from three independent sources: (1) a CRC cohort from the National Cancer Institute's Clinical
672 Proteomic Tumor Analysis Consortium (CPTAC), (2) an in-house MS dataset consisting of 22
673 samples from 11 CRC patients (11 tumor and 11 paired adjacent normal), and (3) an in-house MS
674 data from a colon cancer cell line, HCT116. With the same pipeline as a previous study (Bulik-
675 Sullivan et al., 2019), we searched MS2 spectra from the MS data against a reference by integrating
676 annotated protein sequence from UniProt (release 2022-02) and novel ORFs predicted based on
677 isoforms derived from LR-seq using Comet (Eng et al., 2013). Five peptides were reported for each
678 spectrum query and these peptides were further ranked and scored for final identification using
679 Percolator (Käll et al., 2007). The validated novel ORFs were defined by its inclusion of at least
680 one novel peptide (FDR < 0.01) in the MS data that does not match with any protein sequence from
681 the UniProt database.

682 **Proposing putative neoepitopes for cancer vaccine development**

683 To obtain a shortlist of tumor-specific novel transcript isoforms, we extracted the NIC and NNC
684 isoforms identified in our long-read scRNA-seq profiles and quantified their expression levels in
685 the TCGA-COAD bulk samples along with the known isoforms in GENCODE and RefSeq. Tumor-
686 specificity was ensured by filtering for isoforms discovered only from the EpiT cells and not
687 expressed (TPM < 0.5) in the TCGA-COAD normal samples. We further selected the isoforms that
688 were recurrent across multiple patients (>5 samples), which resulted in a shortlist of 116 novel
689 coding isoforms. To ensure active translation of the isoforms, we prioritized 12 isoforms from the

690 shortlist whose ORF subsequences were supported at least once by the MS data from tumor samples
691 but without any peptides detected in MS data from normal samples.

692 To derive the set of tumor-specific neoepitopes from the tumor-specific novel transcript isoforms,
693 we assembled a collection of all the *k*-mers ($k = 9$) that are present in the predicted ORF sequences
694 and further curated a subset containing tumor-specific novel *k*-mers by requiring: (1) the *k*-mers
695 are not in the Uniprot reference, (2) the *k*-mers are only present in ORFs in transcript isoforms from
696 EpiT but not from EpiN or other cell types, (3) the *k*-mers are from the transcript isoforms that are
697 expressed (TPM > 0.1) in more than five TCGA-COAD tumor samples and not expressed (TPM <
698 0.1) in the normal samples (4) the *k*-mers are from the ORFs in the transcript isoforms classified as
699 NIC and NNC. In this way, we obtained a set $\mathcal{S}$ of candidate neoepitope *k*-mers that are from
700 recurrent tumor-specific isoforms across tumor samples.

701 For a given population, we predicted the binding affinity of each neoepitope *k*-mer against the
702 frequent HLA alleles in the population (represented as the set $\mathcal{H}$) using netMHCpan (v4.1) (Jurtz
703 et al., 2017). We used the rank of elution score, EL_rank < 2 as the threshold of binding/non-
704 binding for a *k*-mer to a specific HLA allele. In this way, we obtained a mapping $\mathcal{M}$ that
705 represented the HLA alleles that a neoepitope *k*-mer can bind to. For example, $\mathcal{M}[s]$ ($\mathcal{M}[s] \subseteq \mathcal{H}$)
706 contains the HLA alleles to which neoepitope s ($s \in \mathcal{S}$) binds.

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

711 Input: The size of the shortlist n and the sets of HLA alleles covered by each epitope
712 $\{M[s_1], M[s_2], \dots, M[s_{|\mathcal{S}|}]\}$.

713 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.

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

716 Input: The size of the shortlist n, the sets of HLA alleles covered by each epitope
717 $\{M[s_1], M[s_2], \dots, M[s_{|\mathcal{S}|}]\}$, and the weight of each HLA allele a ($a \in \mathcal{H}$), $\mathcal{W}[a]$.

718 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 (Nemhauser et al., 1978), 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. (Thorsson et al., 2018) 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'
    $S' \leftarrow \emptyset$
    $c' \leftarrow 0$
    $\mathcal{U} \leftarrow \mathcal{H}$ // the set of uncovered HLA alleles
    for $i = 1$ to n **do**
      $s_i \leftarrow$ SELECT-BEST-KMER ($\mathcal{S}$, $\mathcal{M}$, $\mathcal{U}$, $\mathcal{W}$)
      $S' \leftarrow S' \cup \{s_i\}$, $S \leftarrow S \backslash \{s_i\}$
      $\mathcal{U} \leftarrow \mathcal{U} \backslash \mathcal{M}(s_i)$ **if** $\mathcal{U} \backslash \mathcal{M}(s_i) \neq \emptyset$ **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 ($\mathcal{S}$, $\mathcal{M}$, $\mathcal{P}$, n)
inputs:

$\mathcal{S}$, the set of neoepitope k-mers from recurrent tumor-specific isoforms
$\mathcal{M}$, the mapping from k-mers to the patients whose HLA they bind to
$\mathcal{P}$, 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'
$S' \leftarrow \emptyset$
$c' \leftarrow 0$
$\mathcal{U} \leftarrow \mathcal{P}$ // the set of uncovered patients
for $i = 1$ to n **do**
$s_i \leftarrow$ SELECT-BEST-KMER ($\mathcal{S}$, $\mathcal{M}$, $\mathcal{U}$, $\{p: 1$ **foreach** p **in** $\mathcal{P}\}$)
$S' \leftarrow S' \cup \{s_i\}$, $S \leftarrow S \backslash \{s_i\}$
$\mathcal{U} \leftarrow \mathcal{U} \backslash \mathcal{M}(s_i)$ **if** $\mathcal{U} \backslash \mathcal{M}(s_i) \neq \emptyset$ **else** $\mathcal{U} \leftarrow \mathcal{H}$ // reset uncovered patients if all covered
end
$c' =$ SIZE(S')
return S', c'

---

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

Acknowledgments

We thank all past and present YT and XG 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 from 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 whole or 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 FCC/1/1976-44-01, FCC/1/1976-45-01 and REI/1/5234-01-01.

Authors' Contributions

Z.L. and B.Z. performed computational analysis. J.J.C. performed experiments, single cell isolation and library preparations, as well as analyzed data. H.T., Q.Y.T. and X.H.C. performed single cell isolation, RNA extractions and tissue sample preparation for mass spectrometry. P.D. constructed PacBio libraries and performed sequencing. X.F. assisted with computational analysis. C.C. and D.K. performed the mass spectrometry acquisition. F.C. assisted in mass spectrometry sample preparation. S.W., B.E.S., I.J.W.T., K.Y.L., B. L., W.K.C. and K.K.T provided COAD clinical samples. Z.L, B.Z., J.J.C., X.G. and Y.T. designed the study and prepared the manuscript. All authors reviewed and commented on the manuscript.

Data and Code Availability

Raw sequencing data used in this study are deposited in European Nucleotide Archive under session number PRJEB61296. 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 .

Declaration of Interests

The authors have declared no competing interests.

767 **Figure legends**

768 **Figure 1 The short-read single-cell transcriptomic atlas of human CRC.**
769 (A) Schematic illustration of the workflow of long-read (PacBio) and short-read (Illumina) single-
770 cell RNA sequencing (scRNA-seq) on tumor and normal samples from 12 CRC patients.

771 (B and C) tSNE (t-distributed stochastic neighbor embedding) plots illustrating (B) the major cell
772 types and (C) the source of cells from in-house tumor and normal samples, and the public dataset
773 (c295).

774 (D) Number of detected cells for each major cell type by in-house Illumina scRNA-seq data.

775 (E) Proportion of detected cells from each major cell type in each sample.

776 (F) Copy number variation (CNV) profiling of each cell from each sample.

777 **Figure 2 The long-read single-cell transcriptomic atlas of human CRC.**
778 (A) Schematic showing four structural types of isoforms identified by PacBio scRNA-seq data.
779 FSM: full splice match, ISM: incomplete splice match, NIC: novel in catalog, NNC: novel not in
780 catalog. The black box denotes a novel splice site.

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

783 (C) Percentage of the identified isoforms with supporting Cap Analysis of Gene Expression
784 (CAGE) peaks for the 5' ends and PolyA peaks for the 3' ends.

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

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

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

790 **Figure 3 Dysregulated transcript structures in epithelial tumor cells.**
791 (A) Illustration of dysregulated gene expression (DGE) and dysregulated transcript structure (DTS)
792 in epithelial tumor (EpiT) compared to normal (EpiN) cells.

793 (B) Proportion of upregulated (up), downregulated (down) and not significantly changed (NS)
794 genes for those with DTS, without DTS and those with only one detected isoform. P-value from
795 Chi-squared test.

796 (C) Enrichment of Gene Ontology (GO) terms for genes with DTS.

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

799 (E) The number of significantly up- and downregulated isoforms and their overlap with isoforms
800 with increased, decreased and not significantly changed (NS) percentages in EpiT compared to
801 EpiN.

802 (F) Numbers of DTS isoforms with co-occurrence of two types of AS events (or single type of
803 events on the diagonal) and their percentage of the total corresponding isoforms. P-value from two-
804 tailed binomial test. AF: alternative first exon, SE: skipped exon, MX: mutually exclusive exon,
805 A5: alternative 5' splice site, A3: alternative 3' splice site, RI: retained intron, AL: alternative last
806 exon.

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

810 **Figure 4 Dysregulated RNA editing in epithelial tumor cells.**
811 (A) Illustration of two approaches to calculate RNA editing levels: RNA editing level per site
812 (REPS) and RNA editing level per isoform (REPI).

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

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

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

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

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

823 (G) Heatmap showing the number of isoforms with detected RNA editing versus the total number
824 of detected isoforms of the gene. Each dot represents an editing event at a gene locus. Dots are

825 colored according to their REPS and those for the three RNA editing events in **Figure 4E** are
826 highlighted.

827 **Figure 5 Transcriptome profiling of normal epithelial cell subtypes from multiple lineages.**
828 (A) tSNE (t-distributed stochastic neighbor embedding) plot illustrating subtypes of epithelial
829 normal cells (EpiN) and the three main lineages of differentiation (indicated by arrows), including
830 enterocytes (green), goblet (blue) and *BEST4* (purple).

831 (B) Proportion of EpiN subtypes in each sample.

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

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

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

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

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

843 **Figure 6 Dysregulation of genes and isoforms in the epithelial tumor cell subtypes.**
844 (A) Schematic showing the comparisons between each EpiT subtype and their corresponding EpiN
845 cell subtype.

846 (B) Proportion of EpiT subtypes in each sample.

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

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

851 (E) Proportion of the *KRT8* transcripts encoding each CK8 protein isoform in each EpiN and EpiT
852 subtype.

853 (F) Correlation between the usage of the 56.6kDa transcript isoform and *SMAD4* expression in each
854 EpiN and EpiT subtype.

855 (G) Fold changes (log2 transformed) of top significant genes with dysregulated expression in each
856 EpiT subtype compared to the corresponding EpiN subtype.

857 (H) Overall survival of TCGA-COAD patients with different expression levels of *REG4*.

858 (I) Progression-free survival of TCGA-COAD patients with different scores of cE02 and cE03
859 signature genes. H: high, L: low.

860 **Figure 7 Identification of common neoantigens from recurrent tumor-specific transcripts.**
861 (A) Workflow for the identification of neoepitopes from novel tumor-specific recurrent transcript
862 isoforms for cancer vaccine development.

863 (B) Estimated maximum expression level of candidate isoforms (MS-supported) in the normal and
864 tumor TCGA-COAD samples.

865 (C) HLA binding profile of the selected neoepitopes. Each dot represents the best binding affinity
866 of MHC molecule(s) from an HLA allele (columns 1-3) or a TCGA-COAD patient (column 4)
867 against the panel of neoepitopes.

868 (D) Percentage of HLA alleles or patients with at least n (indicated by the x-axis) strong-binding
869 neoepitopes in the panel.

870 (E-F) Selection of neoepitopes for a novel isoform of (E) *STMN3* and (F) *CPNE7*. The selected
871 MS-supported neoepitopes are encircled by solid boxes and the non-MS-supported ones by dashed
872 line boxes.

873 (G) HLA binding profile of the panel of 24 neoepitopes against the HLA alleles (HLA-A, HLA-B,
874 HLA-C) among the TCGA-COAD patients. The right panel summarizes the number of patients
875 with at least one HLA allele showing 'weak/strong binding' affinities to the neoepitope of the
876 corresponding row. 'Strong binding & detected by RNA-seq' denotes neoepitopes derived from
877 isoforms with TPM>0.5 (as in **Figure 7B**). The last row summarizes the percentage of TCGA-
878 COAD patients with at least one neoepitope satisfying the two criteria above.

879 (H) The binding affinity of the 24 neoepitopes for the 12 in-house patients. 'Weak/strong binding'
880 denotes the binding affinity of the neoepitopes to at least one patient HLA allele (each row). 'Strong
881 binding & detected by LR-seq' neoepitopes are indicated by red boxes in the corresponding
882 heatmap. The top panel summarizes the total number of neoepitopes that satisfies each criterion

883 above per patient. The right panel summarizes the total number of patients for whom the epitope
884 satisfies the above criteria.

885 (I) PCR and Sanger sequencing validation of the unique splice junctions for the recurrent tumor-
886 specific isoforms from which the neoepitope panel is derived. Splice junctions from AF, AL, A5SS
887 and A3SS transcripts are depicted by black dotted lines and the 5' junction of RIs by red dotted
888 lines. Representative chromatograms and images are shown.

References

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.

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, 2021.2010.2001.462818.

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.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths‐Jones, S., Khanna, A., Marshall, M., Moxon, S., and Sonnhammer, E.L. (2004). The Pfam protein families database. Nucleic acids research *32*, D138-D141.

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. Nature Genetics *54*, 985-995.

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.* (2019). Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. Nature Biotechnology *37*, 55-63.

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.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC bioinformatics *10*, 1-9.

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. Nature Cell Biology *24*, 928-939.

Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

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

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.

Dietrich, D.R. (1993). Toxicological and pathological applications of proliferating cell nuclear antigen (PCNA), a novel endogenous marker for cell proliferation. Critical reviews in toxicology *23*, 77-109.

Ding, F., Su, C.J., Edmonds, K.K., Liang, G., and Elowitz, M.B. (2022). Dynamics and functional roles of splicing factor autoregulation. Cell Reports *39*, 110985.

Dong, X., Chen, G., Cai, Z., Li, Z., Qiu, L., Xu, H., Yuan, Y., Liu, X., and Liu, J. (2018). CDK13 RNA Over-Editing Mediated by ADAR1 Associates with Poor Prognosis of Hepatocellular Carcinoma Patients. Cellular Physiology and Biochemistry *47*, 2602-2612.

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.

Elizabeth Tseng (2017). cDNA Cupcake.

Ellis, M.J., Gillette, M., Carr, S.A., Paulovich, A.G., Smith, R.D., Rodland, K.K., Townsend, R.R., Kinsinger, C., Mesri, M., and Rodriguez, H. (2013). Connecting genomic alterations to cancer

938 biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. Cancer
939 discovery *3*, 1108-1112.
940 Eng, J.K., Jahan, T.A., and Hoopmann, M.R. (2013). Comet: An open-source MS/MS sequence
941 database search tool. PROTEOMICS *13*, 22-24.
942 Fang, J., Wang, H., Liu, Y., Ding, F., Ni, Y., and Shao, S. (2017). High KRT 8 expression promotes
943 tumor progression and metastasis of gastric cancer. Cancer science *108*, 178-186.
944 Galeano, F., Tomaselli, S., Locatelli, F., and Gallo, A. (2012). A-to-I RNA editing: the "ADAR"
945 side of human cancer. Paper presented at: Seminars in cell & developmental biology (Elsevier).
946 Golob-Schwarzl, N., Bettermann, K., Mehta, A.K., Kessler, S.M., Unterluggauer, J., Krassnig, S.,
947 Kojima, K., Chen, X., Hoshida, Y., Bardeesy, N.M., *et al.* (2019). High Keratin 8/18 Ratio Predicts
948 Aggressive Hepatocellular Cancer Phenotype. Translational Oncology *12*, 256-268.
949 Gonzalez-Galarza, F.F., McCabe, A., Santos, E.J.M.d., Jones, J., Takeshita, L., Ortega-Rivera,
950 N.D., Cid-Pavon, G.M.D., Ramsbottom, K., Ghattaoraya, G., and Alfirevic, A. (2020). Allele
951 frequency net database (AFND) 2020 update: gold-standard data classification, open access
952 genotype data and new query tools. Nucleic acids research *48*, D783-D788.
953 Gu, T., Buaas, F.W., Simons, A.K., Ackert-Bicknell, C.L., Braun, R.E., and Hibbs, M.A. (2012).
954 Canonical A-to-I and C-to-U RNA Editing Is Enriched at 3′ UTRs and microRNA Target Sites in
955 Multiple Mouse Tissues. PLOS ONE *7*, e33720.
956 Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L.,
957 Roepman, P., Nyamundanda, G., Angelino, P., *et al.* (2015). The consensus molecular subtypes of
958 colorectal cancer. Nature Medicine *21*, 1350-1356.
959 Guo, J., Jia, J., and Jia, R. (2015). PTBP1 and PTBP2 impaired autoregulation of SRSF3 in cancer
960 cells. Scientific Reports *5*, 14548.
961 Han, H., Irimia, M., Ross, P.J., Sung, H.-K., Alipanahi, B., David, L., Golipour, A., Gabut, M.,
962 Michael, I.P., Nachman, E.N., *et al.* (2013). MBNL proteins repress ES-cell-specific alternative
963 splicing and reprogramming. Nature *498*, 241-245.
964 Hendrix, M., Seftor, E.A., Seftor, R., and Trevor, K.T. (1997). Experimental co-expression of
965 vimentin and keratin intermediate filaments in human breast cancer cells results in phenotypic
966 interconversion and increased invasive behavior. The American journal of pathology *150*, 483.
967 Herrmann, C.J., Schmidt, R., Kanitz, A., Artimo, P., Gruber, A.J., and Zavolan, M. (2020).
968 PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. Nucleic
969 Acids Research *48*, D174-D179.
970 Huang, K.K., Huang, J., Wu, J.K.L., Lee, M., Tay, S.T., Kumar, V., Ramnarayanan, K.,
971 Padmanabhan, N., Xu, C., and Tan, A.L.K. (2021). Long-read transcriptome sequencing reveals
972 abundant promoter diversity in distinct molecular subtypes of gastric cancer. Genome biology *22*,
973 1-24.
974 Joanito, I., Wirapati, P., Zhao, N., Nawaz, Z., Yeo, G., Lee, F., Eng, C.L.P., Macalinao, D.C.,
975 Kahraman, M., Srinivasan, H., *et al.* (2022). Single-cell and bulk transcriptome sequencing
976 identifies two epithelial tumor cell states and refines the consensus molecular classification of
977 colorectal cancer. Nature Genetics *54*, 963-975.
978 Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-
979 4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide
980 binding affinity data. The Journal of Immunology *199*, 3360-3368.
981 Kahles, A., Lehmann, K.-V., Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O.,
982 Kohlbacher, O., Sander, C., Caesar-Johnson, S.J., *et al.* (2018). Comprehensive Analysis of
983 Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell *34*, 211-224.e216.
984 Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. (2007). Semi-supervised
985 learning for peptide identification from shotgun proteomics datasets. Nature Methods *4*, 923-925.

986 Keskin, D.B., Anandappa, A.J., Sun, J., Tirosh, I., Mathewson, N.D., Li, S., Oliveira, G., Giobbie-
987 Hurder, A., Felt, K., Gjini, E., *et al.* (2019). Neoantigen vaccine generates intratumoral T cell
988 responses in phase Ib glioblastoma trial. Nature *565*, 234-239.
989 Kiran, A.M., O'Mahony, J.J., Sanjeev, K., and Baranov, P.V. (2012). Darned in 2013: inclusion of
990 model organisms and linking with Wikipedia. Nucleic acids research *41*, D258-D261.
991 Lee, H.-O., Hong, Y., Etlioglu, H.E., Cho, Y.B., Pomella, V., Van den Bosch, B., Vanhecke, J.,
992 Verbandt, S., Hong, H., Min, J.-W., *et al.* (2020). Lineage-dependent gene expression programs
993 influence the immune landscape of colorectal cancer. Nature Genetics *52*, 594-603.
994 Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-
995 3100.
996 Li, H., Courtois, E.T., Sengupta, D., Tan, Y., Chen, K.H., Goh, J.J.L., Kong, S.L., Chua, C., Hon,
997 L.K., Tan, W.S., *et al.* (2017). Reference component analysis of single-cell transcriptomes
998 elucidates cellular heterogeneity in human colorectal tumors. Nature Genetics *49*, 708-718.
999 Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis
1000 toolkit with revamped UIs and APIs. Nucleic acids research *47*, W199-W205.
1001 Maaten, L.v.d., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning
1002 research *9*, 2579-2605.
1003 Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N.,
1004 Coupland, P., Shirley, L.M., *et al.* (2015). G&T-seq: parallel sequencing of single-cell genomes
1005 and transcriptomes. Nature Methods *12*, 519-522.
1006 Meacham, C.E., and Morrison, S.J. (2013). Tumour heterogeneity and cancer cell plasticity. Nature
1007 *501*, 328-337.
1008 ModernaTX, I., Sharp, M., and LLC, D. (2019). An Efficacy Study of Adjuvant Treatment With
1009 the Personalized Cancer Vaccine mRNA-4157 and Pembrolizumab in Participants With High-Risk
1010 Melanoma (KEYNOTE-942) (https://ClinicalTrials.gov/show/NCT03897881).
1011 Moran, P.A. (1950). Notes on continuous stochastic phenomena. Biometrika *37*, 17-23.
1012 Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., Kovar,
1013 C.L., Lewis, L.R., Morgan, M.B., Newsham, I.F., *et al.* (2012). Comprehensive molecular
1014 characterization of human colon and rectal cancer. Nature *487*, 330-337.
1015 Nemhauser, G.L., Wolsey, L.A., and Fisher, M.L. (1978). An analysis of approximations for
1016 maximizing submodular set functions—I. Mathematical programming *14*, 265-294.
1017 Nishikura, K. (2010). Functions and Regulation of RNA Editing by ADAR Deaminases. Annual
1018 Review of Biochemistry *79*, 321-349.
1019 Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-
1020 Hurder, A., Peter, L., *et al.* (2017). An immunogenic personal neoantigen vaccine for patients with
1021 melanoma. Nature *547*, 217-221.
1022 Pacific Biosciences (2018). IsoSeq v3.
1023 Pelka, K., Hofree, M., Chen, J.H., Sarkizova, S., Pirl, J.D., Jorgji, V., Bejnood, A., Dionne, D., Ge,
1024 W.H., Xu, K.H., *et al.* (2021). Spatially organized multicellular immune hubs in human colorectal
1025 cancer. Cell.
1026 Pertea, G., and Pertea, M. (2020). GFF utilities: GffRead and GffCompare. F1000Research *9*.
1027 Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level
1028 expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature
1029 protocols *11*, 1650.
1030 Philpott, M., Watson, J., Thakurta, A., Brown, T., Brown, T., Oppermann, U., and Cribbs, A.P.
1031 (2021). Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. Nature
1032 Biotechnology *39*, 1517-1520.
1033 Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed
1034 graph embedding resolves complex single-cell trajectories. Nature methods *14*, 979-982.
1035 Ramaswami, G., and Li, J.B. (2014). RADAR: a rigorously annotated database of A-to-I RNA
1036 editing. Nucleic Acids Research *42*, D109-D113.

1037 Rice, G.I., Kasher, P.R., Forte, G., Mannion, N.M., Greenwood, S.M., Szynkiewicz, M., Dickerson,
1038 J.E., Bhaskar, S.S., Zampini, M., and Briggs, T.A. (2012). Mutations in ADAR1 cause Aicardi-
1039 Goutieres syndrome associated with a type I interferon signature. Nature genetics *44*, 1243-1248.
1040 Sellars, M.C., Wu, C.J., and Fritsch, E.F. (2022). Cancer vaccines: Building a bridge over troubled
1041 waters. Cell *185*, 2770-2788.
1042 Smart, A.C., Margolis, C.A., Pimentel, H., He, M.X., Miao, D., Adeegbe, D., Fugmann, T., Wong,
1043 K.-K., and Van Allen, E.M. (2018). Intron retention is a source of neoepitopes in cancer. Nature
1044 Biotechnology *36*, 1056-1058.
1045 Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., III, Hao, Y.,
1046 Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data.
1047 Cell *177*, 1888-1902.e1821.
1048 Stühler, K., Köper, K., Pfeiffer, K., Tagariello, A., Souquet, M., Schwarte-Waldhoff, I., Hahn, S.A.,
1049 Schmiegel, W., and Meyer, H.E. (2006). Differential proteome analysis of colon carcinoma cell
1050 line SW480 after reconstitution of the tumour suppressor Smad4. Analytical and Bioanalytical
1051 Chemistry *386*, 1603-1612.
1052 Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F.
1053 (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality
1054 Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians *71*, 209-249.
1055 Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R.,
1056 Ramaswami, G., and Ariyoshi, K. (2017). Dynamic landscape and regulation of RNA editing in
1057 mammals. Nature *550*, 249-254.
1058 Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in
1059 RNA transcripts. Nucleic acids research *43*, e78-e78.
1060 Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., Del Risco, H., Ferrell,
1061 M., Mellado, M., Macchietto, M., Verheggen, K*., et al.* (2018). SQANTI: extensive
1062 characterization of long-read transcript sequences for quality control in full-length transcriptome
1063 identification and quantification. Genome research *28*, 396-411.
1064 The Trinity CTAT Project (2019). InferCNV of the Trinity CTAT Project.
1065 The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic
1066 acids research *49*, D480-D489.
1067 Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.-H., Porta-Pardo, E.,
1068 Gao, G.F., Plaisier, C.L., Eddy, J.A*., et al.* (2018). The Immune Landscape of Cancer. Immunity
1069 *48*, 812-830.e814.
1070 Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., and Eyras, E. (2018).
1071 SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple
1072 conditions. Genome biology *19*, 40.
1073 Veiga Diogo, F.T., Nesta, A., Zhao, Y., Mays Anne, D., Huynh, R., Rossi, R., Wu, T.-C., Palucka,
1074 K., Anczukow, O., Beck Christine, R*., et al.* (2022). A comprehensive long-read isoform analysis
1075 platform and sequencing resource for breast cancer. Science Advances *8*, eabg6711.
1076 Venables, R.S., McLean, S., Luny, D., Moteleb, E., Morley, S., Quinlan, R.A., Lane, E.B., and
1077 Hutchison, C.J. (2001). Expression of individual lamins in basal cell carcinomas of the skin. British
1078 Journal of Cancer *84*, 512-519.
1079 Wang, R., Zheng, D., Yehia, G., and Tian, B. (2018). A compendium of conserved cleavage and
1080 polyadenylation events in mammalian genes. Genome research *28*, 1427-1441.
1081 Wang, T.-Y., Liu, Q., Ren, Y., Alam, S.K., Wang, L., Zhu, Z., Hoeppner, L.H., Dehm, S.M., Cao,
1082 Q., and Yang, R. (2021). A pan-cancer transcriptome analysis of exitron splicing identifies novel
1083 cancer driver genes and neoepitopes. Molecular cell *81*, 2246-2260.e2212.
1084 Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M.,
1085 Barber, T., Ptak, J*., et al.* (2007). The Genomic Landscapes of Human Breast and Colorectal
1086 Cancers. Science *318*, 1108-1113.

1087 Xia, Z., Donehower, L.A., Cooper, T.A., Neilson, J.R., Wheeler, D.A., Wagner, E.J., and Li, W.
1088 (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR
1089 landscape across seven tumour types. Nature Communications *5*, 5274.
1090 Zhang, L., Li, Z., Skrzypczynska, K.M., Fang, Q., Zhang, W., O'Brien, S.A., He, Y., Wang, L.,
1091 Zhang, Q., Kim, A.*, et al.* (2020). Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted
1092 Therapies in Colon Cancer. Cell *181*, 442-459.e429.
1093 Zhang, M., Fritsche, J., Roszik, J., Williams, L.J., Peng, X., Chiu, Y., Tsou, C.-C., Hoffgaard, F.,
1094 Goldfinger, V., Schoor, O.*, et al.* (2018). RNA editing derived epitopes function as cancer antigens
1095 to elicit immune responses. Nature Communications *9*, 3919.
1096 Zhang, R., Deng, P., Jacobson, D., and Li, J.B. (2017). Evolutionary analysis reveals regulatory
1097 and functional landscape of coding and non-coding RNA editing. PLOS Genetics *13*, e1006563.

1098

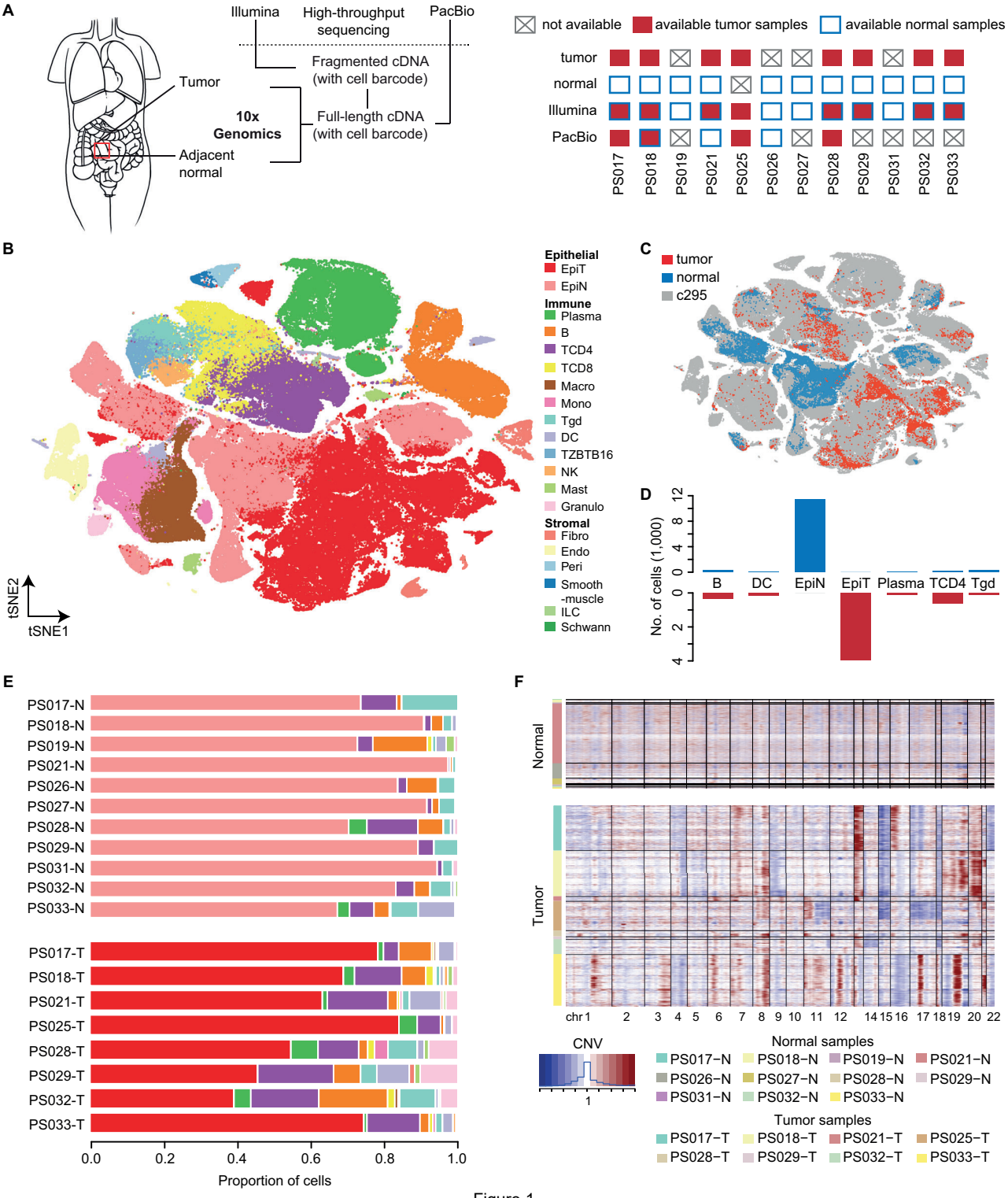
A
Illumina
High-throughput sequencing
PacBio
Fragmented cDNA (with cell barcode)
Tumor
10x Genomics
Full-length cDNA (with cell barcode)
Adjacent normal
not available
available tumor samples
available normal samples
tumor
normal
Illumina
PacBio
PS017
PS018
PS019
PS021
PS025
PS026
PS027
PS028
PS029
PS031
PS032
PS033
B
Epithelial
EpiT
EpiN
Immune
Plasma
B
TCD4
TCD8
Macro
Mono
Tgd
DC
TZBTB16
NK
Mast
Granulo
Stromal
Fibro
Endo
Peri
Smooth -muscle
ILC
Schwann
tSNE2
tSNE1
C
tumor
normal
c295
D
No. of cells (1,000)
B
DC
EpiN
EpiT
Plasma
TCD4
Tgd
E
PS017-N
PS018-N
PS019-N
PS021-N
PS026-N
PS027-N
PS028-N
PS029-N
PS031-N
PS032-N
PS033-N
PS017-T
PS018-T
PS021-T
PS025-T
PS028-T
PS029-T
PS032-T
PS033-T
Proportion of cells
F
Normal
Tumor
chr 1
CNV
Normal samples
PS017−N
PS018−N
PS019−N
PS021−N
PS026−N
PS027−N
PS028−N
PS029−N
PS031−N
PS032−N
PS033−N
Tumor samples
PS017−T
PS018−T
PS021−T
PS025−T
PS028−T
PS029−T
PS032−T
PS033−T

Figure 1

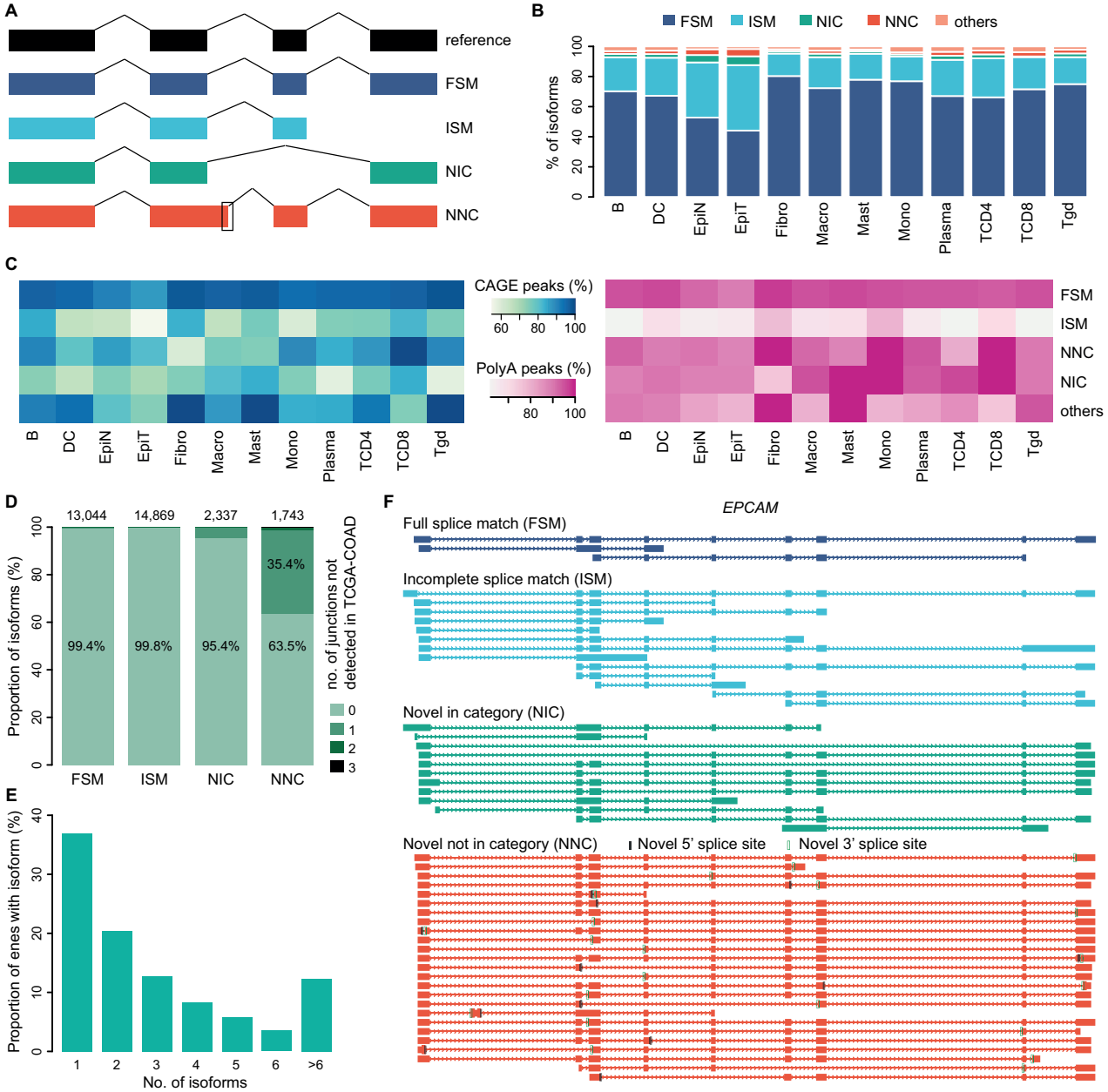
A
reference
FSM
ISM
NIC
NNC
B
FSM
ISM
NIC
NNC
others
% of isoforms
0
20
40
60
80
100
B
DC
EpiN
EpiT
Fibro
Macro
Mast
Mono
Plasma
TCD4
TCD8
Tgd
C
CAGE peaks (%)
60
80
100
PolyA peaks (%)
80
100
FSM
ISM
NNC
NIC
others
D
13,044
14,869
2,337
1,743
99.4%
99.8%
95.4%
63.5%
35.4%
Proportion of isoforms (%)
no. of junctions not detected in TCGA-COAD
0
1
2
3
FSM
ISM
NIC
NNC
E
Proportion of enes with isoform (%)
0
10
20
30
40
1
2
3
4
5
6
>6
No. of isoforms
F
EPCAM
Full splice match (FSM)
Incomplete splice match (ISM)
Novel in category (NIC)
Novel not in category (NNC)
Novel 5' splice site
Novel 3' splice site

Figure 2

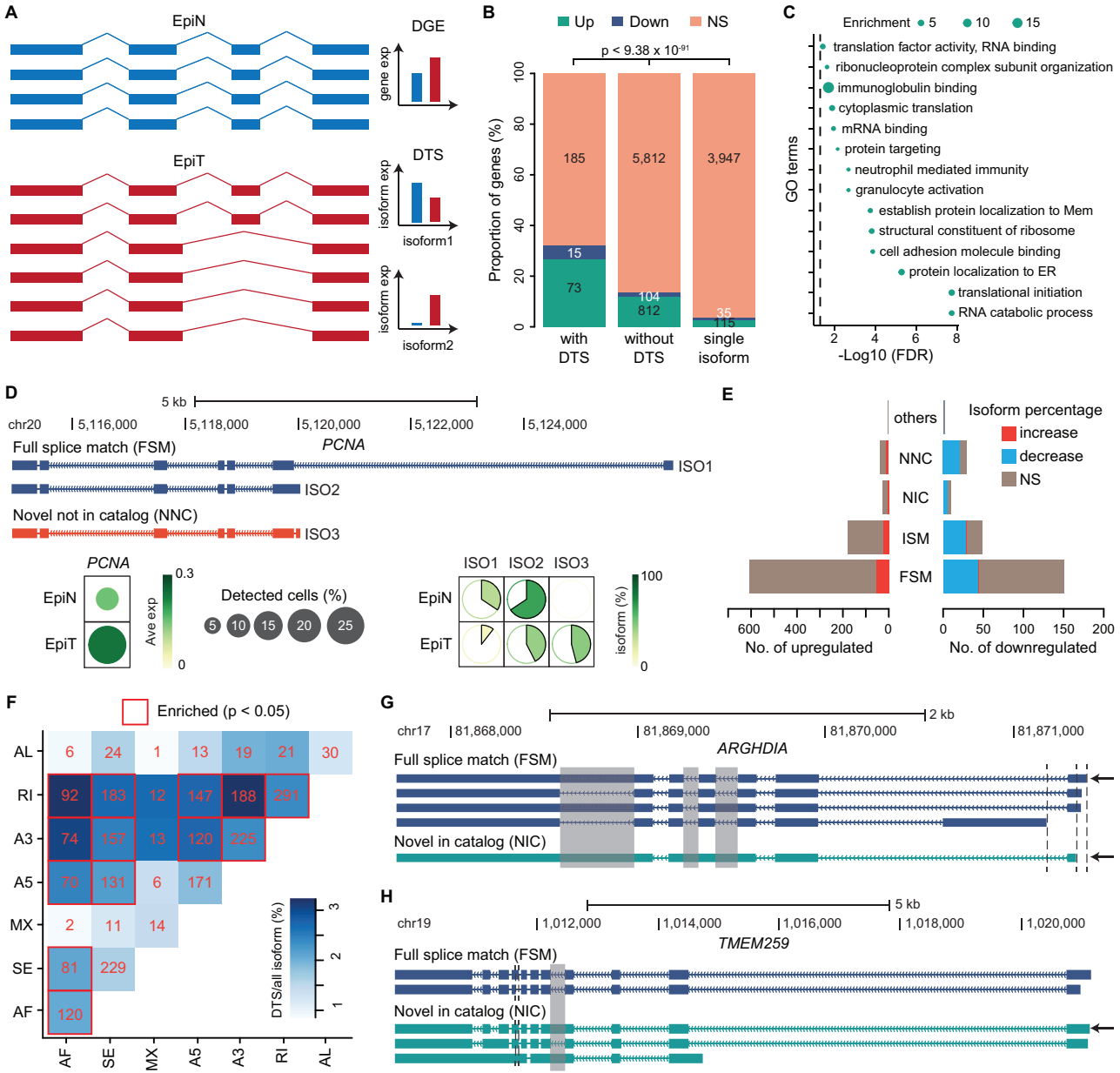
A
EpiN
DGE
gene exp
EpiT
DTS
isoform exp
isoform1
isoform2
B
Up
Down
NS
p < 9.38 x 10-91
Proportion of genes (%)
185
5,812
3,947
15
73
104
812
35
115
with DTS
without DTS
single isoform
C
Enrichment 5 10 15
translation factor activity, RNA binding
ribonucleoprotein complex subunit organization
immunoglobulin binding
cytoplasmic translation
mRNA binding
protein targeting
neutrophil mediated immunity
granulocyte activation
establish protein localization to Mem
structural constituent of ribosome
cell adhesion molecule binding
protein localization to ER
translational initiation
RNA catabolic process
GO terms
-Log10 (FDR)
D
5 kb
chr20
5,116,000
5,118,000
5,120,000
5,122,000
5,124,000
PCNA
Full splice match (FSM)
ISO1
ISO2
Novel not in catalog (NNC)
ISO3
PCNA
EpiN
EpiT
Ave exp
0.3
0
Detected cells (%)
5
10
15
20
25
ISO1 ISO2 ISO3
EpiN
EpiT
isoform (%)
100
0
E
others
Isoform percentage
increase
decrease
NS
NNC
NIC
ISM
FSM
600 400 200 0
No. of upregulated
0 50 100 150 200
No. of downregulated
F
Enriched (p < 0.05)
AL 6 24 1 13 19 21 30
RI 92 183 12 147 188 291
A3 74 157 13 120 225
A5 79 131 6 171
MX 2 11 14
SE 81 229
AF 120
AF SE MX A5 A3 RI AL
DTS/all isoform (%)
3
2
1
G
2 kb
chr17
81,868,000
81,869,000
81,870,000
81,871,000
ARGHDIA
Full splice match (FSM)
Novel in catalog (NIC)
H
5 kb
chr19
1,012,000
1,014,000
1,016,000
1,018,000
1,020,000
TMEM259
Full splice match (FSM)
Novel in catalog (NIC)

Figure 3

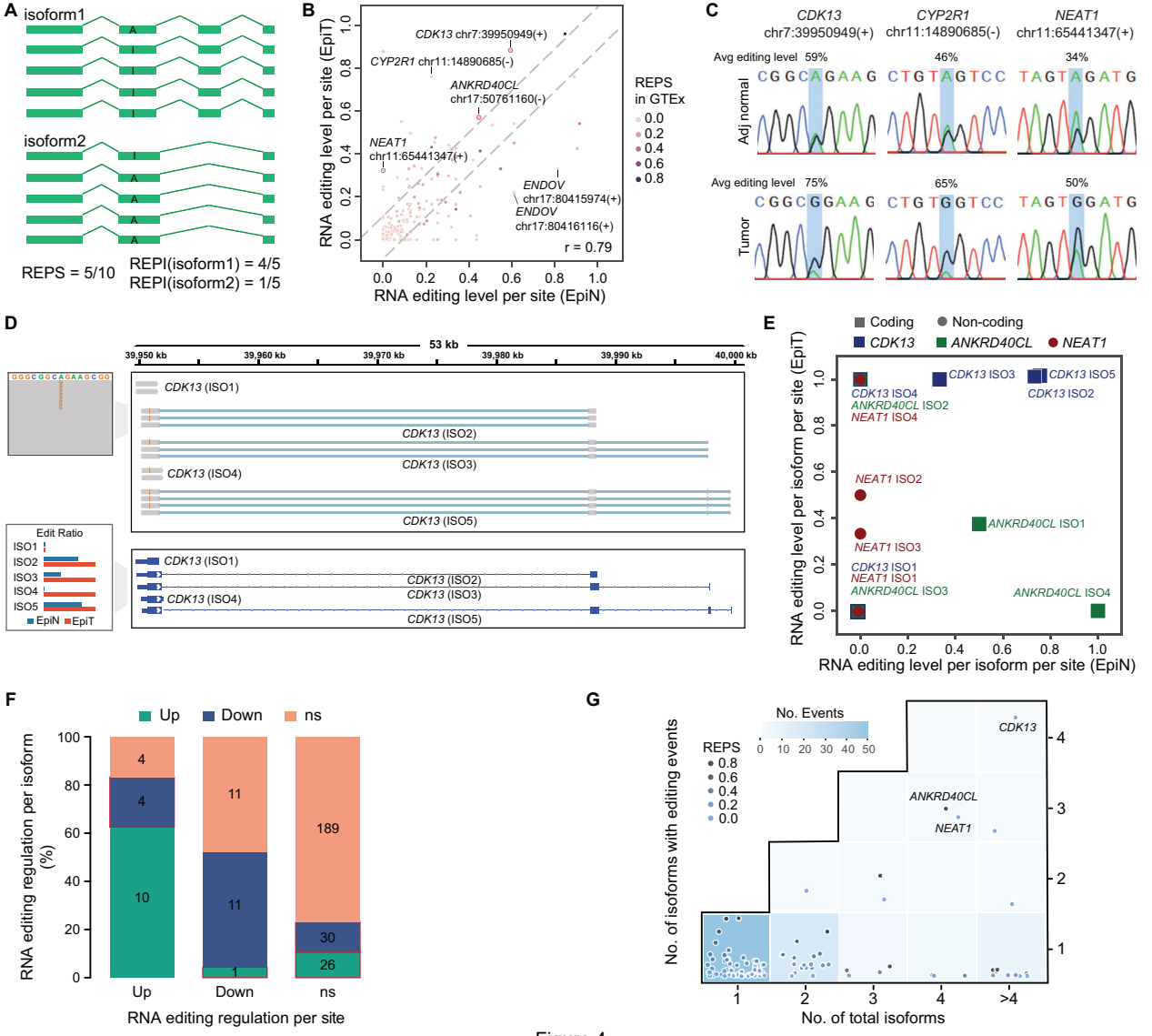
A
isoform1
isoform2
REPS = 5/10
REPI(isoform1) = 4/5
REPI(isoform2) = 1/5
B
CDK13 chr7:39950949(+)
CYP2R1 chr11:14890685(-)
ANKRD40CL chr17:50761160(-)
NEAT1 chr11:65441347(+)
ENDOV chr17:80415974(+)
ENDOV chr17:80416116(+)
r = 0.79
RNA editing level per site (EpiT)
RNA editing level per site (EpiN)
REPS in GTEx
0.0
0.2
0.4
0.6
0.8
C
CDK13 chr7:39950949(+)
CYP2R1 chr11:14890685(-)
NEAT1 chr11:65441347(+)
Avg editing level 59% 46% 34%
Adj normal
Avg editing level 75% 65% 50%
Tumor
D
53 kb
39,950 kb
39,960 kb
39,970 kb
39,980 kb
39,990 kb
40,000 kb
CDK13 (ISO1)
CDK13 (ISO2)
CDK13 (ISO3)
CDK13 (ISO4)
CDK13 (ISO5)
Edit Ratio
ISO1
ISO2
ISO3
ISO4
ISO5
EpiN
EpiT
E
Coding
Non-coding
CDK13
ANKRD40CL
NEAT1
CDK13 ISO3
CDK13 ISO5
CDK13 ISO4
CDK13 ISO2
ANKRD40CL ISO2
NEAT1 ISO4
NEAT1 ISO2
ANKRD40CL ISO1
NEAT1 ISO3
CDK13 ISO1
NEAT1 ISO1
ANKRD40CL ISO3
ANKRD40CL ISO4
RNA editing level per isoform per site (EpiT)
RNA editing level per isoform per site (EpiN)
F
Up
Down
ns
RNA editing regulation per isoform (%)
RNA editing regulation per site
G
No. Events
REPS
CDK13
ANKRD40CL
NEAT1
No. of isoforms with editing events
No. of total isoforms
>4

Figure 4

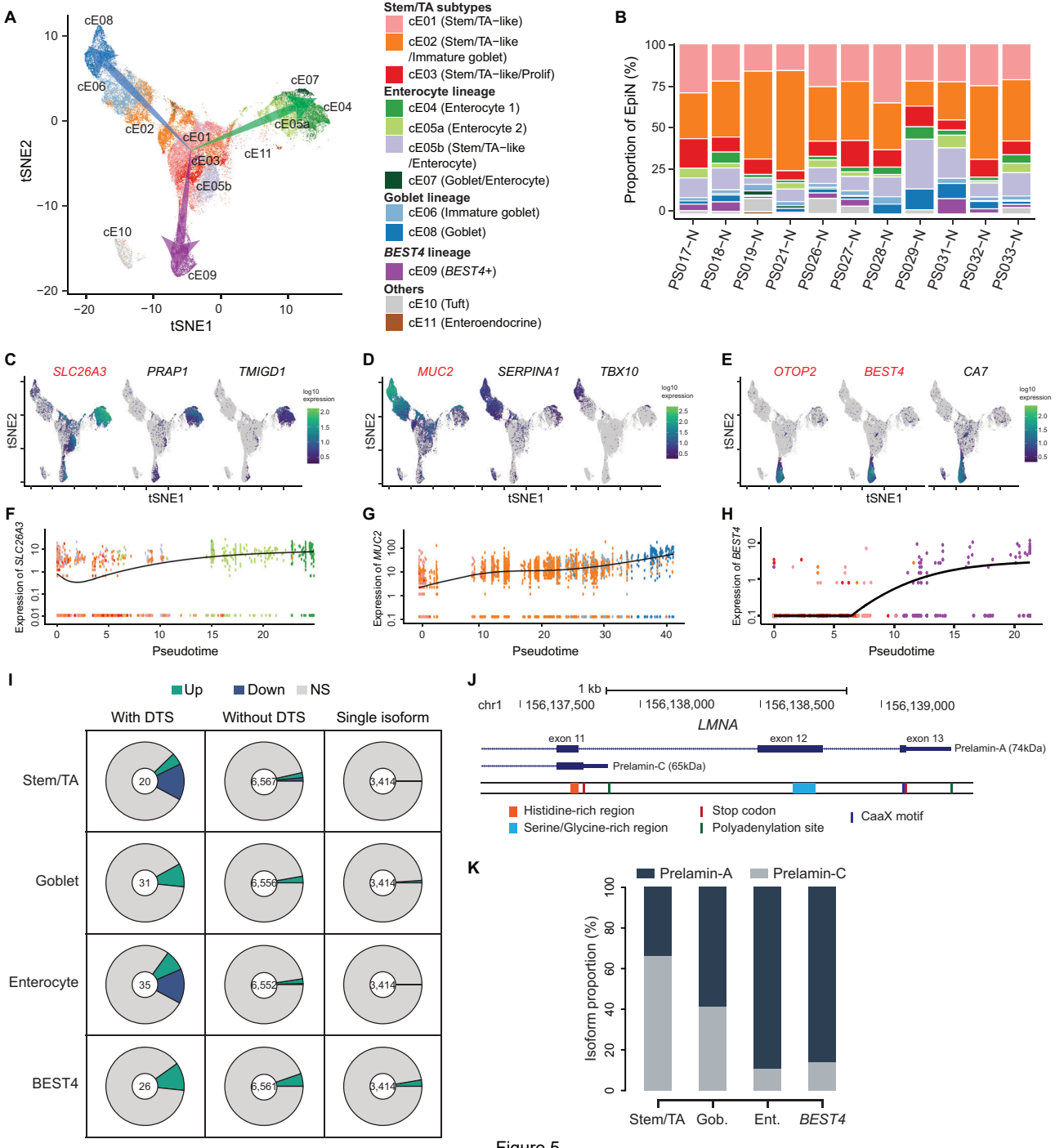
A
cE08
cE06
cE02
cE01
cE03
cE05b
cE11
cE07
cE04
cE05a
cE10
cE09
tSNE2
tSNE1
Stem/TA subtypes
cE01 (Stem/TA−like)
cE02 (Stem/TA−like /Immature goblet)
cE03 (Stem/TA−like/Prolif)
Enterocyte lineage
cE04 (Enterocyte 1)
cE05a (Enterocyte 2)
cE05b (Stem/TA−like /Enterocyte)
cE07 (Goblet/Enterocyte)
Goblet lineage
cE06 (Immature goblet)
cE08 (Goblet)
BEST4 lineage
cE09 (BEST4+)
Others
cE10 (Tuft)
cE11 (Enteroendocrine)
B
Proportion of EpiN (%)
PS017−N
PS018−N
PS019−N
PS021−N
PS026−N
PS027−N
PS028−N
PS029−N
PS031−N
PS032−N
PS033−N
C
SLC26A3
PRAP1
TMIGD1
log10 expression
D
MUC2
SERPINA1
TBX10
E
OTOP2
BEST4
CA7
F
Expression of SLC26A3
Pseudotime
G
Expression of MUC2
H
Expression of BEST4
I
Up
Down
NS
With DTS
Without DTS
Single isoform
Stem/TA
Goblet
Enterocyte
BEST4
20
6,567
3,414
31
6,556
35
6,552
26
6,561
J
1 kb
chr1
156,137,500
156,138,000
156,138,500
156,139,000
LMNA
exon 11
exon 12
exon 13
Prelamin-A (74kDa)
Prelamin-C (65kDa)
Histidine-rich region
Serine/Glycine-rich region
Stop codon
Polyadenylation site
CaaX motif
K
Prelamin-A
Prelamin-C
Isoform proportion (%)
Stem/TA
Gob.
Ent.
BEST4

Figure 5

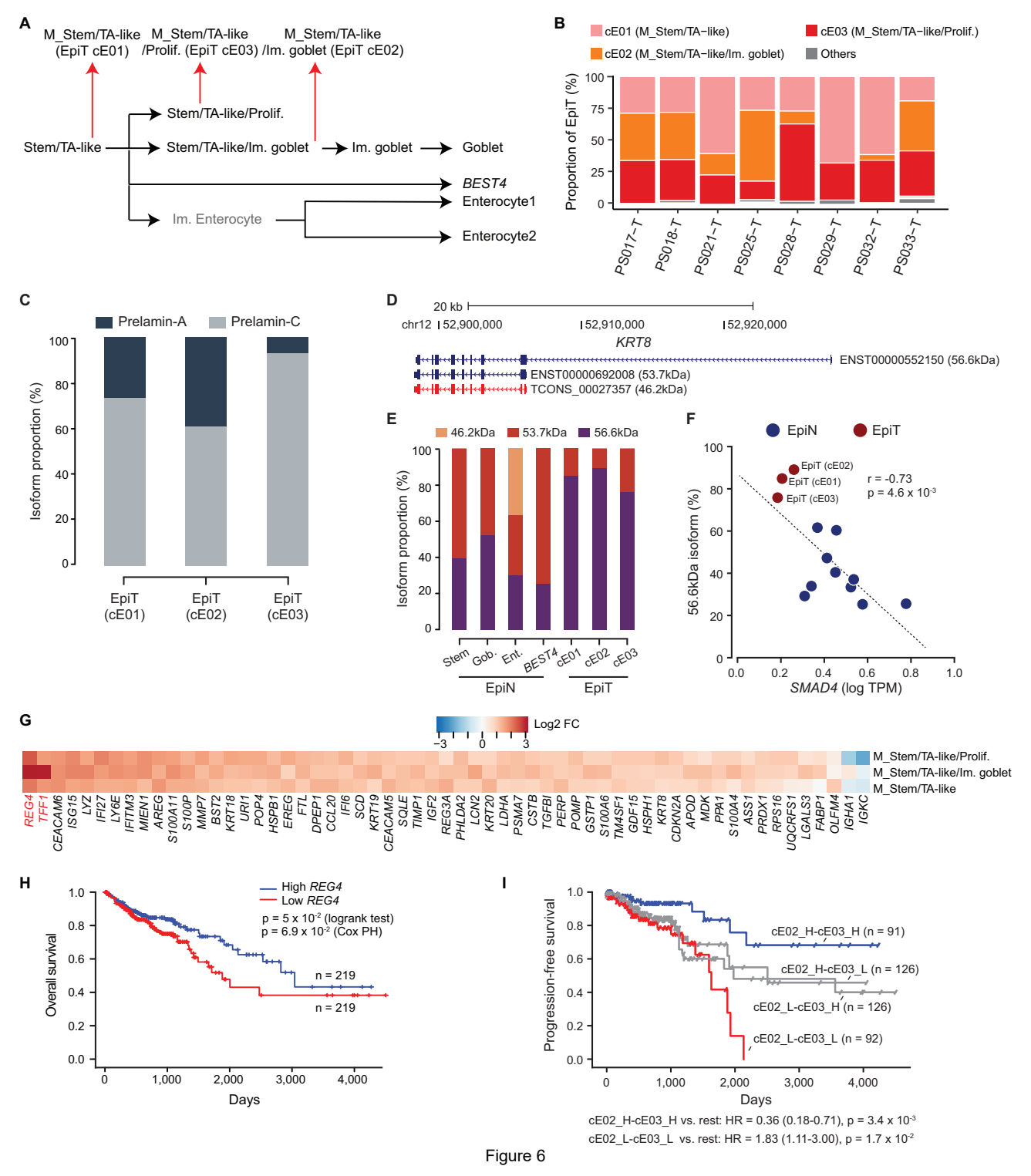
A
M_Stem/TA-like (EpiT cE01)
M_Stem/TA-like /Prolif. (EpiT cE03)
M_Stem/TA-like /Im. goblet (EpiT cE02)
Stem/TA-like
Stem/TA-like/Prolif.
Stem/TA-like/Im. goblet
Im. goblet
Goblet
BEST4
Enterocyte1
Im. Enterocyte
Enterocyte2
B
cE01 (M_Stem/TA−like)
cE02 (M_Stem/TA−like/Im. goblet)
cE03 (M_Stem/TA−like/Prolif.)
Others
Proportion of EpiT (%)
PS017−T
PS018−T
PS021−T
PS025−T
PS028−T
PS029−T
PS032−T
PS033−T
C
Prelamin-A
Prelamin-C
Isoform proportion (%)
EpiT (cE01)
EpiT (cE02)
EpiT (cE03)
D
20 kb
chr12
52,900,000
52,910,000
52,920,000
KRT8
ENST00000552150 (56.6kDa)
ENST00000692008 (53.7kDa)
TCONS_00027357 (46.2kDa)
E
46.2kDa
53.7kDa
56.6kDa
Isoform proportion (%)
Stem
Gob.
Ent.
BEST4
cE01
cE02
cE03
EpiN
EpiT
F
EpiN
EpiT
EpiT (cE02)
EpiT (cE01)
EpiT (cE03)
r = -0.73
p = 4.6 x 10-3
56.6kDa isoform (%)
SMAD4 (log TPM)
G
Log2 FC
−3
0
3
M_Stem/TA-like/Prolif.
M_Stem/TA-like/Im. goblet
M_Stem/TA-like
REG4
TFF1
CEACAM6
ISG15
LYZ
IFI27
LY6E
IFITM3
MIEN1
AREG
S100A11
S100P
MMP7
BST2
KRT18
URI1
POP4
HSPB1
EREG
FTL
DPEP1
CCL20
IFI6
SCD
KRT19
CEACAM5
SQLE
TIMP1
IGF2
REG3A
PHLDA2
LCN2
KRT20
LDHA
PSMA7
CSTB
TGFBI
PERP
POMP
GSTP1
S100A6
TM4SF1
GDF15
HSPH1
KRT8
CDKN2A
APOD
MDK
PPA1
S100A4
ASS1
PRDX1
RPS16
UQCRFS1
LGALS3
FABP1
OLFM4
IGHA1
IGKC
H
High REG4
Low REG4
p = 5 x 10-2 (logrank test)
p = 6.9 x 10-2 (Cox PH)
n = 219
n = 219
Overall survival
Days
I
cE02_H-cE03_H (n = 91)
cE02_H-cE03_L (n = 126)
cE02_L-cE03_H (n = 126)
cE02_L-cE03_L (n = 92)
Progression-free survival
Days
cE02_H-cE03_H vs. rest: HR = 0.36 (0.18-0.71), p = 3.4 x 10-3
cE02_L-cE03_L vs. rest: HR = 1.83 (1.11-3.00), p = 1.7 x 10-2

Figure 6

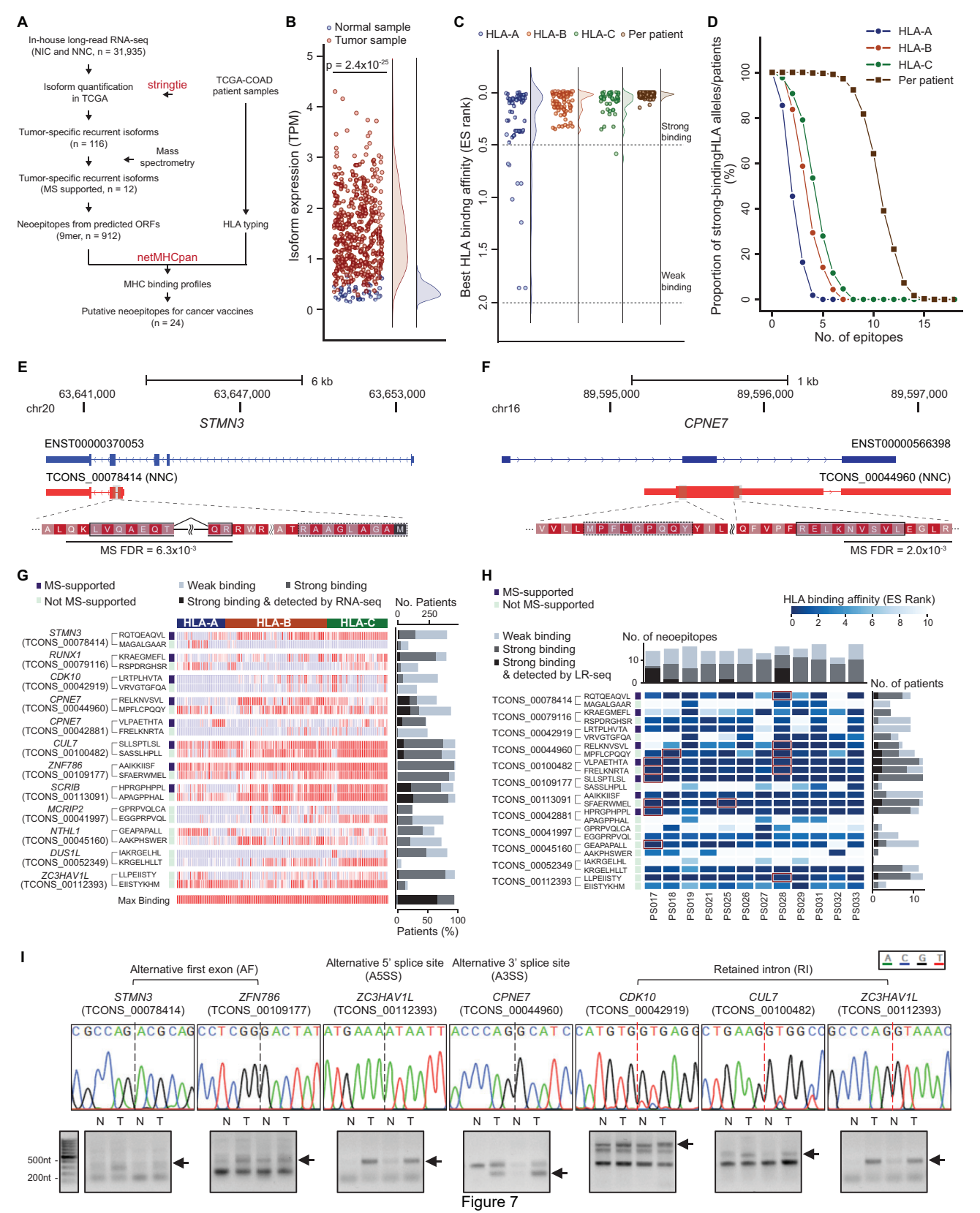
A
In-house long-read RNA-seq (NIC and NNC, n = 31,935)
Isoform quantification in TCGA
stringtie
TCGA-COAD patient samples
Tumor-specific recurrent isoforms (n = 116)
Mass spectrometry
Tumor-specific recurrent isoforms (MS supported, n = 12)
Neoepitopes from predicted ORFs (9mer, n = 912)
HLA typing
netMHCpan
MHC binding profiles
Putative neoepitopes for cancer vaccines (n = 24)
B
Normal sample
Tumor sample
p = 2.4x10-25
Isoform expression (TPM)
C
HLA-A
HLA-B
HLA-C
Per patient
Best HLA bindng affinity (ES rank)
Strong binding
Weak binding
D
HLA-A
HLA-B
HLA-C
Per patient
Proportion of strong-binding HLA alleles/patients (%)
No. of epitopes
E
6 kb
63,641,000
63,647,000
63,653,000
chr20
STMN3
ENST00000370053
TCONS_00078414 (NNC)
MS FDR = 6.3x10-3
F
1 kb
89,595,000
89,596,000
89,597,000
chr16
CPNE7
ENST00000566398
TCONS_00044960 (NNC)
MS FDR = 2.0x10-3
G
MS-supported
Not MS-supported
Weak binding
Strong binding
Strong binding & detected by RNA-seq
HLA-A
HLA-B
HLA-C
No. Patients
STMN3 (TCONS_00078414)
RUNX1 (TCONS_00079116)
CDK10 (TCONS_00042919)
CPNE7 (TCONS_00044960)
CPNE7 (TCONS_00042881)
CUL7 (TCONS_00100482)
ZNF786 (TCONS_00109177)
SCRIB (TCONS_00113091)
MCRIP2 (TCONS_00041997)
NTHL1 (TCONS_00045160)
DUS1L (TCONS_00052349)
ZC3HAV1L (TCONS_00112393)
Max Binding
Patients (%)
H
MS-supported
Not MS-supported
HLA binding affinity (ES Rank)
Weak binding
Strong binding
Strong binding & detected by LR-seq
No. of neoepitopes
No. of patients
PS017
PS018
PS019
PS021
PS025
PS026
PS027
PS028
PS029
PS031
PS032
PS033
I
Alternative first exon (AF)
Alternative 5' splice site (A5SS)
Alternative 3' splice site (A3SS)
Retained intron (RI)
STMN3 (TCONS_00078414)
ZFN786 (TCONS_00109177)
ZC3HAV1L (TCONS_00112393)
CPNE7 (TCONS_00044960)
CDK10 (TCONS_00042919)
CUL7 (TCONS_00100482)
ZC3HAV1L (TCONS_00112393)
N T N T
500nt
200nt

Figure 7