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

Abstract

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 [1,2]. 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 [3-6]. Based on the gene expression profiles in each cell, scRNA-seq can distinguish malignant from infiltrated cells and group them into subpopulations with distinct molecular properties. Furthermore, the varying compositions of each cell subpopulation for both malignant and infiltrated cells have been utilized to classify tumors into subtypes to predict prognosis and response to therapy [7,8].

To date, most scRNA-seq cancer studies only quantify expression at the gene level. However, the majority of human genes can generate multiple transcript isoforms, particularly in tumor cells in which widespread alterations in transcript structure arising from 5' and 3' ends, and alternative splicing (AS) differences are frequently observed [9-11]. Recently, we reported that pervasive splicing within the 3'-untranslated region (3' UTR) is upregulated in cancer to promote oncogene expression and tumorigenesis [12]. These results demonstrate that in addition to gene expression, changes in transcript structures also contribute to cancer pathogenesis. Traditionally, bulk RNA sequencing (RNA-seq) detects cancer-associated aberrations in specific regions of transcripts reconstructed from short reads, whereas the third-generation long-read sequencing (LR-seq), such as PacBio Iso-Seq, is able to directly capture dysregulations of full-length transcripts. Recently, two studies employed PacBio Iso-Seq with bulk RNA from breast cancer samples and gastric cancer cell lines to characterize their full-length transcripts, revealing extensive aberrant transcript structures that may be involved in tumor-related dysfunctions [13,14]. However, only a handful of studies have combined LR-seq and scRNA-seq to dissect tumor cell heterogeneities and transcript dysregulations at the isoform level.

In addition, tumors specifically express RNA with dysregulated transcript structure (DTS) that may contain new open read frames (ORF) encoding unique proteins that serve as a potential source of neoantigen in tumor cells [15]. Several studies have reported that tumor-associated intron retention, neo-splice junction, exitron and RNA editing could generate neoantigens [9,16-18]. To date, only neoantigens derived from DNA mutations with the potential to activate the immune response have been used to generate personalized neoantigen vaccines in clinical trials for melanoma and glioma treatment [19,20]. Studies for neoantigens arising from aberrant RNA processing are limited and

primarily based on short-read RNA-seq data that rely heavily on annotations that could compromise the precision of full ORF predictions. Conversely, the identification of full-length tumor-specific anomalous transcripts using LR-seq may enable predictions of complete novel ORFs to more accurately and exhaustively derive neoepitopes. Nevertheless, this concept and its feasibility have not been explored.

Here, we present a study integrating short-read and PacBio Iso-Seq-based long-read scRNA-seq of matched CRC clinical samples to build a full-length transcriptomic atlas and investigate isoform-specific dysregulations in CRC. We identified 394 DTSs from 273 genes in the tumor epithelial cells, many of which were caused by the coupling of multiple splicing events. Additionally, we detected tumor cell-associated and isoform-specific RNA editing events. By classifying normal (EpiN) and tumor epithelial (EpiT) cells into subpopulations, we unveiled genes and transcripts associated with three EpiN lineages, as well as three EpiT subpopulations comprising more than 90% of malignant cells, each linked to a distinct prognosis. Finally, we predicted the proteome of CRC based on the full-length transcriptome. By integrating the novel ORFs from recurrent tumor-specific transcripts and mass spectrometry (MS) data, we developed an algorithm to build a panel of 22 neoepitopes with strong binding affinities to MHC molecules for a wide range of CRC patients, which may be exploited for the development of universal neoantigen-based cancer vaccines.

Results

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

To study the landscape of full-length transcript isoforms in human CRC, we isolated single cells from eight tumor and eleven adjacent normal samples derived from 12 CRC patients, generated cDNA libraries using the 10x Genomics droplet-based method and performed matched Illumina short-read sequencing and PacBio Iso-Seq (**Fig. 1A, Extended Data Fig. 1a, Supplementary Table 1**). We curated a total of 18,966 high-quality cells with Illumina scRNA-seq data after applying conventional filtering (**Methods**, **Extended Data Fig. 1b**) with the number of cells per sample being slightly lower but comparable to related studies (**Extended Data Fig. 1c**) [5,6]. To accurately identify the cell types, considering the lower cell number in our dataset, we did not perform clustering from scratch. Instead, we mapped these cells to the Human Colon Cancer Atlas (c295), which is currently the most comprehensive single-cell atlas of CRC [6]. By integrating these two datasets (**Extended Data Fig. 1a**, **Methods**), the cells were grouped into 20 major clusters (**Fig. 1b**). We found that cells from the in-house tumor and normal samples were embedded into the c295 cell clusters without obvious batch effects (**Fig. 1c**). Sixteen out of 20 cell types from c295

67 were detected in the in-house dataset, of which eight consisted of more than 100 cells
68 (**Supplementary Table 1c**). Among these, most of the cells were epithelial cells, followed by the
69 immune and stromal cells (**Fig. 1d**), and the gene expression of representative cell markers were
70 validated in each cell type (**Extended Data Fig. 1d-f**).

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

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

96 With long-read scRNA-seq, we were able to perform the quantification of RNA expression at the
97 isoform level in each cell. Overall, we matched 1,994 out of 2,197 cell barcodes by comparing
98 those identified in long-read and short-read scRNA-seq (**Extended Data Fig. 2c**). These cells were
99 from 12 cell types, of which EpiN and EpiT are the most abundant, while immune and stromal cells

100 constituted only a small fraction of the cell population (**Extended Data Fig. 3a**). Across all cell
101 types, most of the isoforms were either FSM or ISM (**Fig. 2b**). Interestingly, we found that EpiT
102 expressed a slightly higher proportion of novel isoforms (NIC and NNC), suggesting more
103 transcriptomic abnormalities in tumor cells.

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

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

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

123 To systematically investigate the dysregulation of RNA isoforms in CRC, we defined two types of
124 dysregulations by comparing gene and isoform expression between EpiN and EpiT (**Fig. 3a**).
125 Dysregulated gene expression (DGE) refers to genes that have significantly different numbers of
126 overall transcripts between the two cell types, while dysregulated transcript structure (DTS) is
127 defined as a multi-isoform gene that expresses significantly different proportions of its isoforms
128 between EpiN and EpiT (**Methods**). The fold changes of the genes showing significant DGE as
129 measured by the two sequencing methods are strongly correlated (r = 0.77), with only three genes
130 (<1%) showing discordance (**Fig. 3b**). This high concordance indicates the reliability of gene
131 expression quantification using PacBio long-read sequencing data. We further experimentally
132 validated the top three DGE events, demonstrating the exclusive expression of *MMP7* and *REG3A*

133 in CRC tumor samples, while *REG1A* was significantly more highly expressed in the tumor samples
134 compared to the adjacent normal (**Fig. 3c**). In total, we identified 898 DGE events and 273 genes
135 with DTS from the PacBio data (**Fig. 3d**). We observed more DGEs that were upregulated (747)
136 than downregulated (151), suggesting high transcriptional activity in tumor cells (**Fig. 3d**).
137 Interestingly, we found that a higher frequency of genes with DTS had DGE, while single-isoform
138 genes tend to be less dysregulated (**Fig. 3d**). Gene ontology (GO) enrichment analysis of the genes
139 with DTSs showed enrichment of pathways related to RNA regulation, such as mRNA binding,
140 translational initiation and RNA catabolic process (**Fig. 3e**). This suggests potential autoregulation
141 of these genes, which is frequently observed for RNA-binding proteins, particularly splicing factors
142 [24-26]. Several immunological pathways were also implicated, indicating that DTS may contribute to
143 dysregulated immunological activity in tumor cells. For example, the proliferating cell nuclear
144 antigen gene, *PCNA*, presented both DGE and DTS. As a cell proliferation marker [27], its expression
145 was upregulated in EpiT and detected in a higher proportion of cells (**Fig. 3f**). The EpiN and EpiT
146 cells preferentially utilized distinct alternative first exons to differentially express the three *PCNA*
147 isoforms. Overall, isoform proportion changes accounted only for a small fraction of transcripts
148 with differential absolute expression levels between EpiT and EpiN for all four isoform categories,
149 suggesting that transcriptional activity may have a more profound impact on isoform expression
150 (**Extended Data Fig. 4a**).

151 We further applied SUPPA2 [28] to extract AS events for the identified transcript isoforms, which
152 were categorized into alternative 3' splice site (A3SS), alternative 5' splice site (A5SS), alternative
153 first exon (AF), alternative last exon (AL), retained intron (RI), skipped exon (SE) and mutually
154 exclusive exon (MX) (**Extended Data Fig. 4b**, **Methods**). Consistent with previous observations
155 [13,14], the most abundant AS events were from the AF, RI and SE categories (**Extended Data Fig.**
156 **4c**). We inspected how each isoform with DTS could arise from the different AS categories and
157 found that isoforms derived from AF, A3 and RI were more enriched with DTS (**Fig. 3g**, diagonal).
158 More importantly, our long-read scRNA-seq captured the structure of full-length isoforms,
159 allowing us to investigate the coupling between different AS categories. We observed that the
160 isoforms derived from coupled AF-RI, AF-A3 and A3-RI had the highest enrichment of DTS
161 isoforms (**Fig. 3g**, off-diagonal). For example, among the *ARGHDIA* isoforms, those utilizing the
162 distal first exon can retain any one of the three downstream introns (**Fig. 3h**), while the *TMEM259*
163 isoforms using a proximal downstream 3' splice site can retain an upstream intron (**Fig. 3i**). Our
164 findings underline the unique advantage of LR-seq in capturing isoform complexity that have thus
165 far eluded the short-read sequencing technology.

166 In addition to alternative splicing, post-transcriptional RNA modifications on single nucleotides
167 could also impact the mRNA sequence and expression without altering transcript structures.
168 Specifically, ADAR-mediated A-to-I RNA editing can modify both coding and non-coding regions
169 of mRNAs, as well as long non-coding RNAs, and has been implicated in several autoimmune
170 disorders [29] and multiple cancers [30].

171 We systematically identified the RNA editing events from isoforms sequenced using long-read
172 scRNA-seq by counting the number of edited reads (**Methods**). To characterize the RNA editing
173 level of a site, we computed RNA editing level per site (REPS) and per isoform (REPI), which are
174 the ratio of the number of all edited reads against all reads for all isoforms and against all reads for
175 a specific isoform, respectively (**Fig. 4a, Supplementary Table 2**). Of note, REPI calculation is
176 only possible with LR-seq as it allows unambiguous association of each edited mRNA read with
177 the isoform of origin.

178 A total of 196 RNA editing sites with sufficient sequencing read support were identified in EpiN
179 and EpiT cells. Most of these sites were on non-coding transcripts or the untranslated regions of
180 coding transcripts, with more events in the 3'UTRs than 5'UTRs (**Extended Data Fig. 5a**), which
181 is consistent with previous studies [31-33]. Furthermore, 10 sites were found in the coding regions,
182 which could potentially introduce codon alterations. For example, the tumor-associated event at
183 chr7:39950949(+) that results in the amino acid change Q103R in CDK13 (**Fig. 4b**), was also
184 observed in hepatocellular carcinoma [34]. Based on the REPS, RNA editing events in genes such as
185 *CDK13*, *CYP2R1* and *NEAT1* were significantly upregulated in EpiT compared to EpiN (**Fig. 4b**).
186 We validated these events using normal colon, CCD 841 CoN, and CRC cell lines, DLD-1 and
187 HCT116, as well as patient samples. The RNA editing levels for *CDK13, CYP2R1* and *NEAT1*
188 were variable and did not follow any trend in the cell lines (**Extended Data Fig. 5b**). However,
189 these were consistently higher in the patient tumor samples of at least four out of five matched pairs
190 tested (**Fig. 4c** and **Extended Data Fig. 5c**), highlighting the importance of performing
191 transcriptomic and validation studies using clinical samples.

192 To investigate isoform-specific RNA-editing, we computed and compared REPI between EpiN and
193 EpiT. For example, we first extracted the PacBio long reads belonging to each *CDK13* isoform and
194 calculated the editing level of the site on each isoform separately using the corresponding reads
195 (**Fig. 4d**). We observed heterogeneity of RNA editing levels of the same site among different
196 isoforms (**Fig. 4e** and **Extended Data Fig. 5d**). Four of the *CDK13* isoforms showed different
197 editing levels in EpiN cells but consistently high editing levels in EpiT, whereas all four isoforms
198 of *NEAT1* showed low editing levels in EpiN with varying editing levels in EpiT. Overall, the

direction of RNA editing changes between EpiT and EpiN was largely concordant using REPS and REPI (**Fig. 4f**). However, due to lower read coverage, we only observed a maximum of four edited isoforms per gene for a specific editing site while the per-isoform RNA editing events were only observed on one isoform for most genes (**Fig. 4g**). Taken together, our results suggest isoform specificity of the RNA editing mechanism, which is still poorly understood.

Genes and isoforms associated with normal epithelial cells from different lineages

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

Using monocle 3 [36], we inferred diffusion pseudotime along the differentiation trajectory of the three epithelial lineages (**Methods**) and identified 46, 31 and 38 lineage-specific genes (moran's I > 0.5, adjusted p-value $< 1.0 \times 10^{-5}$) for the enterocyte, goblet and *BEST4* lineages, respectively, including both well-known and novel markers (**Fig. 5c-e**). These showed a steady increase in expression along the lineage with respect to pseudotime (**Fig. 5f-h**). We next performed differential gene and isoform expression analysis between the stem/transient amplifying (TA) and differentiated subtypes along the three lineages. Overall, we observed a similar number of up- and downregulated genes in the stem/TA subtypes, but mostly upregulation of lineage-specific genes in the three differentiated lineages (**Extended Data Fig. 6a** and **6b, Supplementary Table 3**). The genes with differential regulation were almost mutually exclusive between the three differentiated subtypes (**Extended Data Fig. 6c**), indicating distinct transcriptomic regulatory mechanisms to commit to a specific differentiation lineage.

By categorizing the genes in each differential gene and isoform analysis with DGE and DTS (defined in **Fig. 3a**), we observed a higher percentage of genes with DTSs subjected to DGE compared to those without DTSs or with only one isoform (**Fig. 5i**), similar to the trend observed between EpiN and EpiT (**Fig. 3b**). The gene *LMNA* encoding two protein isoforms, Prelamin-A and Prelamin-C, through AS and alternative polyadenylation (APA), serves as a representative of

232 differential isoform usage between cell subtypes (**Fig. 5j**). We observed a trend of Prelamin-C to
233 Prelamin-A switching when stem cells differentiated into enterocytes, goblet and *BEST4*+ cells
234 (**Fig. 5k**), consistent with the findings in basal cell carcinomas, where Prelamin-A was negatively
235 correlated to cell proliferation and appeared in the later stage of differentiation [37]. Overall, our
236 analysis provides the first comprehensive characterization of isoforms associated with three colon
237 epithelial lineages.

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

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

251 We further identified dysregulated genes in the EpiT subtypes by comparing them to the
252 corresponding EpiN subtypes (**Extended Data Fig. 7c, Supplementary Table 4**). Most of these
253 were shared among the three EpiT subtypes (**Extended Data Fig. 7d**), indicating a common
254 underlying dysregulated mechanism of tumorigenesis. In addition to DGE, we performed
255 differential isoform analysis for the EpiT subpopulations (**Supplementary Table 5**). Differential
256 usage of the *LMNA* isoforms was also observed in the EpiT subtypes, with the highest usage of the
257 Prelamin-C isoform in the most proliferative subtype, cE03, and Prelamin-A isoform in cE02 (**Fig.**
258 **6c**). Furthermore, we noticed that several cytokeratin genes, including *KRT8* and *KRT19*, were
259 among the most significant DTS in all three EpiT subtypes (**Supplementary Table 5**). The
260 overexpression of *KRT8* has been linked to increased tumor progression and invasiveness in several
261 epithelial tumors [38-40]. We detected three major *KRT8* isoforms, including two annotated and one
262 novel isoform (**Fig. 6d**). The tumor subtypes mainly expressed the full-length isoform encoding a
263 56.6kDa protein, while the normal subtypes primarily utilized the short isoforms (**Fig. 6e**).
264 Consistent with a previous study showing that the tumor suppressor *SMAD4* regulated *KRT8*

splicing [41], we observed a negative correlation between *SMAD4* expression and the 56.6kDa isoform usage among the epithelial subtypes (**Fig. 6f**), further confirming the reliability of our isoform analysis. Overall, around 20 genes with DTS were identified in each EpiT subtype and consistent with those in the EpiN cells, genes with DTS were more likely to be DGE (**Extended Data Fig. 8a**). Contrary to the DGEs, the genes with DTS barely overlapped among the three subtypes (**Extended Data Fig. 8b**), suggesting potential dysregulation of isoform switching in the EpiT subpopulations.

Among the top dysregulated genes in the EpiT subtypes, *REG4* and *TFF1* showed the most striking upregulation in cE02 (**Fig. 6g**). Interestingly, we found that *REG4* and *TFF1* expression was highly correlated across the TCGA-COAD patients ($r = 0.75$, $p = 3.81 \times 10^{-71}$) and both were associated with better overall survival (OS) (**Fig. 6h** and **Extended Data Fig. 8c**). We speculate that the tumor samples with high *REG4* and *TFF1* expression may consist of a high proportion of cE02 cells, leading to better patient survival. To this end, we identified the marker genes in each subtype to define a signature score for the three subtypes (**Methods, Supplementary Table 5**). By classifying the TCGA tumors with cE02 and cE03 scores (**Extended Data Fig. 8d**), we found that the patient groups with both high cE02 and cE03 scores, and low cE02 and cE03 scores, had the best and worst OS and progression-free survival (PFS), respectively (**Fig. 6i** and **Extended Data Fig. 8e**), suggesting a synergistic effect of cE02 and cE03 scores on patient prognosis. Similar trends were observed using an independent microarray cohort curated by the CRC Subtyping Consortium (CRCSC) [42] (**Extended Data Fig. 8f** and **Extended Data Fig. 8g**). As cE02 and cE03 have relatively lower stemness features and higher proliferative activity compared to cE01, these results indicate that CRC tumors enriched with high stemness and low replicative cells are associated with more frequent progression and shorter OS time.

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

To investigate the effect of transcript structure alterations on the proteome, we predicted the open reading frames (ORFs) of all the identified isoforms. We performed a series of analyses on the predicted proteome and identified 8,984 novel ORFs from NIC and NNC isoforms (**Extended Data Fig. 9a-e**, **Supplemental Notes Section S1**). After integrating publicly available mass spectrometry (MS) data [43] from TCGA-COAD patients and in-house MS data from matched tumor and normal CRC patient samples and cell line, HCT116, 26 and 44 novel proteins from the NIC and NNC isoforms were supported (**Extended Data Fig. 9f**). There were also more MS-supported novel isoforms from EpiT and tumor samples compared to EpiN and normal samples, suggesting

298 that transcriptomic dysregulation in tumor cells could result in increased abnormalities of the
299 proteome (**Extended Data Fig. 9g**).

300 As abnormal proteins produced in cancer cells are critical for spontaneous anti-tumor immune
301 response, they serve as a potential source for neoantigen-based cancer vaccine development to
302 leverage the immune system for cancer treatment [15]. Therefore, we systematically investigated
303 putative neoepitopes in CRC derived from tumor-specific transcripts. Using the identified novel
304 isoforms together with annotated isoforms as reference, we quantified the expression of each
305 isoform in the TCGA-COAD samples (**Fig. 7a**, **Methods**). We retained 12 recurrent (>5 samples)
306 novel isoforms with significantly higher amount of expression and number of supporting reads of
307 novel junctions in the TCGA-COAD tumor samples than normal samples and whose ORF-derived
308 peptides were MS-supported (**Fig. 7b**, **Methods**). Out of five isoforms selected for experimental
309 confirmation, we validated the unique splice junctions and sequences for three isoforms arising
310 from AF and A5SS AS events, and one with combined A5SS and retained intron (RI) in colon
311 epithelial and CRC cell lines, as well as at least four of the five matched patient samples tested
312 (**Fig. 7c, Extended Data Fig. 10a** and **10b, Supplementary Table 6**). All novel 9-mers derived
313 from the predicted ORFs of the four experimentally validated isoforms were regarded as candidate
314 neoepitopes in CRC.

315 As a limited number of epitopes can be loaded into a cancer vaccine, it is imperative to maximize
316 vaccine efficacy by selecting an optimized subset of neoepitopes with a high binding affinity to the
317 class I MHC molecules that facilitate antigen presentation. To develop cancer vaccines for a broad
318 range of patients, we considered the diversity of HLA alleles across different patients and predicted
319 the MHC binding profiles of 458 TCGA-COAD patients to the neoepitopes from the four recurrent
320 tumor-specific isoforms. We formulated a greedy algorithm to select four to six neoepitopes from
321 each of the four isoforms to maximize their coverage to the HLA alleles or patients, while
322 prioritizing those that overlapped with MS-supported peptides (**Fig. 7d**, **Extended Data Fig. 11a,**
323 **Supplementary Table 7, Methods, Algorithm 1**).

324 Among the 22 selected neoepitopes, there were always one or more showing strong binding
325 affinities (ES rank < 0.5) to 74% of HLA-A, 67% HLA-B and 70% HLA-C MHC molecules, with
326 only four HLA-B alleles not passing the weak binding threshold (ES rank > 2.0) (**Fig. 7e**). For each
327 patient, the highest binding affinities with their three types of MHC molecules were always above
328 the strong binding threshold (**Fig. 7e**). All patients carry at least three strong-binding neoepitopes
329 and 50% of them have at least five strong-binding neoepitopes from the panel (**Fig. 7f**). To better
330 represent the global population, we generated another slightly different panel of neoepitopes

331 (**Extended Data Fig. 11b, Supplementary Table 7**) from the same four isoforms but optimized
332 for the top 50 high-frequency alleles in each HLA type [44], which similarly showed strong HLA
333 binding affinities (**Extended Data Fig. 11c**).

334 To illustrate an example from our panel, the neoepitopes selected for the tumor-specific novel
335 isoform of *STMN3* (TCONS_0078414) that utilizes a coding alternative first exon, were
336 'RQTQEAQVL' that spans the novel splice junction and overlaps with MS-detected peptides, and
337 'ALPRPQAGW' that has a strong HLA binding affinity (**Fig. 7d**). The neoepitopes for *CPNE7*
338 resulting from an alternative 5' splice site were similarly selected (**Fig. 7h**). The HLA binding
339 profile of the neoepitope panel showed heterogeneous affinities to the three types of HLA (**Fig.**
340 **7d**). Overall, all patients have at least one strong-binding neoepitope and 50.4% (n=231) of them
341 with corresponding isoform detection in the tumor samples (**Fig. 7d**). Furthermore, using the same
342 panel of neoepitopes optimized for TCGA patients, we also observed strong HLA binding potential
343 in the 12 in-house patients, with all patients having at least six strong-binding neoepitopes (**Fig. 7i,**
344 **Supplementary Table 8**). In addition to serving as a source of neoantigen, we speculate that these
345 tumor-specific isoforms may also have certain oncogenic functions considering their tumor-
346 specific expression pattern. In line with this, we found that the overexpression of their ORFs
347 promotes cell growth *in vitro* (**Fig. 7j** and **Extended Data Fig. 10c, d**). Collectively, these findings
348 demonstrate a proof-of-concept for the discovery of neoepitopes based on the tumor-specific
349 transcriptome, as well as the invaluable application potential of long-read scRNA-seq in the
350 development of neoepitope-based cancer vaccines.

351 **Discussion**

352 The intestinal epithelium is the fastest self-renewing tissue in mammals, which contains epithelial
353 cells with diverse differentiation statuses, proliferation activities, and other properties and
354 functions. Despite various markers and gene expression profiles that have been characterized for
355 each epithelial cell subtype, the diversity of RNA isoforms in these subtype remains
356 underappreciated. Our study provides a full-length transcriptome across different epithelial
357 subtypes and identifies isoforms associated with enterocyte, goblet and *BEST4* lineages,
358 respectively. Although alternative RNA isoforms have been shown to regulate cell fate
359 determination of stem cells [45], the prevalence and potential role of alternative isoforms in intestinal
360 epithelial differentiation are poorly understood, therefore our data may provide a valuable resource
361 to facilitate further investigations.

362 CRC is the third most common cancer type and ranks as the second leading cause of cancer death
363 around the world [46]. Consistent with the first high-throughput scRNA-seq study on CRC and other

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

379 Despite recent reports of the full transcriptome by bulk RNA LR-seq in breast and gastric cancers
380 [13,14], single-cell LR-seq studies are rare and limited to cancer cell lines [47,48]. Moreover, previous
381 LR-seq studies in cancer focused only on alterations in single AS events. Here, we identify
382 hundreds of DTS arising from combined AS events and calculate isoform- and cell-type-specific
383 RNA editing levels, all of which originate from primary CRC tissues and are only achievable by
384 single-cell LR-seq. Nonetheless, our analyses are limited by technological constraints leading to
385 the relatively low coverage of single-cell PacBio sequencing. Thus, the development of methods to
386 increase the capture efficiency of single-cell isoform sequencing will be beneficial for future
387 transcriptomic studies [49].

388 Therapeutic neoantigen-based cancer vaccines aim to leverage the immune system for the treatment
389 of cancer [15]. Due to somatic mutations and transcriptomic dysregulation, tumor cells produce
390 abnormal proteins that are absent in normal cells. In the event that fragments of these proteins are
391 transported to the cell surface via the MHC I pathway, they may potentially be recognized by the
392 immune system as an antigen. We show that the full-length transcriptome in tumor cells with MS
393 data from matched patient samples can provide a comprehensive source to detect aberrant mRNA-
394 derived neoantigens. Moreover, we propose an algorithm that provides an optimized combination
395 of recurrent tumor-specific neoepitopes with high HLA binding affinities for a wide spectrum of
396 CRC patients. Recently, the phase II clinical trial for a neoantigen-based mRNA vaccine by

397 Moderna has met the endpoint [50]. In contrast to the customized vaccine designed based on a
398 neoepitope identified by exome sequencing of a patient tumor, the panel we propose can potentially
399 target larger CRC patient cohorts without sequencing individual tumors. This could facilitate the
400 development of neoantigen-based cancer vaccines with less economic and time constraints to
401 benefit a larger population.

402 ## Methods

403 **CRC clinical samples**

404 The protocols for the human studies comply with all relevant ethical regulations and are approved
405 by the National Healthcare Group Domain Specific Review Boards (NHG DSRB Ref:
406 2018/01032). All patients gave informed written consent. The CRC tissue samples were derived
407 from 12 patients from National University Hospital (details on the patients and samples are
408 summarized in **Supplementary Table 1**).

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

410 Colon tissues from patients were cut into approximately 5 mm pieces and washed three times with
411 cold PBS until the supernatant became clear. The supernatant was discarded and the tissue pieces
412 were further cut into 1-2 mm pieces. 5 ml of PBS containing 5 mM EDTA (EDTA-PBS) was added
413 to the tissue fragments and incubated at 4°C for 30 minutes with gentle agitation, followed by
414 centrifugation at 200 *x*g, 4°C for 5 minutes. The supernatant was discarded and the tissue fragments
415 were vigorously resuspended in 5 ml of cold EDTA-PBS. The fragments were allowed to settle by
416 gravity and the supernatant was collected in a fresh 15 ml tube. This resuspension/ sedimentation
417 step was repeated five times. The supernatant from each round was collected in a separate 15 ml
418 tube and checked under an inverted microscope for single cells. All fractions containing single cells
419 were combined and the cells were pelleted by centrifugation at 200 *x*g, 4°C for 5 minutes, and
420 washed with cold EDTA-PBS. The single cells were resuspended in serum-free DMEM containing
421 0.05% trypsin and 5 U/ml DNaseI and incubated at 37°C with shaking for 15 minutes. The
422 dissociated cells were centrifuged at 200 *x*g, 4°C for 5 minutes, the supernatant was removed and
423 the pellet was resuspended in 2 ml of serum-free DMEM supplemented with 2 U/ml DNaseI, and
424 filtered using a 35 mm mesh. The cells were pelleted and washed once with cold PBS and
425 resuspended in PBS supplemented with 0.4% BSA, followed by cell quantification.

426 For GEM generation, 10,000 cells from each tissue sample were used to load the Chromium Next
427 GEM Chip G (10x Genomics) and Chromium controller following the manufacturer's protocol
428 (Chromium Next GEM Single Cell 3' v3.1). Post GEM-RT cleanup, cDNA amplification and
429 library construction were performed according to the manufacturer's instructions with some

430 modifications. Prior to the fragmentation step, 20 µl of cDNA from each sample were separately
431 aliquoted for PacBio long-read sequencing. To generate the Single Cell 3' Gene Expression library
432 for Illumina sequencing, 10 µl of cDNA were processed as per the 10x Genomics protocol, with an
433 extension time of 3 minutes in the sample index PCR step. The final Illumina libraries were
434 sequenced on the NovaSeq 6000 platform with S1 or S4 flowcells using the recommended
435 sequencing protocol.

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

443 **Plasmids**

444 The open reading frames (ORFs) of validated neoepitopes followed by a HA tag were cloned into
445 pcDNA3.1+ using the primers and restriction sites listed in **Supplementary Table 9**. All constructs
446 were verified by Sanger sequencing.

447 **Cell culture and transfection**

448 Human normal colon cell line, CCD 841 CoN (ATCC: CRL-1790), was cultured in Dulbecco's
449 Modified Eagle Medium (DMEM). Human CRC cell lines, DLD-1 (Horizon Discovery: HD PAR-
450 086) and HCT116 (ATCC: CCL-247) were cultured in Roswell Park Memorial Institute (RPMI)
451 1640 Medium and DMEM, respectively. All culture media were supplemented with 10% FBS,
452 glutamine and penicillin/streptomycin. The cells were maintained at 37°C and 5% CO_2 in a
453 humidified atmosphere. For overexpression experiments, cells were seeded at 100,000 cells per
454 well in 12-well plates 24 h prior to transfecting 1 µg of each plasmid using Lipofectamine 3000
455 (Thermo Fisher) following the manufacturer's instructions.

456 **Soft agar assay**

457 Cells were transfected 18-24 h prior to seeding as described above. A 0.6% base agarose was
458 prepared in 12-well plates on the day of seeding. Transfected cells were trypsinized, harvested and
459 counted. Seeding densities of 5,000 and 7,000 cells per well were used for CCD 841 CoN and
460 DLD-1, respectively. The cells were resuspended in their respective growth media, mixed with
461 agarose to a final agarose concentration of 0.3% and seeded on the prepared base. After agarose
462 solidification, 0.5 ml of growth medium was added to each well. The cells were maintained in

463 conditions described above and the growth medium was changed every 2-3 days. The colonies were
464 imaged after 7-14 days under 4x magnification using the Olympus IX71 microscope and quantified
465 using ImageJ (v1 51j8).

466 **RNA extraction and PCR**

467 Trizol and the PureLink RNA Mini Kit (Thermo Fisher) were used to extract total RNA from DLD-
468 1 and HCT116 cell lines. The ISOLATE II RNA Mini Kit (Bioline) was used to extract RNA from
469 CRC patient samples following the manufacturer's protocol. cDNA was generated using the High
470 Capacity cDNA Reverse Transcription Kit (Thermo Fisher). PCR experiments were performed
471 using Platinum™ Taq DNA Polymerase (Thermo Fisher). Subsequently, PCR products were
472 subjected to agarose gel electrophoresis, gel extraction and purification using the QIAquick Gel
473 Extraction Kit (Qiagen) and Sanger sequencing. Chromatograms were visualized and multiple
474 sequence alignments were performed using SnapGene (v6.2.1). RNA editing levels were quantified
475 using ImageJ (v1.50). PCR primers used are listed in **Supplementary Table 9**.

476 **Protein extraction and western blot analysis**

477 Cells were harvested 48 h post-transfection, lysed and 10 µg of lysates were fractionated using 12%
478 SDS-PAGE followed by transfer to PVDF membranes as described previously [12]. Specific primary
479 and secondary antibodies in 5% BSA-TBST were incubated with the membranes to probe for genes
480 of interest.

481 **Mass Spectrometry**

482 To generate an in-house mass spectrometry dataset of matched tumor-normal sample pairs, <10mg
483 tissue for each tissue were processed with the iST sample preparation kit (Preomics). Samples were
484 taken up in 100 µL LYSE buffer and mixed with 50 mg of protein extraction beads (Diagenode).
485 Samples were sonicated with 10-20 cycles at 30sec ON/OFF in a Diagenode Bioruptor Plus until
486 complete tissue disruption was observed. Protein concentrations were determined using the Pierce
487 BCA Protein Assay Kit (Thermo Scientific). 20 µg of protein per sample were digested with the
488 iST kit according to the manufacturer's instructions and samples were eluted with the fractionation
489 add-on into 3 fractions. Digested peptides were quantified with the Quantitative Fluorometric
490 Peptide Assay kit (Thermo Scientific), normalized to 0.1 µg/µL and stored at -20°C prior to analysis
491 on an EASY-nLC 1200 Liquid Chromatograph (Thermo Scientific) coupled to a timsTOF fleX
492 (Bruker) mass spectrometer. For each sample, 5 µL digested peptides (500 ng) for each fraction
493 were injected and separated on an Aurora series column (25 cm length, 75 µm inner diameter, C-
494 18 1.7 µm; IonOpticks) with an integrated captive spray emitter. The column was mounted on a
495 captive spray ionization source and temperature controlled by a column oven (Sonation) at 50°C.

496 A 105-min gradient from 2 to 40 % (v/v) acetonitrile in 0.1 % (v/v) formic acid at a flow of 400
497 nL/min was used. Spray voltage was set to 1.65 kV. The timsTOF fleX was operated with data-
498 dependent acquisition (DDA) in PASEF mode with 10 PASEF ramps per topN acquisition cycle
499 (cycle time 1.17s) and a target intensity of 10,000. Singly charged precursor ions were excluded
500 based on their position in the m/z-ion mobility plane and precursor ions that reached the target
501 intensity were dynamically excluded for 24 seconds.

502 **Summary of bioinformatic analysis**

503 The bioinformatic analysis workflow is summarized in **Extended Data Fig. 13**. Additionally, the
504 command lines and parameters used in the analysis are also provided (**Supplemental Notes Section**
505 **S2**).

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

507 CellRanger v4.0.0 (https://support.10xgenomics.com/single-cell-gene-
508 expression/software/downloads/latest) was used for the pre-processing of 10x droplet-based
509 scRNA-seq raw data. Specifically, 'cellranger count' with default parameters was used to align
510 sequencing reads to the human reference genome (GRCh38). The gene expression count matrix
511 was then obtained based on the genomic coordinates defined by the GENCODE annotation (release
512 v37 for GRCh38). Subsequently, 'cellranger aggr' was performed to aggregate the count matrices
513 of samples in multiple 'cellranger count' runs. The aggregated count matrix of all samples was then
514 loaded into Seurat (v3.2.3) [51] for subsequent analysis. A cell was filtered out if it satisfied any of
515 the following criteria: (1) expression of fewer than 250 genes, (2) detected fewer than 300 UMIs,
516 (3) more than 70% of the UMIs mapped to mitochondrial genes. All mitochondrial genes were
517 discarded in subsequent analysis. This analysis resulted in a total of 18,966 high-quality cells from
518 22 tissue samples.

519 The gene expression count matrix was log-normalized using Seurat's 'NormalizeData' function.
520 The top 2000 genes with the highest variations were selected using the 'SelectIntegrationFeatures'
521 subroutine of Seurat, and the 'IntegrateData' subroutine was used to remove batch effects across
522 the samples. Integration anchors were chosen based on the first thirty dimensions of the canonical
523 correlation analysis (CCA) [51]. We performed the analysis with above procedures separately on the
524 normal and tumor samples to prevent the batch effect correction procedure from eliminating
525 differences between the normal and tumor cells (**Extended Data Fig. 1**).

526 **Cell type identification**

527 We integrated the in-house data to a reference dataset, Human Colon Cancer Atlas (c295) [6]from
528 the Broad Institute Single Cell Portal (https://singlecell.broadinstitute.org/), for more robust and

529 reliable identification of the cell types. The c295 atlas contains gene expression profiles of 371,223
530 cells. Due to memory constraint, a subset of the atlas containing 50% of the cells in the c295 atlas
531 was used as the reference based on a stratified sampling procedure across different cell types.
532 Transfer anchors were chosen by the 'FindTransferAnchors' subroutine of Seurat using the first
533 thirty dimensions of the principal component analysis (PCA). The 'TransferData' subroutine was
534 used to transfer the c295's 'ClusterMidway' annotation to the in-house data via the selected
535 anchors. The 'ClusterMidway' labels of the c295 atlas annotate cells into 20 different types,
536 including (with boldface indicating the availability of the cell type in the in-house data):

537 - 2 epithelial cell types: **normal epithelial cells (EpiN**) and **tumor epithelial cells (EpiT**).
538 - 13 immune cell types: **B cells, plasma cells, CD4+ T cells, CD8+ T cells, gamma-delta**
539 **T cells, PLZF+ T cells, natural killer cells, innate lymphoid cells, dendritic cells,**
540 **granulocytes, macrophages, mast cells, monocytes**,
541 - 5 stromal cell types: **fibroblasts**, endothelial cells, pericytes, smooth muscle cells,
542 Schwann cells

543 Several post-processing steps were performed for the result from the above classification: (1) We
544 corrected all the EpiN misclassified as EpiT cells in the normal samples. (2) To improve the purity
545 of the identified epithelial tumor cells, we trained a gradient boosting classifier (xgboost v1.5.0 [35])
546 using the combined EpiN and EpiT cells from the c295 atlas and the in-house dataset. Following a
547 similar 5-fold cross-validation scheme as in [6], we divided the combined dataset into five
548 independent splits and used four folds for training and the rest one fold for testing. The classifier
549 was trained to perform the binary classification of whether a cell is a EpiT or a EpiN cell. Only
550 when a cell came from a tumor sample and with >75% prediction probability as EpiT cells in the
551 test split did we regard it as a candidate for the third step. (3) For all candidates from step (2), we
552 inferred somatic copy number variations (SCNV) of the cells based on their gene expression
553 profiles using inferCNV (v1.11.1, https://github.com/broadinstitute/infercnv) [21]. The final set of
554 tumor epithelial cells was kept as those having >10% genomic SCNV regions. The statistics of
555 epithelial cells that passed the xgboost and the SCNV criteria are reported in **Supplementary Table**
556 **1e**. The epithelial cells from tumor samples that did not pass either the gradient boosting criteria or
557 the SCNV criteria were deemed as undetermined epithelial cells.

558 **Dimensionality reduction and visualization**

559 PCA was performed on the log-normalized and centered gene expression matrix for dimensionality
560 reduction. t-distributed stochastic neighbor embedding (t-SNE) [52] was then performed to generate

561 2D embeddings for visualization. tSNE was supplied with the first 30 components of PCA and run
562 with a perplexity value of 30.

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

564 SMRT Link (v9.0.0, https://www.pacb.com/support/software-downloads/) was utilized to pre-
565 process PacBio Iso-seq long-read scRNA-seq data. We used the Isoseq-deduplication pipeline [53]
566 recommended by the official repository of PacBio (https://github.com/PacificBiosciences/IsoSeq/)
567 for the processing of raw data from the PacBio Sequel II system. The 'ccs' module of SMRT Link
568 was applied to generate consensus reads (CCS Reads) from the BAM file that contained the PacBio
569 subreads. The reads with less than 90% accuracy ('--min-rq 0.9') were discarded. The 'lima'
570 module was utilized in orientation determination and primer removal of the reads. The cell barcodes
571 and UMIs of each read were extracted by 'isoseq3 tag' using the library design 'T-12U-16B'. The
572 processed reads were trimmed of the Poly(A) tail and potential concatemers were removed by the
573 command 'isoseq3 refine'. The resulting full-length non-concatemer (FLNC) reads were clustered
574 together according to their cell barcode and UMI with 'isoseq3 dedup' to deduplicate and generate
575 a consensus sequence of each molecule. Those consensus sequences were then mapped to the
576 human reference genome (GRCh38) using minimap2 [54]. According to the genomic coordinates of
577 the mapped molecules, the structures of each isoform represented in the gtf format were obtained
578 by utilizing the script 'collapse_isoforms_by_sam.py' from the cDNA cupcake toolkit [55] (v27.0.0,
579 https://github.com/Magdoll/cDNA_Cupcake).

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

581 We merged the isoforms identified across different samples with gffcompare [56]. We then utilized
582 SQANTI3 (v3.3) [57] to compare the identified isoforms against the reference transcriptome
583 annotation, which produced quality control and classification information for each isoform. To
584 provide a comprehensive reference for SQANTI3, we merged databases from multiple sources,
585 including GENCODE (Release v37 for GRCh38) and RefSeq (NCBI Homo sapiens Annotation
586 Release 109 for GRCh38), for the annotation of isoform structures. With this assembled reference,
587 SQANTI3 classified each isoform into FSM, ISM, NIC, NNC and several other structural types [57].
588 The rule-based filter functionality of SQANTI3 was used to filter out the isoforms that were likely
589 to be sequencing artifacts such as intra-priming, reverse transcriptase template switching (RTS),
590 and non-canonical splice sites with low read coverage. The remaining isoforms and their associated
591 barcodes are assembled into an isoform-level expression matrix $\mathbf{X}^{(\mathrm{iso})} \in \mathbb{R}^{N_{\mathrm{iso}} \times N_{\mathrm{cells}}}$, where N_{iso} is
592 the number of isoforms, and N_{cells} is the number of all cells associated with those isoforms. Finally,
593 we searched the barcodes from PacBio data against the cell barcodes detected in the same sample

594 in the 10x Illumina sequencing data to calculate the expression (UMI counts) of each isoform in
595 each cell. To evaluate the quality of the identified isoforms, we examined whether the 5' and 3'
596 ends of each isoform are supported by the TSS and PAS reference. The TSS reference was a union
597 of TSSs from GENCODE, RefSeq and refTSS (v3.1) [58], while the PAS reference contains all
598 transcript 3' ends in GENCODE, RefSeq and PAS from PolyASite (v2.0) [59] and PolyADB (v3.2)
599 [60]. Furthermore, splice junctions from each identified isoform were overlapped with junction reads
600 extracted from the RNA-seq data in The Cancer Genome Atlas colon adenocarcinoma cohort
601 (TCGA-COAD).

602 **Splice event extraction and analysis**

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

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

613 Dysregulated gene expression (DGE) was identified by comparing the expression of genes (log-
614 normalized UMI) in cells from one group (cell type or subtype) to that from another group. In brief,
615 for each gene, we first counted the number of cells (n_1) expressing this gene with the expression
616 threshold (0) and the total number of cells (n_2) in each group. Next, we applied Fisher's exact test
617 on the 2×2 contingency table formed by n_1 and n_2 from two groups. In addition, we also applied
618 the Mann-Whitney U test (Wilcoxon Rank-Sum test) on the gene expression (log-normalized UMI)
619 in cells from different groups. A significantly upregulated gene was identified by requiring: (1)
620 Benjamini-Hochberg (BH) - adjusted p values (FDR) from Fisher's exact test < 0.01, (2) odds ratio
621 > 2 and (3) log2 fold change > 0. Significantly downregulated genes were identified using the
622 identical p-value threshold, but with the opposite directional changes (log2 fold change < 0 and
623 odds ratio < 1/2). Dysregulation in isoform expression was identified similarly using the expression
624 of each isoform. For evaluating concordance of DGE quantified by the two sequencing methods,
625 we used genes showing significant changes in both measurements.

626 To identify dysregulated transcript structures (DTSs), we counted the number of PacBio UMIs
627 supporting a specific isoform versus all alternative isoforms of the same gene for cells from the two
628 groups. The significance of DTSs was computed using Fisher's exact test on the 2×2 contingency
629 table formed by four numbers from the two groups of cells. The significant DTSs were identified
630 by requiring: (1) BH-adjusted p-value (FDR) < 0.01, (2) the percentage of isoform changes > 0 and
631 (3) odds ratio > 2, or (1) BH-adjusted p-value (FDR) < 0.01, (2) the percentage of isoform changes
632 < 0 and (3) odds ratio < 1/2.

633 **GO enrichment analysis**

634 The list of genes with DTS from comparing EpiT to EpiN was uploaded to WebGestalt [61] as input
635 and all the genes with detected expression in these two cell types were used as the background to
636 identify the enriched GO terms (FDR < 0.05, enrichment score > 2).

637 **Calling RNA editing events**

638 We assembled a comprehensive reference of potential RNA editing sites from three different
639 sources: (1) the RADAR database [62], (2) the DARNED database [63] and (3) Tan et al., 2017. [64]. For
640 each putative RNA editing site in the reference, we systematically examined the evidence of reads
641 supporting an A>G change at that locus (or a T>C change on the negative strand) using the
642 'mpileup' subcommand of bcftools [65]. We then associated each event with the isoforms and cell
643 barcodes from which the event was discovered. To remove interference from genomic SNVs, we
644 searched the candidate RNA editing sites against the dbSNP database and discarded the sites with
645 an allele frequency greater than 0.01 in any genomic studies it had been concerned. We retained
646 RNA editing sites with more than 10 read coverage and computed an edit ratio for each site, defined
647 as the number of edited reads divided by the number of total read coverage. For these RNA editing
648 sites, we additionally calculated their per isoform edit ratio using the reads that can be
649 unambiguously identified to a specific isoform and if the per isoform read coverage is greater than
650 five. The per isoform edit ratio is defined as the number of edited reads divided by the number of
651 total read coverage from the isoform (**Fig. 4a**).

652 **Lineage and trajectory analysis**

653 We employed monocle 3 [36] for the single-cell lineage and trajectory analysis on the epithelial
654 component of the short-read scRNA-seq dataset. The principal graph was constructed using the
655 'learn_graph' subroutine of monocle 3 with default parameters using the tSNE coordinates and the
656 nodes corresponding to the four stem cells subclusters were designated as the root for diffusion
657 pseudotime inference. The epithelial cells that were also detected in the PacBio long-read scRNA-
658 seq were associated with their isoform level expression values. The expression of genes and

659 isoforms, as well as the percentage of isoforms that were associated with a specific lineage, were
660 identified based on Moran's I statistic [66].

661 **Quantification of isoforms in TCGA samples**

662 Isoform-level expression of the TCGA-COAD bulk samples was quantified using stringtie (v2.2.1)
663 [67] using the merged transcriptomic reference of novel identified isoforms, GENCODE (version 37)
664 and RefSeq (NCBI Homo sapiens Annotation Release 109 for GRCh38).

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

666 We identified the top differentially expressed genes in the EpiT subtypes cE02 and cE03
667 (**Supplementary Table 6**). We then computed the z-scores of those up- and down-regulated genes
668 in each bulk sample. The signature score of an EpiT subtype is defined as the averaged z-scores of
669 the up-regulated genes minus the averaged z-scores of the down-regulated genes.

670 **Survival analysis**

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

675 **Proteomic analysis of isoforms**

676 To comprehensively characterize the proteome in CRC, GeneMarkS-T [68] was used to predict ORFs
677 in the PacBio-identified isoforms. For comparison and annotation of the predicted ORFs, we
678 collected the human reference proteome from Uniprot (release 2022-02) [69], including all sequences
679 from the reviewed (SwissProt), unreviewed (TrEmbl) and spliced isoform (Varsplice) collections
680 at protein existence (PE) level PE=1 and PE=2. Using blastp [70], we searched each predicted ORF
681 sequence against the UniProt reference proteome, and the reference sequence with the highest
682 alignment score was assigned to it as its nearest homolog in the database. The ORFs with a
683 similarity of less than 99% to their nearest homolog in the reference were considered novel protein
684 sequences. Protein domains of each ORF sequence, as defined by the Pfam-A hidden Markov
685 model (HMM)-based multiple sequence alignments [71], were annotated by hmmscan
686 (http://hmmer.org/). We regarded a novel ORF sequence as having a domain gain if it contained a
687 proper superset of the Pfam domains from its nearest Uniprot homolog, or a domain loss if it
688 contained only a proper subset of them. We also inferred the sensitivity of the PacBio-identified
689 isoforms to nonsense-mediated decay (NMD) based on the relative position of the predicted stop
690 codon from the last splice junction. Lastly, the subcellular localization of the ORF-derived proteins
691 was predicted using DeepLoc [72].

692 **Mass spectrometry data analysis**

693 To validate the predicted ORFs, particularly the novel ORFs, we collected mass spectrometry (MS)
694 data from three independent sources: (1) a CRC cohort from the National Cancer Institute's Clinical
695 Proteomic Tumor Analysis Consortium (CPTAC), (2) an in-house MS dataset consisting of 22
696 samples from 11 CRC patients (11 tumor and 11 paired adjacent normal), and (3) an in-house MS
697 data from a colon cancer cell line, HCT116. With the same pipeline as a previous study [73], we
698 searched MS2 spectra from the MS data against a reference by integrating annotated protein
699 sequence from UniProt (release 2022-02) and novel ORFs predicted based on isoforms derived
700 from LR-seq using Comet [74]. Five peptides were reported for each spectrum query and these
701 peptides were further ranked and scored for final identification using Percolator [75]. The validated
702 novel ORFs were defined by its inclusion of at least one novel peptide (FDR < 0.01) in the MS data
703 that does not match with any protein sequence from the UniProt database.

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

705 To obtain a shortlist of tumor-specific novel transcript isoforms, we extracted the NIC and NNC
706 isoforms identified in our long-read scRNA-seq profiles and quantified their expression levels in
707 the TCGA-COAD bulk samples along with the known isoforms in GENCODE and RefSeq. Tumor-
708 specificity was ensured by both filtering for isoforms discovered only from the EpiT cells and not
709 expressed (TPM < 0.5) in the TCGA-COAD normal samples. Additionally, we required each novel
710 isoform to possess at least one novel junction which has supporting reads in <5% of TCGA normal
711 samples and >5% TCGA tumor samples. To ensure active translation of the isoforms, we prioritized
712 12 isoforms from the shortlist whose ORF subsequences were supported at least once by the MS
713 data from tumor samples but without any peptides detected in MS data from normal samples. Out
714 of five isoforms in this shortlist selected for experimental confirmation, we validated the unique
715 splice junction sequences and RT-PCR products for three isoforms arising from AF and A5SS AS
716 events, and one isoform with retained introns (RI) in CRC cell lines, as well as at least two of the
717 five matched patient samples tested.

718 To derive the set of tumor-specific neoepitopes from the tumor-specific novel transcript isoforms,
719 we assembled a collection of all the *k*-mers ($k = 9$) that are present in the predicted ORF sequences
720 and further curated a subset containing tumor-specific novel *k*-mers by requiring: (1) the *k*-mers
721 are not in the Uniprot reference, (2) the *k*-mers are only present in ORFs in transcript isoforms from
722 EpiT but not from EpiN or other cell types, (3) the *k*-mers are from the transcript isoforms that are
723 expressed (TPM > 0.5) in more than five TCGA-COAD tumor samples and not expressed (TPM <
724 0.5) in the normal samples (4) the *k*-mers are from the ORFs in the transcript isoforms classified as

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

727 For a given population, we predicted the binding affinity of each neoepitope *k*-mer against the
728 frequent HLA alleles in the population (represented as the set $\mathcal{H}$) using netMHCpan (v4.1) [76]. We
729 used the rank of elution score, EL_rank < 2 as the threshold of binding/non-binding for a *k*-mer to
730 a specific HLA allele. In this way, we obtained a mapping $\mathcal{M}$ that represented the HLA alleles that
731 a neoepitope *k*-mer can bind to. For example, $\mathcal{M}[s]$ ($\mathcal{M}[s] \subseteq \mathcal{H}$) contains the HLA alleles to
732 which neoepitope s ($s \in \mathcal{S}$) binds.

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

737 Input: The size of the shortlist n and the sets of HLA alleles covered by each epitope
738 $\{M[s_1], M[s_2], \dots, M[s_{|\mathcal{S}|}]\}$.

739 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.

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

742 Input: The size of the shortlist n, the sets of HLA alleles covered by each epitope
743 $\{M[s_1], M[s_2], \dots, M[s_{|\mathcal{S}|}]\}$, and the weight of each HLA allele a ($a \in \mathcal{H}$), $\mathcal{W}[a]$.

744 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.

745 As both the weighted and unweighted maximum coverage problems are NP-hard [77], we used the
746 following greedy algorithm for the selection of an epitope shortlist with size n (**Algorithm 1**,
747 SELECT-KMERS-FOR-POPULATION), **which is also illustrated in Extended Data Fig. 11A**. We
748 omitted the algorithm for the unweighted version in the presentation here as it is a special case of
749 the weighted version when $\mathcal{W}[a] = 1$ ($\forall a \in \mathcal{H}$). On each iteration, we selected an epitope *k*-mer
750 not yet in the shortlist that would result in the maximum gain in the weight of covered HLA alleles
751 (**Algorithm 1**, SELECT-BEST-KMER). If the HLA alleles were already completely covered before the
752 size of the shortlist reaches n, we reset the set of uncovered HLA alleles and continued the iteration.
753 In this way, the remaining epitopes in the shortlist were optimized again for the maximum coverage
754 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. [78] 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 (𝒮, ℳ, 𝒰, 𝒲)
inputs:
    𝒮, the set of k-mers from recurrent tumor-specific isoforms
    ℳ, the mapping from k-mers to the identities (HLA alleles or patients) they bind to
    𝒰, the set of uncovered identities
    𝒲, the weight of the identities
outputs:
    s_best, the selected best k-mer in set 𝒮
        s_best = POP(𝒮) , c_best = SUM({𝒲[a] foreach a in M[s_best]})
        foreach s in 𝒮
            c = SUM( { 𝒲[a] foreach x in M[a] ∩ 𝒰 } )
            if c > c_best
                c_best ← c, s_best ← s
            end
        end
        return s_best
```

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.

Figure legends

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

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

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

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

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

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

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

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

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

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

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

Figure 3 Dysregulated transcript structures in epithelial tumor cells.
(a) Illustration of dysregulated gene expression (DGE) and dysregulated transcript structure (DTS) in epithelial tumor (EpiT) compared to normal (EpiN) cells.
(b) Scatterplot showing the correlation in fold change of genes between EpiT and EpiN quantified by long- and short-read sequencing. Up (down): measured by both sequencing methods as significantly up- (down-) regulated; discordant: measured by both sequencing methods as significant but with inconsistent directions of change; NS: not significant.

823 (c) PCR validation of the top three DGE events, *MMP7*, *REG1A* and *REG3A*, from (b) using four
824 pairs of CRC tumor (T) and adjacent normal (N) patient samples (PS).

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

828 (e) Enrichment of Gene Ontology (GO) terms for genes with DTS.

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

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

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

839 **Figure 4 Dysregulated RNA editing in epithelial tumor cells.**
840 (a) Illustration of two approaches to calculate RNA editing levels: RNA editing level per site
841 (REPS) and RNA editing level per isoform (REPI).

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

844 (c) Sanger sequencing validation of the RNA editing sites and levels as shown in (b).
845 Representative chromatograms are shown.

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

848 (e) REPI for each event on each isoform from the sites highlighted in (b) in *CDK13*, *ANKRD40CL*
849 and *NEAT1*.

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

852 (g) Heatmap showing the number of isoforms with detected RNA editing versus the total number
853 of detected isoforms of the gene. Each dot represents an editing event at a gene locus. Dots are
854 colored according to their REPS and those for the three RNA editing events in **Figure 4E** are
855 highlighted.

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

860 (b) Proportion of EpiN subtypes in each sample.

861 (c-e) The top identified markers with lineage-specific expression for (c) enterocytes, (d) goblet and
862 (e) *BEST4*. Known markers are in red.

863 (f-h) RNA expression of lineage-specific marker genes over the pseudotime was estimated by
864 trajectory analysis along the (f) enterocyte, (g) goblet and (h) *BEST4* lineages.

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

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

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

872 **Figure 6 Dysregulation of genes and isoforms in the epithelial tumor cell subtypes.**
873 (a) Schematic showing the comparisons between each EpiT subtype and their corresponding EpiN
874 cell subtype.

875 (b) Proportion of EpiT subtypes in each sample.

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

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

880 (e) Proportion of the *KRT8* transcripts encoding each CK8 protein isoform in each EpiN and EpiT
881 subtype.

882 (f) Correlation between the usage of the 56.6kDa transcript isoform and *SMAD4* expression in each
883 EpiN and EpiT subtype.

884 (g) Fold changes (log2 transformed) of top significant genes with dysregulated expression in each
885 EpiT subtype compared to the corresponding EpiN subtype.

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

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

889 **Figure 7 Identification of common neoantigens from recurrent tumor-specific transcripts.**
890 (a) Workflow for the identification of neoepitopes from novel tumor-specific recurrent transcript
891 isoforms for cancer vaccine development.

892 (b) Estimated maximum expression level (quantified by Stringtie) of MS-supported candidate
893 isoforms and the number of supporting reads of novel junctions in the normal and tumor TCGA-
894 COAD samples.

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

899 (d) HLA binding profile of the panel of 22 neoepitopes against the HLA alleles (HLA-A, HLA-B,
900 HLA-C) among the TCGA-COAD patients. The right panel summarizes the number of patients
901 with at least one HLA allele showing 'weak/strong binding' affinities to the neoepitope of the
902 corresponding row. 'Strong binding & detected by RNA-seq' denotes neoepitopes derived from
903 isoforms with TPM>0.5 (as in **Fig. 7B**). The last row summarizes the percentage of TCGA-COAD
904 patients with at least one neoepitope satisfying the two criteria above.

905 (e) HLA binding profile of the selected neoepitopes. Each dot represents the best binding affinity
906 of MHC molecule(s) from an HLA allele (columns 1-3) or a TCGA-COAD patient (column 4)
907 against the panel of neoepitopes. Neoepitopes whose sequences overlap with MS are marked as
908 'MS supported' and vice versa.

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

911 (g-h) Selection of neoepitopes for a novel isoform of (g) *STMN3* (TCONS_0078414) and (h)
912 *CPNE7* (TCONS_00044960). The selected MS-supported neoepitopes are encircled by solid boxes
913 and the non-MS-supported ones by dashed line boxes.

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

920 (j) Effect of overexpressing the open reading frames derived from the validated neoepitopes in (c)
921 on anchorage-independent growth in DLD-1 cells.

922 **References**

923 1 Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of
924 genetic heterogeneity in cancer evolution. *Nature* **501**, 338-345, doi:10.1038/nature12625
925 (2013).
926 2 Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature*
927 **501**, 328-337, doi:10.1038/nature12624 (2013).
928 3 Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular
929 heterogeneity in human colorectal tumors. *Nature Genetics* **49**, 708-718,
930 doi:10.1038/ng.3818 (2017).
931 4 Becker, W. R. *et al.* Single-cell analyses define a continuum of cell state and composition
932 changes in the malignant transformation of polyps to colorectal cancer. *Nature Genetics*
933 **54**, 985-995, doi:10.1038/s41588-022-01088-x (2022).
934 5 Zhang, L. *et al.* Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies
935 in Colon Cancer. *Cell* **181**, 442-459.e429, doi:https://doi.org/10.1016/j.cell.2020.03.048
936 (2020).
937 6 Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal cancer.
938 *Cell*, doi:https://doi.org/10.1016/j.cell.2021.08.003 (2021).
939 7 Lee, H.-O. *et al.* Lineage-dependent gene expression programs influence the immune
940 landscape of colorectal cancer. *Nature Genetics* **52**, 594-603, doi:10.1038/s41588-020-
941 0636-z (2020).
942 8 Joanito, I. *et al.* Single-cell and bulk transcriptome sequencing identifies two epithelial
943 tumor cell states and refines the consensus molecular classification of colorectal cancer.
944 *Nature Genetics* **54**, 963-975, doi:10.1038/s41588-022-01100-4 (2022).
945 9 Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from
946 8,705 Patients. *Cancer Cell* **34**, 211-224.e216,
947 doi:https://doi.org/10.1016/j.ccell.2018.07.001 (2018).

10 Xia, Z. *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nature Communications* **5**, 5274, doi:10.1038/ncomms6274 (2014).
11 Demircioğlu, D. *et al.* A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. *Cell* **178**, 1465-1477.e1417, doi:https://doi.org/10.1016/j.cell.2019.08.018 (2019).
12 Chan, J. J. *et al.* Pan-cancer pervasive upregulation of 3′ UTR splicing drives tumourigenesis. *Nature Cell Biology* **24**, 928-939, doi:10.1038/s41556-022-00913-z (2022).
13 Veiga Diogo, F. T. *et al.* A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Science Advances* **8**, eabg6711, doi:10.1126/sciadv.abg6711 (2022).
14 Huang, K. K. *et al.* Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. *Genome biology* **22**, 1-24 (2021).
15 Sellars, M. C., Wu, C. J. & Fritsch, E. F. Cancer vaccines: Building a bridge over troubled waters. *Cell* **185**, 2770-2788, doi:https://doi.org/10.1016/j.cell.2022.06.035 (2022).
16 Smart, A. C. *et al.* Intron retention is a source of neoepitopes in cancer. *Nature Biotechnology* **36**, 1056-1058, doi:10.1038/nbt.4239 (2018).
17 Wang, T.-Y. *et al.* A pan-cancer transcriptome analysis of exitron splicing identifies novel cancer driver genes and neoepitopes. *Molecular cell* **81**, 2246-2260.e2212, doi:https://doi.org/10.1016/j.molcel.2021.03.028 (2021).
18 Zhang, M. *et al.* RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nature Communications* **9**, 3919, doi:10.1038/s41467-018-06405-9 (2018).
19 Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217-221, doi:10.1038/nature22991 (2017).
20 Keskin, D. B. *et al.* Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* **565**, 234-239, doi:10.1038/s41586-018-0792-9 (2019).
21 InferCNV of the Trinity CTAT Project (2019).
22 Wood, L. D. *et al.* The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science* **318**, 1108-1113, doi:10.1126/science.1145720 (2007).
23 Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).
24 Dredge, B. K., Stefani, G., Engelhard, C. C. & Darnell, R. B. Nova autoregulation reveals dual functions in neuronal splicing. *The EMBO journal* **24**, 1608-1620 (2005).
25 Guo, J., Jia, J. & Jia, R. PTBP1 and PTBP2 impaired autoregulation of SRSF3 in cancer cells. *Scientific Reports* **5**, 14548, doi:10.1038/srep14548 (2015).
26 Ding, F., Su, C. J., Edmonds, K. K., Liang, G. & Elowitz, M. B. Dynamics and functional roles of splicing factor autoregulation. *Cell Reports* **39**, 110985, doi:https://doi.org/10.1016/j.celrep.2022.110985 (2022).
27 Dietrich, D. R. Toxicological and pathological applications of proliferating cell nuclear antigen (PCNA), a novel endogenous marker for cell proliferation. *Critical reviews in toxicology* **23**, 77-109 (1993).
28 Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome biology* **19**, 40, doi:10.1186/s13059-018-1417-1 (2018).
29 Rice, G. I. *et al.* Mutations in ADAR1 cause Aicardi-Goutieres syndrome associated with a type I interferon signature. *Nature genetics* **44**, 1243-1248 (2012).
30 Galeano, F., Tomaselli, S., Locatelli, F. & Gallo, A. in *Seminars in cell & developmental biology.* 244-250 (Elsevier).

997 31 Zhang, R., Deng, P., Jacobson, D. & Li, J. B. Evolutionary analysis reveals regulatory and
998 functional landscape of coding and non-coding RNA editing. *PLOS Genetics* **13**,
999 e1006563, doi:10.1371/journal.pgen.1006563 (2017).
1000 32 Nishikura, K. Functions and Regulation of RNA Editing by ADAR Deaminases. *Annual*
1001 *Review of Biochemistry* **79**, 321-349, doi:10.1146/annurev-biochem-060208-105251
1002 (2010).
1003 33 Gu, T. *et al.* Canonical A-to-I and C-to-U RNA Editing Is Enriched at 3′ UTRs and
1004 microRNA Target Sites in Multiple Mouse Tissues. *PLOS ONE* **7**, e33720,
1005 doi:10.1371/journal.pone.0033720 (2012).
1006 34 Dong, X. *et al.* CDK13 RNA Over-Editing Mediated by ADAR1 Associates with Poor
1007 Prognosis of Hepatocellular Carcinoma Patients. *Cellular Physiology and Biochemistry* **47**,
1008 2602-2612, doi:10.1159/000491656 (2018).
1009 35 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference*
1010 *on knowledge discovery and data mining.* 785-794.
1011 36 Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nature*
1012 *methods* **14**, 979-982 (2017).
1013 37 Venables, R. S. *et al.* Expression of individual lamins in basal cell carcinomas of the skin.
1014 *British Journal of Cancer* **84**, 512-519, doi:10.1054/bjoc.2000.1632 (2001).
1015 38 Fang, J. *et al.* High KRT 8 expression promotes tumor progression and metastasis of gastric
1016 cancer. *Cancer science* **108**, 178-186 (2017).
1017 39 Golob-Schwarzl, N. *et al.* High Keratin 8/18 Ratio Predicts Aggressive Hepatocellular
1018 Cancer Phenotype. *Translational Oncology* **12**, 256-268,
1019 doi:https://doi.org/10.1016/j.tranon.2018.10.010 (2019).
1020 40 Hendrix, M., Seftor, E. A., Seftor, R. & Trevor, K. T. Experimental co-expression of
1021 vimentin and keratin intermediate filaments in human breast cancer cells results in
1022 phenotypic interconversion and increased invasive behavior. *The American journal of*
1023 *pathology* **150**, 483 (1997).
1024 41 Stühler, K. *et al.* Differential proteome analysis of colon carcinoma cell line SW480 after
1025 reconstitution of the tumour suppressor Smad4. *Analytical and Bioanalytical Chemistry*
1026 **386**, 1603-1612, doi:10.1007/s00216-006-0803-9 (2006).
1027 42 Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine*
1028 **21**, 1350-1356, doi:10.1038/nm.3967 (2015).
1029 43 Ellis, M. J. *et al.* Connecting genomic alterations to cancer biology with proteomics: the
1030 NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer discovery* **3**, 1108-1112
1031 (2013).
1032 44 Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: gold-
1033 standard data classification, open access genotype data and new query tools. *Nucleic acids*
1034 *research* **48**, D783-D788 (2020).
1035 45 Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and
1036 reprogramming. *Nature* **498**, 241-245, doi:10.1038/nature12270 (2013).
1037 46 Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and
1038 Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for*
1039 *Clinicians* **71**, 209-249, doi:https://doi.org/10.3322/caac.21660 (2021).
1040 47 Philpott, M. *et al.* Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq.
1041 *Nature Biotechnology* **39**, 1517-1520, doi:10.1038/s41587-021-00965-w (2021).
1042 48 Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and
1043 transcriptomes. *Nature Methods* **12**, 519-522, doi:10.1038/nmeth.3370 (2015).
1044 49 Al'Khafaji, A. M. *et al.* High-throughput RNA isoform sequencing using programmable
1045 cDNA concatenation. *bioRxiv*, 2021.2010.2001.462818, doi:10.1101/2021.10.01.462818
1046 (2021).

50 ModernaTX, I., Sharp, M. & LLC, D. (https://ClinicalTrials.gov/show/NCT03897881, 2019).
51 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821, doi:10.1016/j.cell.2019.05.031 (2019).
52 Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).
53 Pacific Biosciences. IsoSeq v3. (2018).
54 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
55 cDNA Cupcake (2017).
56 Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Research* **9** (2020).
57 Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome research* **28**, 396-411, doi:10.1101/gr.222976.117 (2018).
58 Abugessaisa, I. *et al.* refTSS: a reference data set for human and mouse transcription start sites. *J Mol Biol* **431**, 2407-2422 (2019).
59 Herrmann, C. J. *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Research* **48**, D174-D179, doi:10.1093/nar/gkz918 (2020).
60 Wang, R., Zheng, D., Yehia, G. & Tian, B. A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome research* **28**, 1427-1441 (2018).
61 Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic acids research* **47**, W199-W205 (2019).
62 Ramaswami, G. & Li, J. B. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research* **42**, D109-D113, doi:10.1093/nar/gkt996 (2014).
63 Kiran, A. M., O'Mahony, J. J., Sanjeev, K. & Baranov, P. V. Darned in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic acids research* **41**, D258-D261 (2012).
64 Tan, M. H. *et al.* Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550**, 249-254 (2017).
65 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
66 Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17-23 (1950).
67 Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols* **11**, 1650, doi:10.1038/nprot.2016.095 (2016).
68 Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic acids research* **43**, e78-e78 (2015).
69 The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic acids research* **49**, D480-D489 (2021).
70 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1-9 (2009).
71 Bateman, A. *et al.* The Pfam protein families database. *Nucleic acids research* **32**, D138-D141 (2004).
72 Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387-3395 (2017).
73 Bulik-Sullivan, B. *et al.* Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nature Biotechnology* **37**, 55-63, doi:10.1038/nbt.4313 (2019).

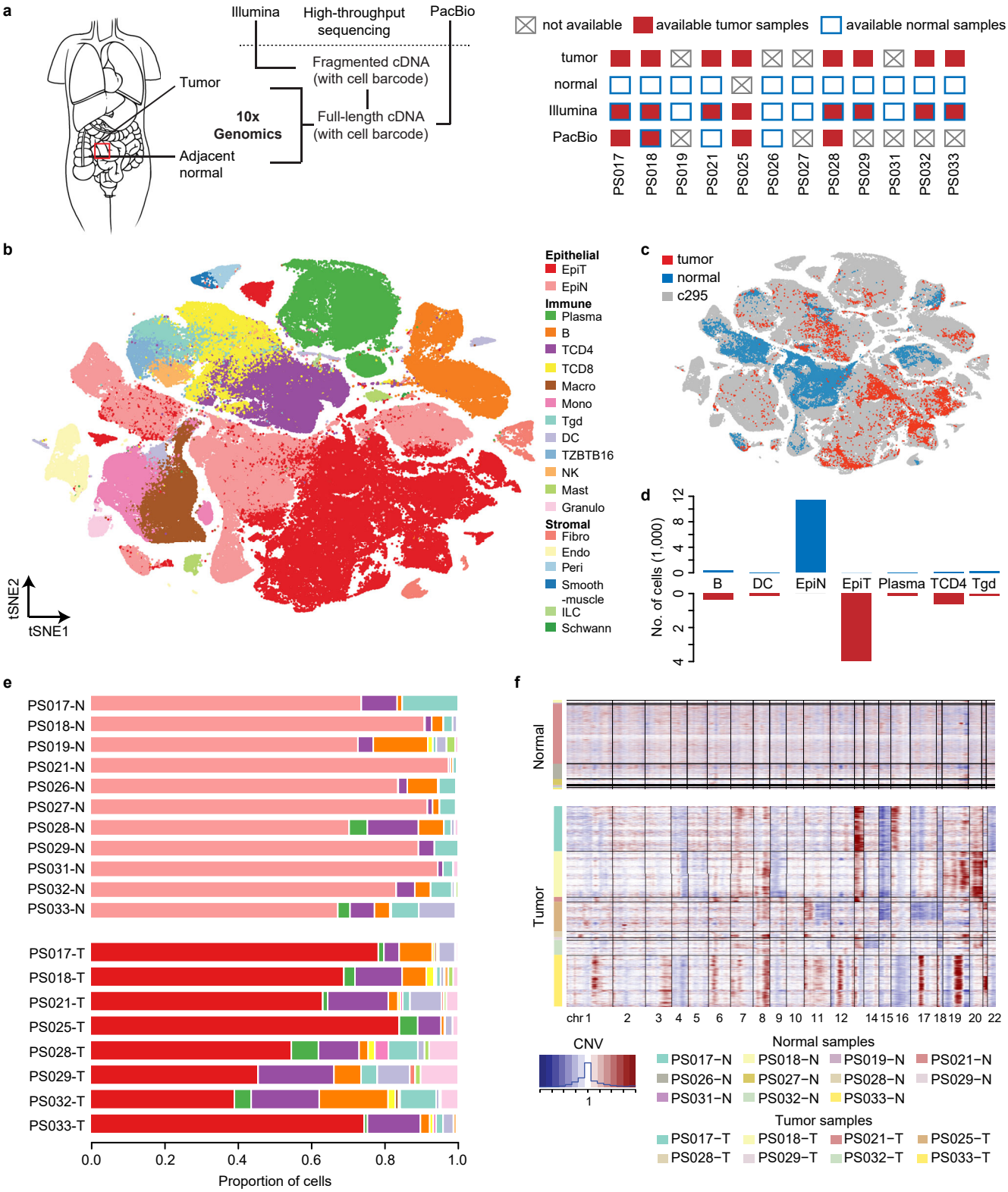
74 Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS* **13**, 22-24, doi:https://doi.org/10.1002/pmic.201200439 (2013).

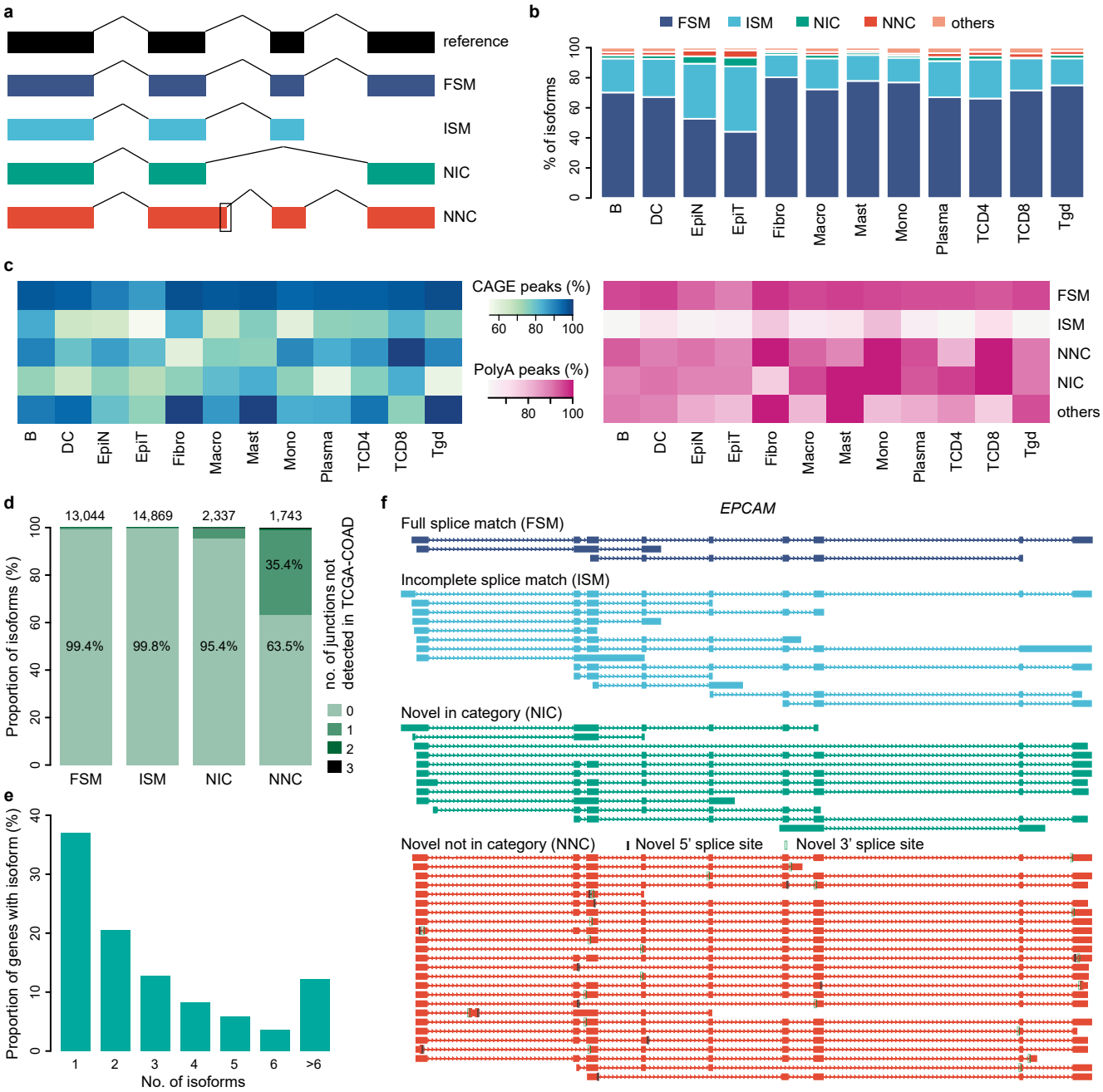
75 Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4**, 923-925, doi:10.1038/nmeth1113 (2007).

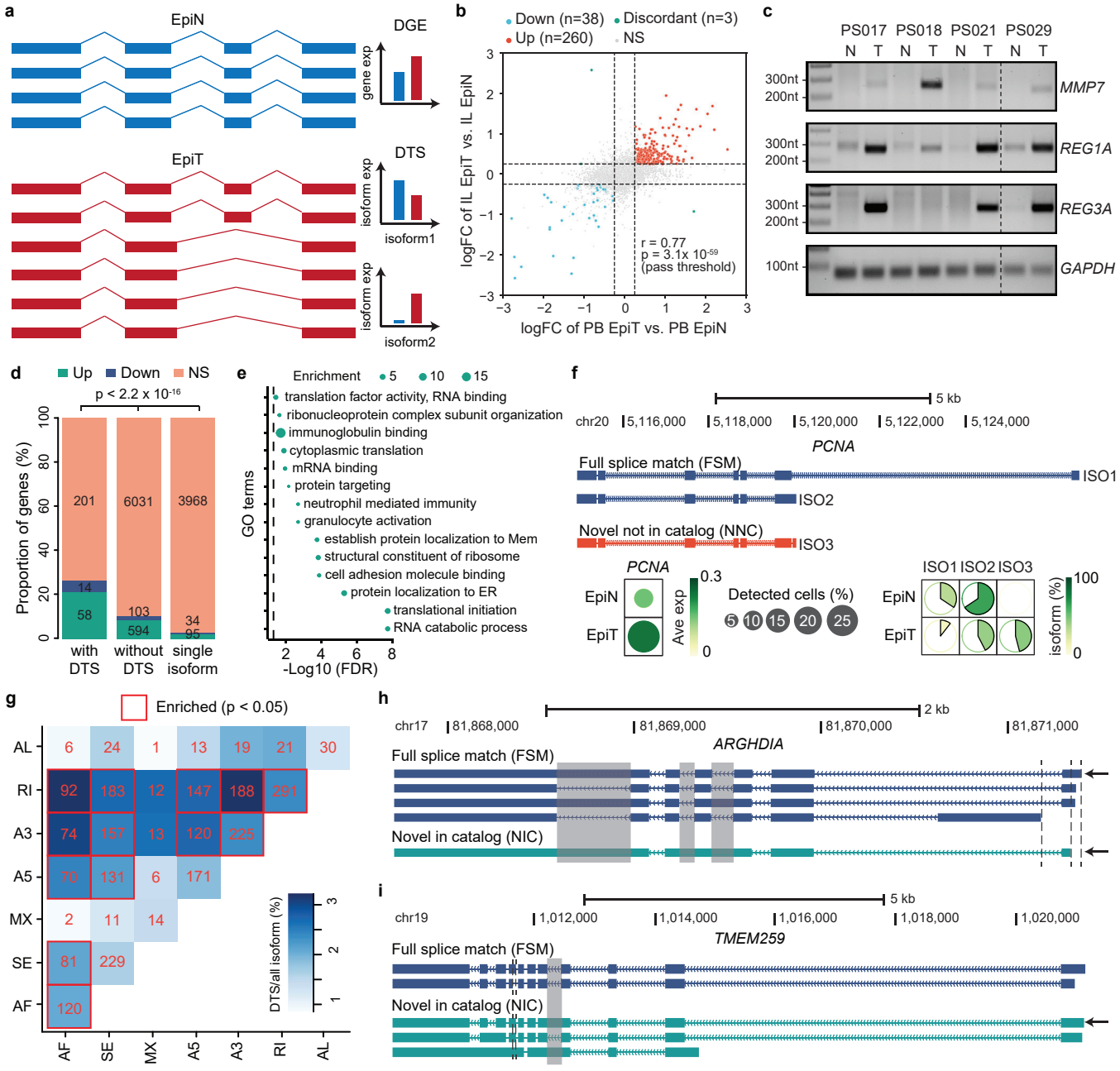
76 Jurtz, V. *et al.* NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *The Journal of Immunology* **199**, 3360-3368 (2017).

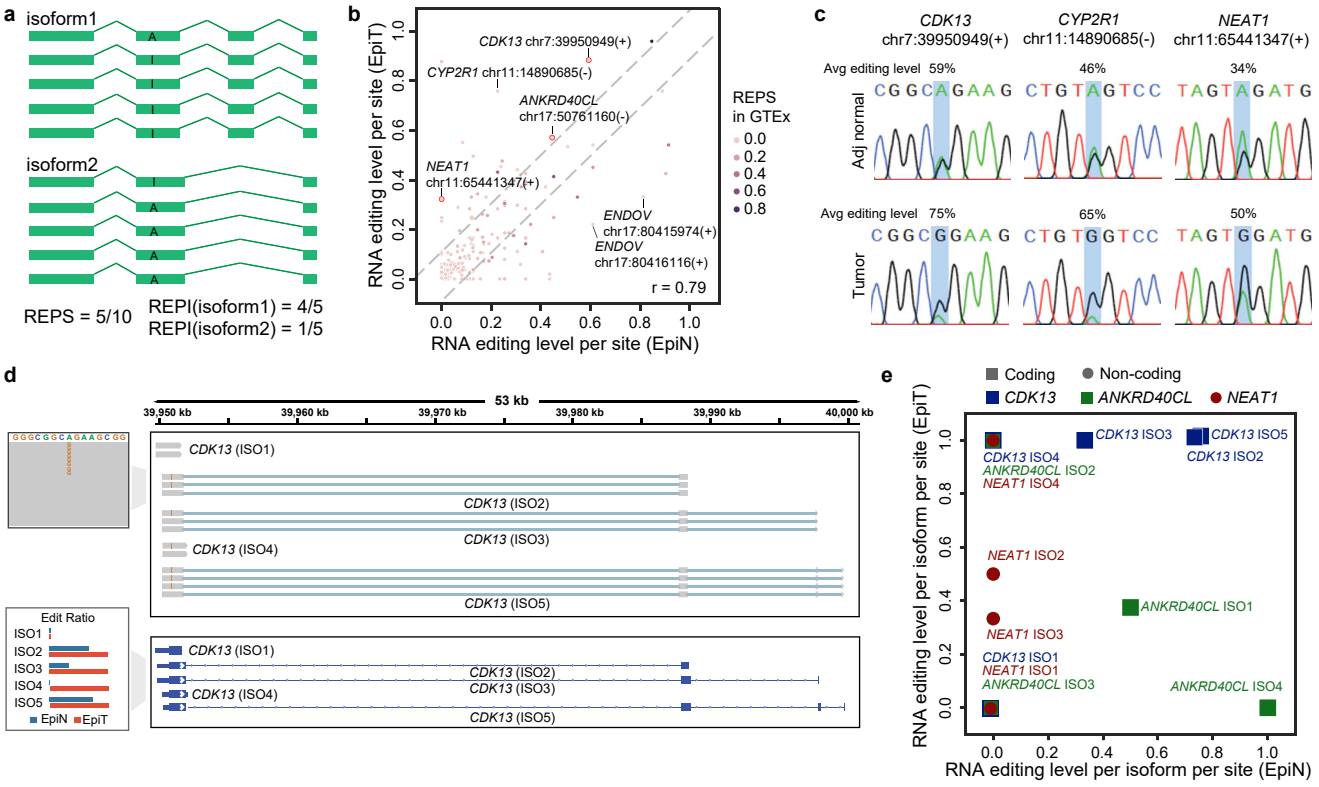
77 Nemhauser, G. L., Wolsey, L. A. & Fisher, M. L. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming* **14**, 265-294 (1978).

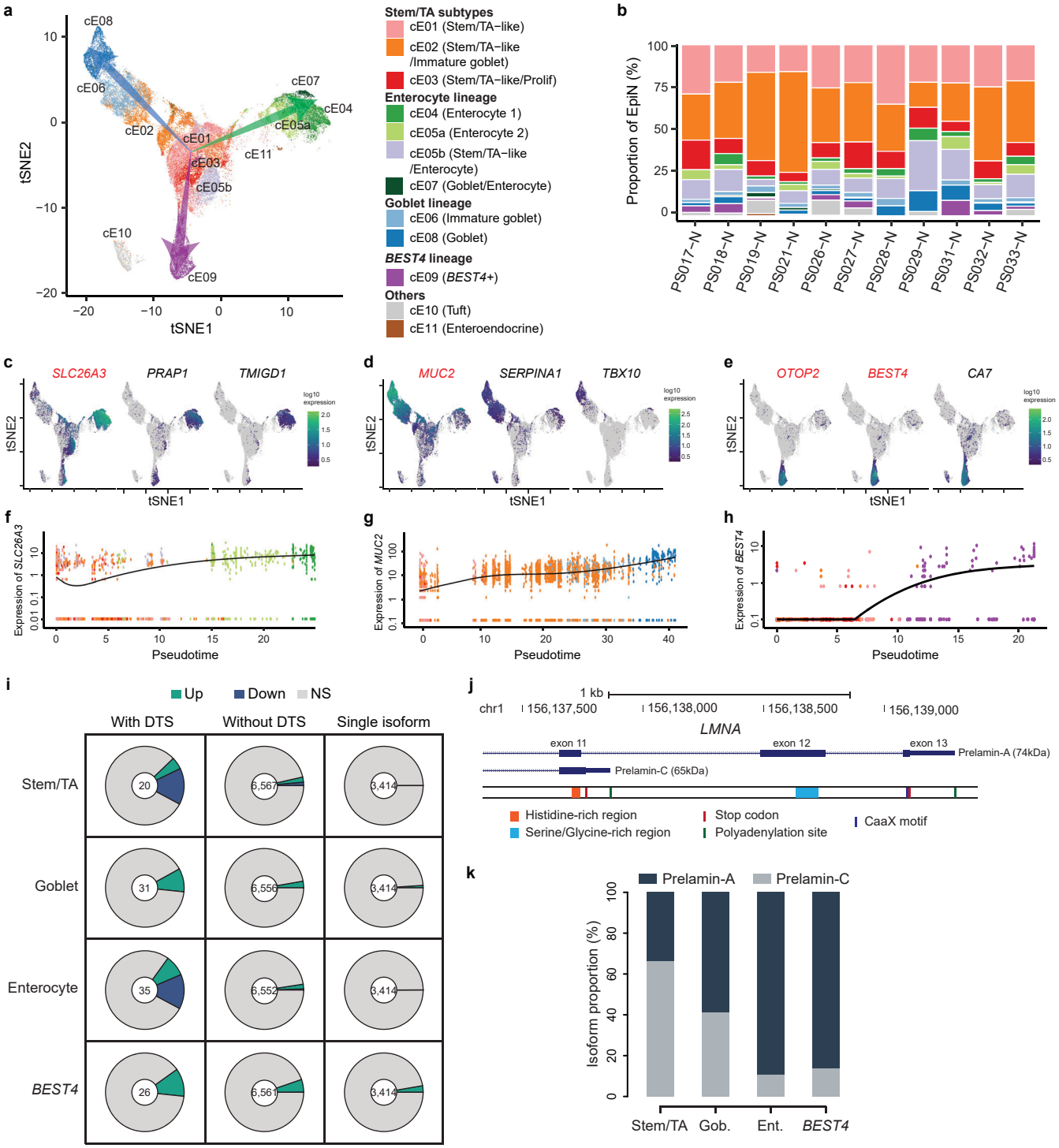
78 Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e814, doi:https://doi.org/10.1016/j.immuni.2018.03.023 (2018).

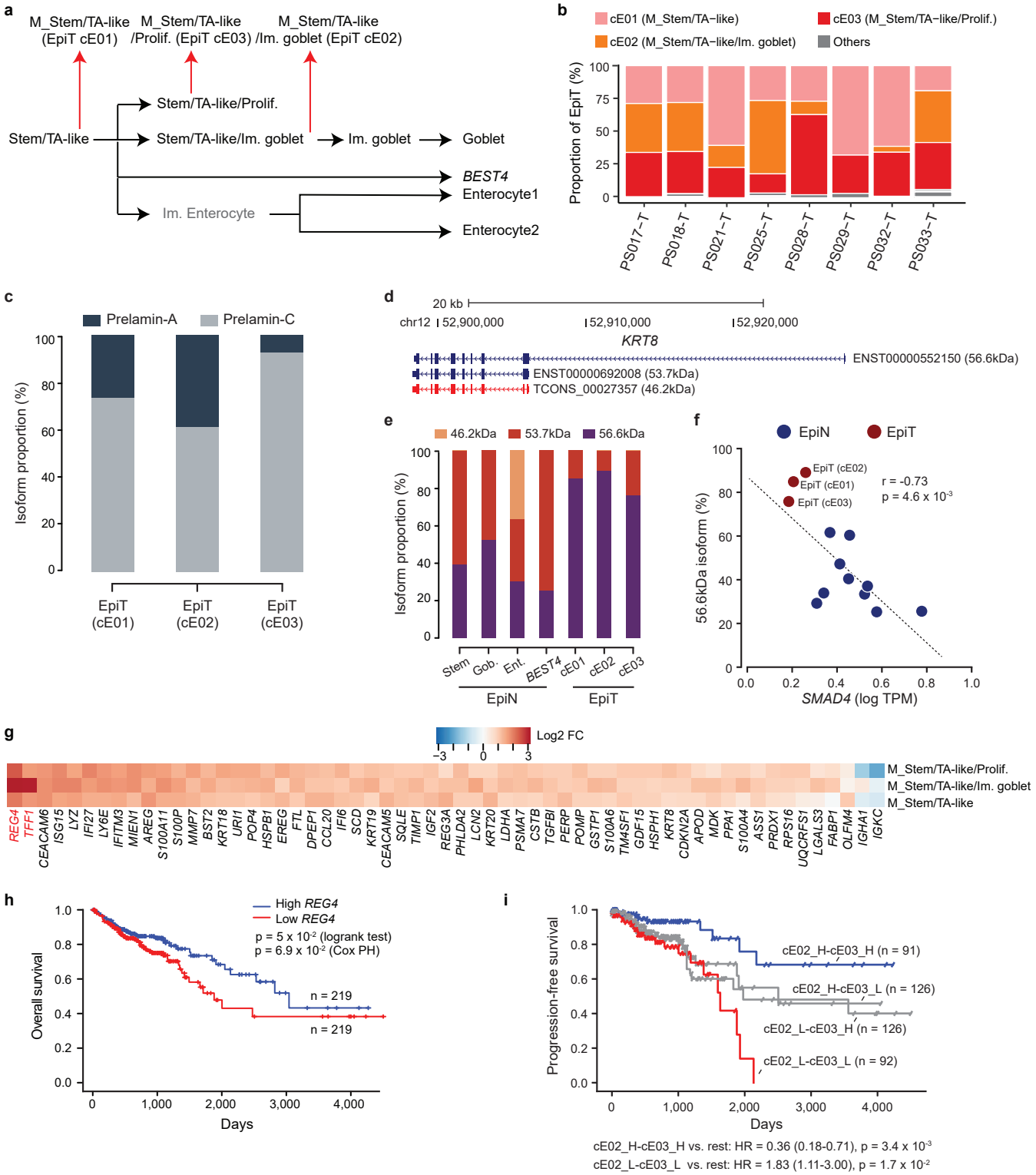

a
Illumina
High-throughput sequencing
PacBio
Fragmented cDNA (with cell barcode)
Tumor
10x Genomics
Full-length cDNA (with cell barcode)
Adjacent normal
not available
available tumor samples
available normal samples
tumor
normal
Illumina
PacBio
PS017
PS018
PS019
PS021
PS025
PS026
PS027
PS028
PS029
PS031
PS032
PS033
b
tSNE2
tSNE1
Epithelial
EpiT
EpiN
Immune
Plasma
B
TCD4
TCD8
Macro
Mono
Tgd
DC
TZBTB16
NK
Mast
Granulo
Stromal
Fibro
Endo
Peri
Smooth -muscle
ILC
Schwann
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
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of cells
f
Normal
Tumor
chr 1
2
3
4
5
6
7
8
9
10
11
12
14
15
16
17
18
19
20
22
CNV
1
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

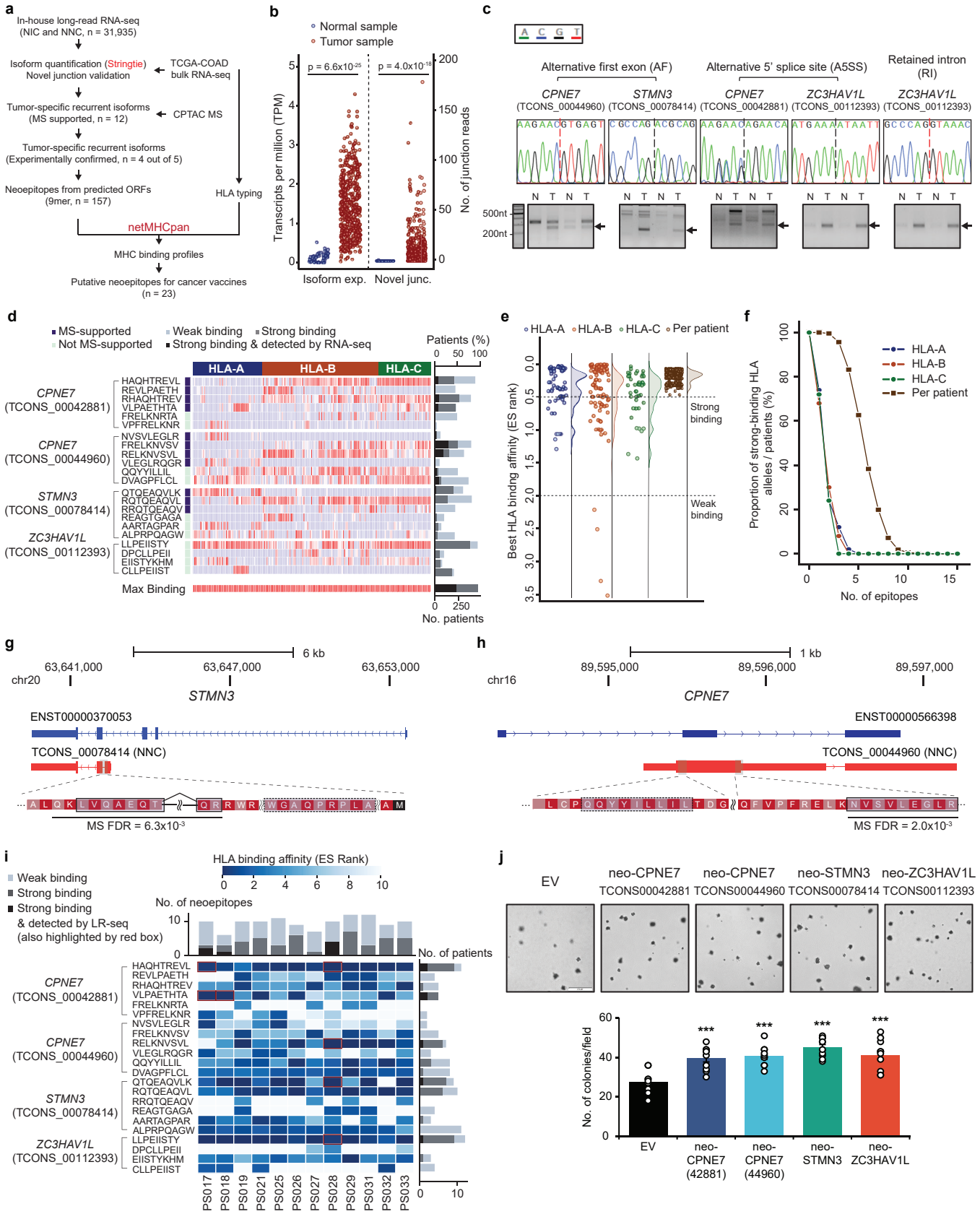

a
reference
FSM
ISM
NIC
NNC
b
FSM
ISM
NIC
NNC
others
% of isoforms
0
20
40
60
80
100
B
DC
EpiN
EpiT
Fibro
Macro
Mast
Mono
Plasma
TCD4
TCD8
Tgd
c
CAGE peaks (%)
60
80
100
PolyA peaks (%)
80
100
FSM
ISM
NNC
NIC
others
d
13,044
14,869
2,337
1,743
Proportion of isoforms (%)
99.4%
99.8%
95.4%
63.5%
35.4%
FSM
ISM
NIC
NNC
no. of junctions not detected in TCGA-COAD
0
1
2
3
e
Proportion of genes 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


a
EpiN
DGE
gene exp
EpiT
DTS
isoform exp
isoform1
isoform exp
isoform2
b
Down (n=38)
Discordant (n=3)
Up (n=260)
NS
logFC of IL EpiT vs. IL EpiN
logFC of PB EpiT vs. PB EpiN
r = 0.77
p = 3.1x 10-59
(pass threshold)
c
PS017 PS018 PS021 PS029
N T N T N T N T
300nt
200nt
100nt
MMP7
REG1A
REG3A
GAPDH
d
Up Down NS
p < 2.2 x 10-16
Proportion of genes (%)
201
6031
3968
14
58
103
594
34
95
with DTS
without DTS
single isoform
e
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)
f
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
isoform (%)
100
0
g
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 70 131 6 171
MX 2 11 14
SE 81 229
AF 120
AF SE MX A5 A3 RI AL
DTS/all isoform (%)
h
2 kb
chr17 81,868,000 81,869,000 81,870,000 81,871,000
ARGHDIA
Full splice match (FSM)
Novel in catalog (NIC)
i
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)


a
isoform1
isoform2
REPS = 5/10
REPI(isoform1) = 4/5
REPI(isoform2) = 1/5
b
RNA editing level per site (EpiT)
RNA editing level per site (EpiN)
CDK13 chr7:39950949(+)
CYP2R1 chr11:14890685(-)
ANKRD40CL chr17:50761160(-)
NEAT1 chr11:65441347(+)
ENDOV chr17:80415974(+)
ENDOV chr17:80416116(+)
r = 0.79
REPS in GTEx
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
RNA editing level per isoform per site (EpiT)
RNA editing level per isoform per site (EpiN)
CDK13 ISO4
ANKRD40CL ISO2
NEAT1 ISO4
CDK13 ISO3
CDK13 ISO5
CDK13 ISO2
NEAT1 ISO2
ANKRD40CL ISO1
NEAT1 ISO3
CDK13 ISO1
NEAT1 ISO1
ANKRD40CL ISO3
ANKRD40CL ISO4
f
Up
Down
ns
RNA editing regulation per isoform (%)
RNA editing regulation per site
g
No. Events
REPS
No. of isoforms with editing events
No. of total isoforms
CDK13
ANKRD40CL
NEAT1


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
d
MUC2
SERPINA1
TBX10
e
OTOP2
BEST4
CA7
log10 expression
f
Expression of SLC26A3
Pseudotime
g
Expression of MUC2
Pseudotime
h
Expression of BEST4
Pseudotime
i
Up
Down
NS
With DTS
Without DTS
Single isoform
Stem/TA
Goblet
Enterocyte
BEST4
20
6,567
3,414
31
6,556
3,414
35
6,552
3,414
26
6,561
3,414
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


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)
cE03 (M_Stem/TA−like/Prolif.)
cE02 (M_Stem/TA−like/Im. goblet)
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
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


a
In-house long-read RNA-seq (NIC and NNC, n = 31,935)
Isoform quantification (Stringtie)
Novel junction validation
TCGA-COAD bulk RNA-seq
Tumor-specific recurrent isoforms (MS supported, n = 12)
CPTAC MS
Tumor-specific recurrent isoforms (Experimentally confirmed, n = 4 out of 5)
Neoepitopes from predicted ORFs (9mer, n = 157)
HLA typing
netMHCpan
MHC binding profiles
Putative neoepitopes for cancer vaccines (n = 23)
b
Normal sample
Tumor sample
p = 6.6x10-25
p = 4.0x10-18
Transcripts per million (TPM)
No. of junction reads
Isoform exp.
Novel junc.
c
Alternative first exon (AF)
Alternative 5' splice site (A5SS)
Retained intron (RI)
CPNE7 (TCONS_00044960)
STMN3 (TCONS_00078414)
CPNE7 (TCONS_00042881)
ZC3HAV1L (TCONS_00112393)
ZC3HAV1L (TCONS_00112393)
N T N T
500nt
200nt
d
MS-supported
Not MS-supported
Weak binding
Strong binding
Strong binding & detected by RNA-seq
Patients (%)
HLA-A
HLA-B
HLA-C
CPNE7 (TCONS_00042881)
CPNE7 (TCONS_00044960)
STMN3 (TCONS_00078414)
ZC3HAV1L (TCONS_00112393)
Max Binding
No. patients
e
HLA-A
HLA-B
HLA-C
Per patient
Best HLA binding affinity (ES rank)
Strong binding
Weak binding
f
HLA-A
HLA-B
HLA-C
Per patient
Proportion of strong-binding HLA alleles / patients (%)
No. of epitopes
g
6 kb
chr20
63,641,000
63,647,000
63,653,000
STMN3
ENST00000370053
TCONS_00078414 (NNC)
MS FDR = 6.3x10-3
h
1 kb
chr16
89,595,000
89,596,000
89,597,000
CPNE7
ENST00000566398
TCONS_00044960 (NNC)
MS FDR = 2.0x10-3
i
Weak binding
Strong binding
Strong binding & detected by LR-seq (also highlighted by red box)
HLA binding affinity (ES Rank)
No. of neoepitopes
No. of patients
CPNE7 (TCONS_00042881)
CPNE7 (TCONS_00044960)
STMN3 (TCONS_00078414)
ZC3HAV1L (TCONS_00112393)
PS017
PS018
PS019
PS021
PS025
PS026
PS027
PS028
PS029
PS031
PS032
PS033
j
EV
neo-CPNE7 TCONS00042881
neo-CPNE7 TCONS00044960
neo-STMN3 TCONS00078414
neo-ZC3HAV1L TCONS00112393
No. of colonies/field
EV
neo-CPNE7 (42881)
neo-CPNE7 (44960)
neo-STMN3
neo-ZC3HAV1L