# Somatic repetitive element insertions Define New Biomarkers in Pan-cancer genome

Lin Xia[1#], Kailing Tu[1#], Huan Wang[1#], Xuyan Liu[2], Yahui Liu[2], Qilin Zhang[2], Ziting Feng[2], Chunyan Yu[1], Maolin Li[1], Yan Huang[1], Jing Wang[1], Ranlei Wei[1], Tianfu Zeng[2], Xuenan Pi[1], Chen Ling[3], Qianwen Zhang[3], Yixin Qiao[4], Shengzhong Hou[5], Zhuoyuan Zhang[6,7], Ailin Wei[8], Haotian Liao[9], Wenhao Guo[10], Jifeng Liu[4,11], Zhong Wu[5], Zhoufeng Wang[12,13], Ang Li[5], Weiming Li[12,13], Dan Xie[1,2*]

1. Laboratory of Omics Technology and Bioinformatics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
2. State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan, China
3. Department of Gynecology and Obstetrics, Key Laboratory of Birth Defects and Related Diseases of Women and Children, Ministry of Education, West China Second Hospital, Sichuan University, Chengdu, China
4. Department of Otolaryngology Head and Neck Surgery, West China Hospital, Sichuan University, Chengdu, China.
5. Division of Pancreatic Surgery, Department of General Surgery, West China Hospital, Sichuan University
6. Department of Head and Neck Cancer Surgery, West China School of Stomatology, Sichuan University, Chengdu, Sichuan, China
7. State Key Laboratory of Oral Diseases, West China School of Stomatology, Sichuan University, Chengdu, Sichuan, China
8. Guangan People's Hosital, Guangan, Sichuan, China
9. Division of Liver Surgery, Department of General Surgery and Laboratory of Liver Surgery, and State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan, China.

10. Department of Abdominal Oncology, Cancer Center and State Key Laboratory of Biotherapy, West China Hospital, West China Medical School, Sichuan University,Chengdu, China.
11. Deep Underground Space Medical Center, West China Hospital, Sichuan University, Chengdu, China.
12. Institute of Respiratory Health, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
13. Precision Medicine Key Laboratory of Sichuan Province, West China Hospital, Sichuan University, Chengdu, Sichuan, China

* Corresponding authors. Email: danxie@scu.edu.cn

# These authors contributed equally

# Abstract

Repetitive elements insertions (REIs), including tandem repeat insertions (TRIs) and transposable element insertions (TEIs), constitute >90% of large-scale (>50bp) somatic insertions and are linked to diseases and cancer biomarkers. However, the genomic landscape of large-scale REIs in pan-cancer has not been well-characterized due to short-read sequencing limitations. Here, we detected somatic REIs in 325 cancer genomes across 12 cancer types using long-read whole-genome sequencing and customized algorithms. The diverse sequence patterns of somatic TRIs and the insertion mechanisms of somatic TEIs were characterized across different cancer types. We identified 152 high-frequency somatic TRIs, with 10 showing pan-cancer characteristics, including TRIs in *SHROOM2, PALMD*, and the enhancer of *PTPRZ1*. The MHC class II gene cluster exhibited the highest somatic TEI abundance, affecting ~40% tumors. Adenocarcinomas and squamous cell carcinomas exhibited differing TEI distributions within this region. Our study highlights REIs as important but under-explored markers for personalized cancer diagnosis and immunotherapy.

# Introduction

Cancer is formidable in its vast diversity, although all cancers share common hallmarks[1,2]. Understanding both the commonalities and variations among cancer types is crucial for uncovering the fundamental mechanisms of oncogenesis and advancing precision medicine. Pan-cancer studies, which aim at identifying shared genomic patterns and alterations driving cancer progression across cancers [3-7], not only enhances our understanding of cancer biology, but also holds considerable promise for advancing early diagnosis and developing personalized treatment strategies. These efforts have significantly expanded the knowledge of cancer driver genes, with *TP53*, *PIK3CA*, *KRAS*, *PTEN*, and *ARID1A* among the most frequently mutated genes across various cancers [8]. Furthermore, pan-cancer studies have uncovered new non-coding driver genes [9,10], mutational signatures [11], and evolutionary patterns [12-14], as well as their downstream effects on gene expression [15] and cancer-specific characteristics [16,17]. Despite these advances, many findings have yet to translate into clinical benefits, highlighting the pressing need for the identification of other critical genetic variants in cancer progression.

Previous pan-cancer research has predominantly focused on small-scale mutations, such as single nucleotide variations (SNVs) and small insertions/deletions (INDELs, < 50bp). Structural variations (SVs), which impact a broader range of nucleotide change [18], have increasingly garnered attention due to their significant role in cancer progression [19]. With short-read sequencing (SRS) data, researchers have identified various genomic SVs, including whole-genome duplications [20], chromothripsis [21], chromosomal instability [17], and complex structural variants [22], all associated with poorer prognoses. Nevertheless, large-scale insertions (>50 bp) in tumors remain less explored, particularly for repetitive element insertions (REIs), including tandem repeat insertions (TRIs) and transposable element insertions (TEIs) [23]. Previous population studies and tumor analyses indicate that over 80% of large-scale insertions are repetitive element insertions (REIs) [24-26]. Both TRIs [27,28] and TEIs [29-32] are drawing increasing attention as potential diagnostic and therapeutic target for various types of cancer. Despite extensive mapping efforts to characterize TRI

88 and TEI events in cancer genomes [27,28,33-35], only a limited number of recurrent somatic
89 REIs have been identified. The short read lengths of SRS technologies have hindered the
90 in-depth exploration of REI variation patterns in cancer genomes [23]. These challenges
91 underscore the urgent need for advancements in sequencing technologies and
92 bioinformatics approaches to enhance the detection and analysis of somatic REIs in
93 cancer.
94
95 The continuous advancement of third-generation sequencing (TGS) technology,
96 particularly Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), has
97 significantly enhanced long-read sequencing (LRS). Technology of ONT produces reads
98 exceeding 10 kb, with some extending to millions of bases, enabling the complete capture
99 of large-scale SVs and their flanking regions in a single read. This improvement enhanced
100 the sensitivity and accuracy of SV detection [26,36-38]. LRS has enabled researchers to
101 accurately map TR alternations across populations, revealing complex TR alternations
102 that were previously undetectable by SRS [39,40]. LRS were also used in population studies
103 to uncover the widespread presence of short interspersed nuclear element (SINE)
104 insertions, which are more difficult to capture than long interspersed nuclear element
105 (LINE) insertions typically identified by SRS [36,37]. However, most population studies have
106 focused on germline genomic alterations, leaving the study of somatic REIs in tumor
107 genome largely unexplored [41-43]. While some studies have applied long-read sequencing
108 to examine structural variations in tumors, the lack of large tumor sample cohorts and
109 suitable bioinformatics tools has limited the comprehensive analyses.
110
111 In this study, we conducted long-read nanopore whole genome sequencing (LRS WGS)
112 on 325 paired tumor and blood samples from 12 cancer types. With specialized algorithms
113 developed, we identified 39,661 somatic REIs. The diversity of somatic TRIs sequence
114 patterns and variable TEI insertion mechanisms across cancer types were elucidated. In
115 total, we identified 152 high-frequency somatic TRIs, where 10/152 are shared in
116 diverse cancer types, including somatic TRI events were located within the intron of
117 *SHROOM2* and *PALMD*, and in the distal enhancer region upstream of *PTPRZ1*.

118 Regarding TEIs, approximately 40% of tumors harbored somatic TEIs in the MHC class II
119 gene cluster, with adenocarcinomas showing greater TEI enrichment than squamous cell
120 carcinomas. Notably, adenocarcinoma cells with a higher TEI burden within this region
121 showed significantly enhanced T cell activation compared to those with a lower burden,
122 indicating that TEIs may contribute to the regulation of tumor immune responses. Overall,
123 this study provides a valuable resource to the scientific and clinical community and a
124 paradigm for character the somatic genomic alternations in pan-cancer genomes using
125 LRS.

# Results

## Section 1: Specimens, Tumor Types and Dataset

This pan-cancer study encompassed 325 tumor samples and 300 paired blood samples from 12 distinct cancer types (Figure 1a). Solid tumor types were from gynecologic (ovarian [OV], uterine corpus endometrial carcinoma [UCEC], cervical squamous cell carcinoma [CCA]), core gastrointestinal (stomach adenocarcinoma [GC], colon adenocarcinoma [IC]), developmental gastrointestinal (liver hepatocellular carcinoma [HCC], pancreatic adenocarcinoma [PAAD]), head and neck (Oral squamous cell carcinoma [OSCC], Laryngocarcinom [LC]), thoracic (lung adenocarcinoma [LUAD], lung squamous cell carcinoma [LUSC]), and central nervous system cancers (glioblastoma multiforme [GBM]). 4, 2, 27 and 2 multiple synchronous tumors were obtained from 9 patients with LUAD (2), GBM (1), HCC (5) and PAAD (1), separately (Table S1). All patients provided written informed consent for genomic studies approved by local institutional review boards.

Whole-genome sequencing (WGS) was performed on all tumor and paired blood samples using PromethION nanopore sequencer, yielding long sequencing reads (Table S2). A total of 45.73 terabases (Tb) of long clean reads were generated, averaging 93.03 gigabases (Gb) per tumor sample and 51.63 Gb per paired blood sample. The mean read depth reached 30 X (13 X~46 X) for tumor samples and 17 X (3 X~43 X) for paired blood samples (Figure S1a), with mean read N50 length across all samples was 24.9 kbp (4.5 kbp~43.1 kbp) for clean long reads (Figure S1b). Due to an update in ONT technology during data acquisition, two versions of sequencing chips were employed in this study. Specifically, 83.06% (510/614) of the LRS WGS data were generated using the R9 chips, while the remaining data were obtained using the R10 chips (Figure S1c). Notably, the data produced with the R10 chips exhibited significantly higher quality compared to that generated with the R9 chips (Wilcoxon test, $p < 2.2e-16$) (Figure S1d). Additionally, RNA-seq datasets were generated for 89 NSCLC samples on an NGS platform [25] (Table S3).

155 ## Section 2: Accurate identification of somatic REIs in
156 ## pan-cancer genomes

157 In our previous research, nearly 90% somatic insertions in NSCLC genomes were
158 identified as somatic TRIs and TEIs, where somatic TRIs were predominantly located
159 within annotated simple repeat regions (77.8%) [25]. The high repetitiveness of these
160 sequences complicates alignment accuracy, often resulting in a high false-positive rate. To
161 address this issue, we designed a more precise algorithm for detecting somatic TRIs
162 across tumor genomes (STAR Methods) (Figure 1b and S2a). First, cleaned LRS reads
163 were mapped to the hg38 reference genome using Minimap2 software (version 2.17) [44]],
164 achieving an average mapping rate of 99.91% (93.07%-100%). Next, widely used SV
165 detection softwares were applied to identify loci harboring somatic TRIs. Specifically,
166 Sniffles2 (version 2.0.6) [45] was employed to detect INSs within tandem repeat regions
167 annotated by RepeatMasker in the hg38 reference genome. Repeat expansions (REs) in
168 tumor genomes were then identified using Straglr (version 1.4.1) [46]. Regions identified by
169 either tool were combined to form a final list of candidate somatic tandem repeat insertion
170 windows for further analysis. For each paired tumor-blood samples, somatic TRIs within
171 these candidate windows were detected based on the clustering of reads according to
172 editing distance and the TDScope algorithm [47] (Figure S2a, STAR Methods).
173
174 Somatic transposable element insertions (TEIs) were identified by optimizing an in-house
175 pipeline for somatic structural variant (SV) detection [25] (Figure 1c and S2b). First, tumor
176 sample alignments were merged with those of the paired blood sample. Next, large-scale
177 INSs were detected using Sniffles2 (version 2.0.6) [45]. A somatic INS was considered valid
178 if more than 3 supporting reads from the tumor sample, but no read from the paired blood
179 sample (STAR Methods). Next, Iris software (version 1.0.5)[48] was used to correct the
180 sequences and breakpoints of the candidate somatic INSs. To false positives due to
181 reference-individual genome differences, a custom pipeline was developed. The
182 reference genome at each INS loci was polished using Racon (version 1.4.16)[49], based on
183 alignments from the paired blood samples. Tumor read were then realigned to the

184 polished reference, and those with more than 3supporting reads were retained as
185 candidate somatic INSs. Finally, consensus sequences of somatic INSs were categorized
186 using RepeatMasker (version 4.1.2)[50], Tandem Repeat Finder (version 4.10.0) [51], and
187 Sdust (version 0.1)[52] (STAR Methods). Somatic INSs with consensus sequences
188 containing TE segments were classified as somatic TEIs (STAR Methods).
189
190 A total of 16,382 somatic TRIs (Table S4) were identified across 325 tumor genomes, with
191 counts ranging from 1 to 849, and an average of 50. Similarly, 23,279 somatic TEIs (Table
192 S5) were identified across 325 tumor genomes, with counts ranging from 0 to 1,776, and
193 an average of 72 (Figure 1d and S2c). No correlation was identified between tumor purity
194 (mean 46.33%, Table S1) and the count of somatic TRIs or somatic TEIs (Figure S3a-S3d,
195 STAR Methods). Squamous carcinomas (SCC) exhibited a significantly higher burden of
196 somatic TEIs compared to other cancer types (Welch two sample t-test, p value=0.0014
197 with Other, p value=1.6e-05 with Adenocarcinoma, p value=2.9e-07 with high burden)
198 (Figure S3e and S3f). Notably, 75.88% (3351/4416) of the somatic TEIs were specific to
199 SCC. However, only 43of the 3351 of the SCC-specific somatic TEIs were shared
200 between two SCC samples, suggesting that while TEs are more active in SCC genomes,
201 but SCC-specific TEIs were mostly low-frequency mutations. The distribution of somatic
202 TRI and TEI counts across various cancer types did not show significant differences
203 (Figure 1c, S3g).
204
205 Interestingly, a subset of 11 non-SCC tumor samples exhibited markedly higher numbers
206 of both somatic TRIs (ranging from 592 to 849) and somatic TEIs (ranging from 621 to
207 1776) compared to the remaining tumors (TRIs: ranging from 1 to 139, TEIs: ranging from
208 0 to 279) (Figure 1d and S2c). This group, termed high-burden samples, included 2 HCC
209 samples, 1 GBM sample, 2 GC samples, and 6 LUAD samples. Using the Genomic
210 Regions Enrichment of Annotations Tool (GREAT)[53], the high-burden-specific somatic
211 TRIs (Table S6) were linked to distal cis-regulatory effects on genes involved in cell cycle
212 regulation, telomere length, and chromosome localization (Figure S3h). Additionally,
213 transcriptomic analysis of available NSCLC NGS RNA-seq data revealed that high-burden

214 LUAD samples exhibited significantly lower expression of DNA repair-related genes,
215 including *ERCC3*, *RAD51*, and *XRCC4* (Wilcoxon rank test, $p < 0.05$,
216 $|\log2(\text{FoldChange})| > 1$), compared to other LUAD samples (Figure S3i). Reduced
217 expression of these DNA repair genes may lead to the accumulation of unrepaired DNA
218 damage, contributing to genomic instability [54-56].
219
220 Somatic TRIs were significantly enriched in subtelomeric regions (5.308-fold compared to
221 the rest region of chromosome arm in average, Student t-test, p-value = 9.81e-14),
222 whereas no enrichment of somatic TEIs was detected in subtelomeric regions (Student
223 t-test, p-value = 0.6275) (Figure S4a, Table S7). Similar to the previous LRS study [57], the
224 long arms of human chromosomes were more likely exhibited a broader zone of somatic
225 TRIs enrichment (5–10 Mbp) compared to the short arms (Figure S4b). Similar to a large
226 healthy Chinese cohort genomic study [58] and a recent TE recombination study [59], somatic
227 TEIs were frequently enriched in centromeric and pericentromeric regions, particularly on
228 chromosomes 20, 4, and 9 (Figure S4c). Notably, solo LINE insertions, solo SINE
229 insertions, and complex TEIs containing LTR segments exhibited the most pronounced
230 enrichment in these regions. (Figure S4c). In contrast, a comparable analysis of TEs in
231 the reference genome and all somatic TEIs revealed no such patterns for solo LINE
232 insertions.
233
234 To identify hotspot regions of somatic TRIs and TEIs, the genome was segmented into
235 non-overlapping 1 Mb windows, and their distribution was investigated (Figure 1e).
236 Hotspot regions containing genes that overlapped or were adjacent to somatic TRIs and
237 TEIs were identified. These included *MUC16*, associated with various cancers [60-62] and
238 widely used as a biomarker for ovarian cancer (CA125) [63] *STK24*, a prognostic marker for
239 HCC, LUAD, PAAD, and KIRC ($p < 0.001$, log-rank test), involved in apoptosis regulation
240 and cell migration[64,65]; *NBPF1*, part of the neuroblastoma breakpoint family, linked to
241 genomic instability in neuroblastoma and gliomas[66,67]; and *CELF2*, a prognostic marker
242 for HNSC and KIRC ($p < 0.001$, log-rank test), implicated in apoptosis and cancer
243 progression [68]. Genes such as *ZNF716* and *ZNF717*, which encoding zinc finger proteins

244 involved in transcriptional regulation and signaling [69]; *PTPN7* and *PTPRZ1*, which were
245 members of the protein tyrosine phosphatase family [70,71], were also identified as highly
246 recurrent but whose role in cancer remain to be fully characterized. In addition to genes
247 with established roles in tumor progression, we also identified a number of genes with
248 high-frequency somatic REI events across pan-cancer samples, including those involved
249 in cytoskeleton regulation (e.g., *SHROOM2*, *PALMD*, and *CROCC*), suggesting their
250 potential as biomarkers.
251
252 We next examined the enrichment of somatic TRIs and somatic TEIs in gene and
253 regulatory regions by comparing the real dataset ('O', observed) to datasets representing
254 the genomic coordinates of gene and regulatory regions ('E', expected) (Figure S4d).
255 52.5% somatic TRIs were located within gene regions, while 47.9% somatic TEIs were
256 detected in these regions. Notably, somatic TRIs were significantly enriched in candidate
257 cis-regulatory elements predicted by ENCODE (cCRE)) (log2(O/E) ranging from 0.27 to
258 2.07, binomial test, FDR < 0.05), suggesting their potential roles in chromosomal structure
259 and genome regulation. In contrast, somatic TEIs were enriched in regions proximal to
260 genebody (±2.5 kb, log2(O/E) ranging from 0.1 to 0.2, binomial test, FDR < 0.05),
261 implicating them in noncanonical splicing events, as previously reported [31].
262

## 263 Section 3: The Sequence Features and Mechanisms of 264 Somatic REIs

265 To explore the mutational patterns of somatic TRI, we applied pairwise alignment
266 algorithms to compare the consensus sequences of germline and somatic events within
267 the somatic TRI windows across various samples. Sequences identified exclusively in
268 somatic TRI events, relative to their paired germline events, were classified as somatic
269 TRI mutational sequences. These sequences were classified into three categories based
270 on the repeat motifs involved. The first category, 'Expansion', referred to somatic TRI
271 events where a pre-existing germline repeat motif is expanded in the somatic sample
272 (Figure 2a). For example, in the somatic TRI window (chr18:80,195,548-80,195,600) of

273 sample UCEC12, the AT motif was notably expanded (Figure S5a), reflecting an increase
274 in repeat length. The second category, 'Motif-Change', involved somatic TRI events where
275 a new repeat motif emerges, absent in the germline sequence (Figure 2a). In sample
276 Lung136, the somatic sequence showed the appearance of the AGATATATAT motif, which
277 was not present in the germline TR window (chrX:145,918,007-145,918,062) (Figure S5a).
278 The third category includes somatic TRI events where the inserted sequence did not
279 match any recognizable repeat motif, classified as 'Undefined' (Figure 2a and S5a).
280
281 As expected, 95% of somatic TRI events were primarily characterized by the expansion of
282 repeat motifs (Figure 2b, 2c and S5b). Among these, 67.8% involved the expansion of
283 pre-existing germline motifs, with the proportion of mutational sequences ranging from
284 40.28% to 100% (average 87.89%, median 91.48%). 16.7% were driven by the expansion
285 of newly emerged motifs, with the proportion of mutational sequences ranging from
286 50.15% to 100% (average 81.90%, median 82.12%). Furthermore, 10.5% of somatic TRI
287 events exhibited a combination of both germline motif expansion (average 47.85%,
288 median 49.03%) and the expansion of new motifs (average 45.96%, median 44.98%). In
289 addition to these, 2.8% of somatic TRI events were classified as 'Undefined', where the
290 repeat motifs in the somatic INSs could not be identified. These events had an
291 undetermined insertion frequency ranging from 63.25% to 99.84% (average 86.64%,
292 median 86.84%). Furthermore, 2.2% of somatic TRI events displayed a combination of all
293 three mutational types, preventing classification into a predominant pattern. These events
294 were categorized as "Ambiguous". Given the challenges associated with nanopore
295 sequencing, particularly with long tandem repeat sequences [72], we propose that the
296 "Undefined" and "Ambiguous" categories may be affected by sequencing errors.
297
298 In terms of the sequence pattern of TEIs, we identified two types of somatic TEIs:
299 'complex TEIs', whose inserted sequences comprise multiple TE fragments (Figure 2d),
300 and 'solo TEIs', which consist of a single TE fragment. Among the somatic TEIs, 56%
301 were classified as solo TEIs, while 44% were complex TEIs (Figure 2e). Complex TEIs
302 were significantly longer than solo TEIs (average length: 1,823.8 bp vs. 570.5 bp,

303 Student's t-test, p-value < 2.2e-16) (Figure S5c). Approximately 80% of complex TEIs
304 exceeded 300 bp, beyond the typical read length of short-read sequencing (SRS), making
305 them difficult to resolve with SRS. Notably, 3.4% of complex TEIs matched the known
306 somatic recombination pattern of Alu and L1 [33,59]. The remaining complex TEIs exhibited
307 more intricate compositions of TE families and inverted TE segments (Table S5) —
308 observations not systematically identified in SRS studies.
309
310 Somatic TEIs containing LINE elements were most prevalent (57.8%), followed by those
311 containing SINE elements (44.7%) (Figure 2c, S5d and S5e). The distribution of somatic
312 TEIs varied across different tumor types. In SCC, gastrointestinal cancers (including IC
313 and GC), and UCEC, somatic TEIs predominantly contained LINE segments
314 (45.4%-76.0%), consistent with previous SRS studies showing a higher frequency of
315 somatic solo L1 insertions in SCC [73,74]. These tumors exhibited a typical ~300 bp
316 somatic TEI profile with no distinct peaks (Figure S5f). In contrast, other tumors a higher
317 proportion of somatic SINE-containing somatic TEIs (54.6%-70.4%). Peaks were
318 identified around 300 bp and 6 kb (Figure S5f), corresponding to SINE- and
319 LINE-containing somatic TEIs, respectively. High-burden samples exhibited a noticeable
320 peak at approximately 300 bp, with 48.2% (5422/11238) being solo SINEs. These findings
321 suggested differential TE activity across cancer types. Notably, younger TE subfamilies,
322 such as AluY and L1HS, significantly contributed to the tumor genome (Figure S5g),
323 highlighting their activity in tumor development and their key role in shaping the somatic
324 TEI composition of the cancer genome.
325
326 To explore the insertion mechanisms, we analyzed the sequences features of somatic
327 TEIs (Supplementary Note). By aligning somatic TEI sequences with the flanking regions
328 of insertion sites (STAR Methods), we identified 82.1% of somatic TEIs contained
329 homologous sequences. Based on the positions and lengths of homologous sequences,
330 we classified somatic TEIs into two mechanisms: retrotransposition and double-strand
331 break (DSB) repair via homologous recombination[59]. 58.2% of somatic TEIs were
332 associated with target site duplications (TSDs), a hallmark of retrotransposition (Figure 2g

333 and Figure S6a). The median TSD length was 14.6 bp (Figure S6h), consistent with
334 previous findings [74]. 31.3% of somatic TEIs exhibited homologous sequences (average:
335 760.8 bp) on the one side of the breakpoint, suggesting the homologous recombination.
336 Among these, 12.3% had homologous sequences longer than 100 bp (average: 262 bp,
337 ranging from 101 bp to 19,659 bp), similar to the non-allelic homologous recombination
338 (NAHR) mechanism described previously ([59,75] (Figure 2g and Figure S6b). 12.9% of
339 somatic TEIs carried shorter homologous sequences (<100 bp, average: 7 bp), indicating
340 microhomology-mediated recombination (Figure 2g and Figure S6c). The remaining
341 somatic TEIs were classified as 'other' indicating unknown mechanisms (Figure 2g and
342 Figure S6d).
343
344 We further uncovered the mechanisms underlying the formation of solo and complex TEIs.
345 Somatic solo TEIs were predominantly generated via retrotransposition (69.2%), with only
346 17.4% arising from recombination (Figure 2e). In contrast, complex TEIs exhibited a more
347 balanced distribution, with 38.9% arising from recombination and 40.3% by
348 retrotransposition (Figure 2e). Notably, the mechanisms of somatic complex TEIs varied
349 across cancer types. GC, GBM, and HCC had a larger proportion of somatic complex
350 TEIs generated by NAHR recombination (Chi-square test, p-value < 2.2e-16, Fold
351 Change = 2.56). Digestive system cancers (GC, IC, PAAD, and HCC) displayed a larger
352 proportion of somatic complex TEIs with unidentified mechanisms (Chi-square test,
353 p-value < 2.2e-16, Fold Change = 1.96). These findings suggest that the formation of
354 somatic complex TEIs is characterized by significant stochastic variability rather than a
355 strictly ordered process.
356

## 357 Section 4: High-Frequency Somatic TRI Events among 358 samples are potential Pan-Cancer Biomarkers

359 To identify non-random somatic TRI events, we compared the observed frequency
360 distribution of somatic TRI events with the expected distribution under random conditions.
361 This analysis revealed that 152 somatic TRI events, each detected in ≥10 patients, are

362 exceptionally rare under the random distribution model (permutation test, p-value ≤ 0.002;
363 STAR Methods, Figure S7a). Notably, three of these events have also been reported as
364 recurrent repeat expansion events (rREs) with subtype-specific characteristics in
365 previous pan-cancer studies of somatic simple repeat expansions[34] (Table S4). We
366 therefore defined these somatic TRI events as high-frequency somatic TRI events. Similar
367 to the genome-wide distribution pattern of somatic TRI events, we identified that
368 high-frequency TRI events are significantly underrepresented in coding sequence (CDS)
369 regions compared to the genome-wide background (binomial test, FDR = 0.0425; Figure
370 S7b). This finding suggested that the high-frequency somatic TRI events are more likely to
371 exert their downstream effects through alterations in non-coding regions, potentially
372 impacting regulatory processes. Previous studies have demonstrated that tandem repeat
373 (TR) sequence variations contribute to genomic regulation primarily by modulating
374 transcription factor (TF) binding sites [76] or by introducing CpG dinucleotide sequences
375 that influence methylation patterns, thereby affecting downstream gene transcription [77].
376
377 High-frequency somatic TRI events are significantly enriched in 10 out of 490 ENCODE
378 TF binding regions (Table S9) compared to low-frequency TRI events (Chi-square test,
379 p-value ≤ 0.05, absolute ratio difference > 5%; Figure S7c). Among these, the most
380 notable enrichment was observed in CTCF binding sites, where high-frequency somatic
381 TRIs displayed an odds ratio of 2.032 (p-value = 3.3e-4). This suggested that
382 high-frequency somatic TRIs may be involved in 3D chromatin structure alterations and
383 distal regulatory processes [78]. Additionally, when comparing somatic sequences of
384 high-frequency TRIs to their germline counterparts, we identified significant differences in
385 CpG dinucleotide counts in 62 out of 152 high-frequency somatic TRIs (Wilcoxon rank test,
386 FDR < 0.05, |log2(Fold Change)|>1, Figure S7d, Table S10). Leveraging the unique ability
387 of nanopore sequencing to simultaneously detect sequence variations and methylation
388 states [79], we found that 18 out of these 62 high-frequency somatic TRIs introduced
389 methylation marks at somatic extended CpG dinucleotides (Wilcoxon rank test, p-value ≤
390 0.05, Figure S7e, Table S11). For instance, in the sample Lung120, we investigated that a
391 somatic TRI region (chr6:64,295,338–64,295,575) within an intron of the *EYS* gene

392 introduced 55 methylated CpG sites (Figures S7f-S7i, Supplementary Note). As
393 mentioned in the previous study [77], these somatic TRIs may act as potential regulatory
394 elements.
395
396 To investigate the association between high-frequency somatic TRIs and cancer-related
397 biological processes, we performed functional enrichment analysis on genes neighboring
398 high-frequency somatic TRIs. Compared to low-frequency somatic TRIs, genes
399 associated with high-frequency somatic TRIs were significantly enriched in functional
400 categories related to focal adhesion, apical junctions, and cytokine-cytokine receptor
401 interactions (Figure 3a, Chi-square test, p-value < 2e-4). Among these, the somatic TRI
402 event located at chrX:9,791,969–9,792,160, within the intron of *SHROOM2,* which is
403 associated with apical junction functions, exhibited the highest shared frequency across
404 patients (57 out of 299, Figure 3b, Table S12). Using targeted PCR and Sanger
405 sequencing, we validated this somatic TRI event in 25 paired tumor and blood samples
406 (Figure 3c, Figure S8a, Table S13 and Table S14). Additionally, this *SHROOM2*-related
407 TRI was detected in liver cancer and intestinal cancer cell lines (Figure S8b). Overall, this
408 somatic TRI event was present in nearly 20% of pan-cancer samples (57 out of 299) and
409 occurred across 11 of the 12 cancer types analyzed. Notably, this somatic TRI exhibited
410 particularly high recurrence rates in HCC (58.82%; 10 out of 17) and IC (40%; 6 out of 15).
411 Intriguingly, all multi-site biopsies from three HCC patients with this somatic TRI
412 consistently contained it, while biopsies from two HCC patients without this somatic TRI
413 consistently lacked it. This result indicated a high degree of intra-tumor consistency for
414 this somatic TRI. Based on length distribution, somatic TRI sequences within this window
415 were categorized into two groups: long-fragment expansions and short-fragment
416 expansions (Figure 3d). The long-fragment expansions ranged from 3,896 to 8,284 bp,
417 whereas the short-fragment expansions ranged from 639 to 3,259 bp (Figure 3e).
418 Interestingly, the two cancer types with high recurrence rates (HCC and IC) predominantly
419 carried somatic TRIs with long-fragment expansions (HCC: 7 out of 10; IC: 5 out of 6;
420 Figure 3f, Fisher's exact test, p-value = 3.29e-6). These long-fragment expansions were

421 primarily characterized by the amplification of the AGAAT repeat motif (Figure S9, Table
422 S15).
423
424 Subsequently, using our NSCLC NGS RNA-seq dataset, we performed differential gene
425 expression analysis within a 300 kb window surrounding all high-frequency somatic TRI
426 events detected in NSCLC. This analysis identified 51 differentially expressed genes
427 associated with 33 high-frequency somatic TRI events (Figure S10a, Table S16). Among
428 these, we identified the somatic TRI event located at chr7:121,603,262–121,603,506,
429 which was previously reported in our study to be present in nearly 40% of LUSC samples
430 (Figure S10b and S9c). This somatic TRI was shown to regulate the expression of its
431 downstream *PTPRZ1* gene by altering the 3D genomic structure [25]. Additionally, we
432 identified two other somatic TRI events associated with the expression of genes in the
433 PTP family: the somatic TRI at chr1:202,204,213–202,204,518, which was correlated with
434 the upregulation of *PTPN7* expression (Figure S10b and S9d), and the somatic TRI at
435 chr7:158,595,097–158,595,342, which was associated with the downregulation of
436 *PTPRN2* expression (Figure S10b and S9e). Notably, while the somatic TRI at
437 chr7:121,603,262–121,603,506 was significantly enriched in LUSC and LC samples, the
438 other two PTP-related somatic TRI events did not exhibit enrichment in specific cancer
439 types (Figure S10f). Furthermore, we noticed the presence of these three TRI events
440 across the pan-cancer cohort, encompassing 66 samples.
441
442 By performing clustering analysis based on the distribution spectrum of somatic TRI
443 events, we identified that samples from HCC, GBM, GC, and LUAD formed four distinct
444 clusters (Figure 4a). Notably, GC, GBM, and LUAD samples did not harbor somatic TRI
445 events with an intra-cancer frequency exceeding 30% (Figure 4b), indicating a high
446 degree of inter-patient heterogeneity for TRI events within these cancer types. In total, we
447 identified 12 somatic TRI events that were detected within single cancer types with high
448 frequency (Figure 4c, STAR Methods). HCC harbored the highest number of intra-cancer
449 high-frequency somatic TRI events (4 out of 12; Table S16, Figures 4d). These included a
450 previously described somatic TRI event located within the SHROOM2 intron, as well as

three additional events: one within the PALMD intron, predominantly characterized by expansions of the AAAG motif (chr1:99,683,035–99,683,173, Figure S11a), which is consistent with previous reports in other pan-cancer somatic tandem repeat study [34] （Table S4）; another within the *LIPG* intron, mainly involving expansions of the GGGAGA motif (chr18:49,579,001–49,579,348, Figure S11b); and an intergenic event featuring expansions of the GGTTT motif (chr1:34,802,227–34,802,447, Figure S11c). Notably, these HCC-specific high-frequency somatic TRI events were also recurrently detected in IC, UCEC, OV, OSCC, and LC. Genes such as *SHROOM2* and *PALMD* are implicated in the regulation of the cytoskeleton [80-83], while *LIPG* is involved in lipid metabolism [84]. High-frequency somatic TRIs shared across different cancer types may suggest common malignant biological behaviors, making them potential diagnostic and therapeutic biomarkers for these cancers.

Additionally, digestive system adenocarcinomas (IC and PAAD), gynecological tumors (CCA, OV, and UCEC), and three SCCs (OSCC, LC, and LUSC) formed three distinct clusters based on the frequency of somatic TRI events (Figure 4a). Each cluster exhibited specific somatic TRI events. Within the IC-PAAD group, we identified three somatic TRI events with significantly higher frequencies compared to other cancer types (Student's t-test, p-value < 0.05): a somatic TRI event located in the *LRBA* intron (chr4:150,658,505–150,658,775; Figure S11d), another in the *SLC22A2* intron (chr6:160,218,690–160,218,725; Figure S11e), and the third upstream of the *SPINK1* gene (chr5:147,837,649–147,837,716, 6073 bp upstream; Figure S12a). These genes are associated with tumor cell growth (*LRBA*) [85], chemotherapeutic resistance (*SLC22A2*) [86], and tumor dedifferentiation (*SPINK1*) [87], respectively. In contrast to the IC-PAAD group, gynecological tumors (OV, CCA, and UCEC) formed a distinct cluster in the somatic TRI frequency spectrum, but exhibited low recurrence rates for their specific somatic TRI events. Two cluster-specific somatic TRI events (Student's t-test, p-value < 0.05) included one upstream of the ENSG00000257323 gene (chr12:74,615,532–74,615,609; Figure S12b) and another within the *MED16* intron (chr19:878,845–879,150; Figure S12c). These events appeared at frequencies below 30% within any individual gynecological

481 tumor type. The SCC cluster, consisting of OSCC, LC, and LUSC, exhibited three
482 cluster-specific high-frequency somatic TRI events (Figure 4f). These included the
483 previously described *PTPRZ1*-related somatic TRI event (Figure S10c), as well as two
484 somatic TRI events located in the promoter of the OBI1 gene
485 (chr13:78,659,295–78,659,339; Figure 4g) and the *SLC10A6* intron
486 (chr4:86,842,786–86,842,877; Figure S12d). In the available samples, we validated the
487 cluster-specific high-frequency TRI events through PCR experiments and Sanger
488 sequencing, highlighting the existence of similar somatic TRI distribution patterns among
489 cancer types within the same cluster (Figure S12, Table S14).
490

## 491 Section 5: somatic TEIs were enriched in the MHC II gene 492 cluster

493 To investigate the somatic TEI distribution across cancer types, we clustered samples
494 based on somatic TEI breakpoints, identifying five distinct clusters (Table S1) with unique
495 histological compositions (Figure 5a). Cluster 1 predominantly comprised SCC samples
496 (67%) (Figure 5b and 5c), including 87.5% of LUSC and 68.8% of OSCC cases (Figure 5c
497 and S14a). Notably, approximately 20% of ADC samples were also assigned to Cluster1,
498 with IC (73.3%) and OV (50%) being the most represented (Figure 5c and S14a). This
499 suggested that IC and OV share TEI characteristics with SCC samples in Cluster 1.
500 Moreover, ADC samples in Cluster 1 exhibited lower somatic TEIs burden (mean: 36.7,
501 median: 19) compared to those in other clusters (mean: 156, median: 23) (Welch
502 two-sample test，$p = 0.005$), indicating the distinct somatic TEI profiles of ADC samples
503 within Cluster 1. 57.7% of Cluster 1 SCC samples further subdivided into two subclusters.
504 In Subcluster 1, 75% of the samples were SCC, while in Subcluster 2, nearly 60% of the
505 samples were SCC (Figure 5d and S14b). Subcluster 1 exhibited a significantly higher
506 somatic TEI burden than Subcluster 2 (Figure S14c), highlighting the substantial
507 molecular heterogeneity within SCC samples.
508

509 Cluster 2 encompassed tumor samples with the highest levels of TEI burden, including all

510 high-burden samples and two multi-site biopsies from a LUAD patient, which ranked just
511 below the high-burden samples. Although multi-site biopsies were collected from 9
512 patients, with samples from the same patient clustering together, only the complete set
513 from patient HCC10 was assigned to Cluster 3, suggesting a potential association
514 between he multi-lesional nature of the patient and their distinct somatic TEI profile.
515
516 Clusters 4 and 5 were predominantly composed of ADC samples (68% and 66%,
517 respectively) but differed in somatic TEI burden and distribution. Cluster 4 exhibited a
518 higher somatic TEI burden and a greater degree of overlap between samples (Figure 5e),
519 indicating a more homogeneous TEI profile. This cluster also included poorly differentiated,
520 non-metastatic CCA samples and early-stage (I/II) HCC biopsies, which clustered with
521 multi-site samples from a stage I LUAD patient. In contrast, cancers commonly associated
522 with high heterogeneity, such as PAAD, UCEC, and GBM, were largely detected in Cluster
523 5, with PAAD samples showing duodenal or peripancreatic invasion (Figure S14a, Table
524 S1).
525
526 We analyzed high- and low-frequency somatic TEI regions across the clusters, covering
527 1,464 1Mb genomic windows. In Cluster 1 and Cluster 5, we identified 15 and 6 regions,
528 respectively, with significantly low-frequency somatic TEIs (Fisher's exact test, FDR <
529 0.01, Fold change < 0.5, and >10% of non-cluster samples showing somatic TEIs in the
530 same window). In contrast, Cluster 2, Cluster 3, and Cluster 4 contained 1384, 45, and 14
531 regions, respectively, with significantly high-frequency somatic TEIs (Fisher's exact test,
532 FDR < 0.01, Fold change > 2, and >10% of cluster samples exhibiting somatic TEIs in the
533 same window) (Figure S14d, Table S17).
534
535 To identify non-randomly enriched TEI windows, we applied a permutation test (STAR
536 Methods). In windows with ≥14 patients, somatic TEI frequency was significantly higher
537 than in randomly selected background regions. This suggested that these windows are
538 non-randomly enriched with somatic TEIs (permutation test, p-value=0.0002, Figure 6a).
539 Given the high sequence variability and repetitive nature of centromeric and

540 pericentromeric regions, as well as the limited functional data and reference genome
541 quality in these areas, we excluded them from further analysis. After exclusion, the
542 chr6:32,000,000-33,000,000 window, containing the MHC II gene cluster, exhibited the
543 highest somatic TEI count and the greatest proportion of TEI-positive samples (40.6%)
544 (Figure 6b). Particularly in Cluster 1 (67% SCC samples), the proportion of samples
545 harboring TEI (27.8%) is significantly lower than that in clusters predominantly composed
546 of non-SCC samples (49.7%) (Fisher's exact test, FDR=0.0007).
547
548 Somatic TEIs in this region were primarily located in the intergenic and coding regions of
549 the *HLA-DRB1*, *HLA-DRB5*, and *HLA-DRB6* genes (Figure 6c). The distribution of
550 somatic TEIs varied across cancer types. High-burden samples showed widespread
551 somatic TEIs around the MHC class II gene cluster, while LUAD samples showed higher
552 frequencies of somatic TEIs than LUSC (Figure 6c). Notably, in the *HLA-DRB1* genebody
553 and its 5 kb flanking regions, somatic TEIs were absent in LUSC, but frequent in LUAD
554 (Fisher's exact test, FDR = 0.00066). Most somatic TEIs in this region were solo SINEs
555 (Figure S15a), consistent with previous findings of LINE enrichment in SCC samples.
556
557 Next, we examined the impact of somatic TEIs in the chr6:32,000,000-33,000,000 region
558 on gene expression using transcriptomic data from LUAD tumor samples. Comparison of
559 32 LUAD samples with somatic TEIs in this region to 39 samples without revealed 1,798
560 differentially expressed genes (p value < 0.05, |log2(FC)| > 1, Table S18, Figure S15b).
561 These genes were enriched in cytokine signaling and immune response pathways (Figure
562 S16a and S16b). 6 immunoglobulin-related and 4 T-cell receptor-related genes were
563 significantly downregulated in LUAD samples with somatic TEIs (Figure S15c). These
564 results suggested that TEIs in this region may contribute to immune evasion and tumor
565 progression.
566
567 Next, we further validated our bioinformatics findings through experiments. First, in a
568 subset of samples that underwent LRS, the authenticity of the TEI event was gauged by
569 Polymerase Chain Reaction (PCR). With blood DNA serving as a reference, when a TEI

570 event was present in the tumor tissue sample, the target band obtained from the tumor
571 tissue sample exceeded that of the blood tissue under identical primer amplification
572 conditions (Figure S15 d and e, Table S19). Moreover, through the utilization of specific
573 fluorescence in situ hybridization (FISH) probes (Table S20), TEI events were also
574 discerned in certain sample tissue sections (Figure 6e). It is widely recognized that
575 tumor tissues encompass a rich and diverse spectrum of cell types, inclusive of infiltrating
576 immune cells. Considering that the TEI insertion site is located in the MHC II gene cluster
577 region, we gathered additional fresh clinical samples (Table S21). By sorting CD45
578 positive cells, we further amplified the TEI events identified in this region within the clinical
579 sample LRS sequencing outcomes. The findings confirmed that these insertions occurred
580 in both CD45 positive and negative cells (considered as tumor cells) (Figures S17).
581 Subsequently, we detected TEI events within the MHC II gene cluster in 33 tumor cell
582 lines stored in the laboratory. In alignment with the sequencing results in clinical samples,
583 non-squamous carcinomas harbored a greater number of TEI events in contrast to
584 squamous carcinomas (a total of 63 TEIs were detected, with an average detection of
585 11.07 in squamous carcinomas and 19.35 in non-squamous carcinomas) (Figures S18,
586 S19, and S20a).
587
588 MHC class II molecules play a crucial role in cellular immunity. In antigen-presenting cells
589 (such as macrophages, dendritic cells, and B cells), MHC class II molecules are involved
590 in the processing and presentation of ingested exogenous antigens into peptide
591 fragments, and presenting them to CD4+ T cells, thereby initiating the adaptive immune
592 response[88]. Here, we compared the in vitro phenotypes of CD40L activation by TEI-enrich
593 and TEI-impoverished tumor cells through in vitro co-culture of tumor cells and mouse
594 CD4-positive t cells. The results showed that in squamous cell carcinoma cell lines, the
595 surface expression of CD40L on TEI-rich tumor cells was twice that of TEI-impoverished
596 cells (Figure 6e and Figure S20b); while the activation situation is reversed in
597 adenocarcinoma cells (Figure 6e and Figure S20b), this all suggested that the frequency
598 of TEI events and the sequence type of TEI event insertions may be important factors in
599 regulating the success of tumor adaptive immune initiation.

# Discussion

This study was designed to elucidate high-frequency somatic large-scale REIs (>50bp) in the pan-cancer genome. Previous research has been constrained by the limitations of SRS technology, which hindered the precise identification of these somatic SVs and their distribution across tumor population. LRS technology offers a promising solution to these challenges [89]. In this work, we applied nanopore LRS to sequence 325 tumor samples and their matched blood samples from 12 cancer types, generating nearly 46 Tb of data. This extensive LRS dataset, one of the largest to date, provides a robust resource for exploring somatic large-scale REI events across the pan-cancer landscape.

Current LRS-based somatic SV detection algorithms are hindered by alignment biases, particularly in repetitive genomic regions. These biases, stemming from the seed-and-chain alignment strategy, undermine the accuracy of insertion detection [89,90]. Although assembly-based approaches hold promise, they depend on high-quality individual genome assemblies, which are challenging to produce in the context of heterogeneous tumor genomes [91,92]. To overcome these limitations in detecting somatic REIs, we developed specialized bioinformatics methods tailored for somatic TRI and TEI detection, respectively. By harnessing the technical strengths of long reads that span somatic insertions, and combining unsupervised learning with partial order alignment graphs and statistical analysis of local sequence features, we minimized errors introduced by reference genomes and alignment algorithms, thus enabling more accurate detection of somatic REIs. Recent advancements in long-read sequencing (LRS) technology, alongside continuous improvements to the human genome reference—particularly the completion of the Telomere-to-Telomere (T2T) project [93]—have significantly enhanced the study of large-scale repetitive sequence variations, such as centromeric and segmental duplications [94]. This study also benefits from these developments, uncovering numerous related somatic REIs. However, the structural complexity and dynamic nature of these regions, marked by variation in repeat copy

629 number and sequence divergence, present substantial challenges to accurate detection.
630 Consequently, refined bioinformatics methods and ultra-long sequencing technologies are
631 expected to provide critical solutions to these issues [95].
632
633 Using our developed methodology, we identified a significant number of high-frequency
634 somatic REIs across pan-cancer genomes. In addition to the somatic TRI events located
635 within the introns of the *SHROOM2* and *PALMD* genes, which exhibited remarkably high
636 recurrence in HCC, IC, CCA, UCEC, OV, LC, OSCC, and LUSC. Interestingly, cancers
637 from anatomically neighboring tissues tend to exhibit similar somatic TRI profiles. For
638 example, the gynecological cancers CCA, UCEC, and OV shared a similar somatic TRI
639 spectrum, with a somatic TRI event in the intron of *MED16* being specific to these tumor
640 types. The *MED16* gene is associated with estrogen receptor sensitivity in the previous
641 research [96], the gynecological cancers specific somatic TRI event may play a role as a
642 regulatory element of this gene. Similarly, squamous cell carcinomas, including LUSC, LC,
643 and OSCC, displayed similar somatic TRI profiles. Notably, a somatic TRI event identified
644 in approximately 40% of LUSC samples was found to occur at an even higher frequency
645 in LC samples (nearly 70%). This event has been shown to influence the
646 three-dimensional structure of the genome, thereby modulating the expression of the
647 downstream gene *PTPRZ1* through enhancer activity [25]. In addition to *PTPRZ1*, further
648 analysis of paired RNA-seq data from non-small cell lung cancer (NSCLC) revealed two
649 additional somatic TRI events (chr1:202,204,213-202,204,518 and
650 chr7:158,595,097-158,595,342, Figure S10a and S10b), which were significantly
651 associated with the expression of their downstream genes, *PTPN7* and *PTPRN2*. Notably,
652 20.31% of pan-cancer patients carried at least one somatic TRI event potentially
653 associated with the PTP gene family. As an important receptor kinase family, the PTP
654 genes play a crucial role in immune evasion mechanisms [71]. We propose that somatic TRI
655 events may function as key distal regulatory elements for the expression of this gene
656 family. Further investigation into the transcriptional regulatory mechanisms underlying
657 these somatic TRI events will be essential to understanding the role of the PTP gene

658 family in immune regulation. Understanding the impact of the high-frequency somatic TRI
659 events we identified on the tumor genome network is a key focus for future research.
660
661 The success of immune checkpoint inhibitors in certain cancers has demonstrated that
662 the immune system can target and reject tumors through T cell responses. Tumor cell
663 recognition relies on MHC-mediated presentation of tumor-associated antigens. Our
664 research reveals that somatic TEIs are enriched in MHC II gene regions, with over 40% of
665 pan-cancer samples harboring somatic TEI events in these regions. Notably, LUAD
666 exhibited a significantly higher somatic TEI burden in MHC II regions compared to LUSC
667 (Figure 6c). Although the immunological effects of these somatic transposable element
668 insertions (TEIs) have not been fully elucidated, co-culture experiments indicated that
669 TEI-enriched adenocarcinoma cell lines are more potent in suppressing CD4+ T cell
670 activation compared to TEI-impoverished counterparts. Furthermore, LUSC patients have
671 been shown to derive greater clinical benefit from PD-1 inhibitors than LUAD patients [97].
672 Given that TEI-enriched LUAD cell lines suppress CD40L expression on CD4+ T cells, we
673 hypothesize that immune activation within these tumors may be constrained. Since PD-1
674 inhibitors function to counteract PD-L1-mediated immune suppression of CD8+ T cells,
675 their efficacy in TEI-enriched LUAD may be diminished due to the absence of CD4+ T cell
676 activation. Nevertheless, this hypothesis warrants further in vivo validation. Moreover, we
677 propose that TEI enrichment in MHC II regions could serve as a potential target for tumor
678 immune evasion. Yet, the precise mechanisms by which TEIs modulate immune
679 responses remain poorly understood, highlighting a critical area for future investigation in
680 tumor immunology.
681
682 Here, we leverage LRS technology to investigate somatic REIs and their potential as
683 diagnostic and therapeutic biomarkers in cancer. By integrating optimized bioinformatics
684 algorithms and advancing long - read sequencing technologies, future studies will be able
685 to systematically characterize diverse somatic genomic alterations, such as translocations
686 and inversions, from the dataset we provided. Additionally, integrating multimodal
687 nanopore sequencing data will enable single-cell resolution analysis of large genomic

regions and their associated epigenetic modifications, as demonstrated in recent studies [98]. These efforts will provide critical insights into the regulatory mechanisms underlying large-scale somatic structural variations (SVs) in cancer genomes, shedding light on their functional impact. While our current dataset offers valuable insights into pan-cancer diagnostics, expanding the cohort to include diverse cancer types and enhancing clinical follow-up data—such as treatment responses and patient survival—will enable a comprehensive exploration of the relationship between somatic SVs and cancer outcomes. To achieve this, we are actively expanding our research cohort to encompass a broader spectrum of tumor types and patient demographics, ensuring robust characterization in future studies.

In conclusion, our study demonstrates the transformative potential of long-read sequencing in cancer genomics. By expanding the understanding of the genetic variants of cancer, this work sets the stage for future studies that will explore the role of SVs in human disease. As LRS technology and analysis methods evolve, we anticipate these advancements will enable more accurate SVs detection and a deeper insight into their functional consequences, ultimately guiding the development of targeted therapies and personalized treatment approaches in oncology.

# STAR Methods

## Patient Cohort and Ethics

The matched tumor-blood samples across 8 cancer types (HCC, GBM, PAAD, LUAD, IC, GC, LUSC, and LC) were collected from patients who undergone surgery at our hospital. The matched tumor-blood samples across 3 gynecological cancer types (CCA, OV and UCEC) were collected from patients who undergone surgery at West China Second University Hospital, Sichuan University. And the matched tumor-blood samples of OSCC were collected from patients who undergone surgery at West China Hospital of Stomatology, Sichuan University. None of the patients were received preoperative chemotherapy or radiotherapy. All diagnoses confirmed through histological review by a board-certified pathologist. Cancer clinical stage was determined based on the 8th edition of the American Joint Committee on Cancer (AJCC) TNM staging system. This study protocol was approved by the Institutional Review Board of our hospital (Ethics: Project

720 identification code: 20242416). Written informed consent was obtained from all patients.
721 Demographic and clinical data were collected for analysis(Table S1).
722

723 **PromethION whole genome Sequencing**

724 High-molecular-weight (HMW) genomic DNA (gDNA) was extracted from tissue or blood
725 samples using the MagAttract HMW DNA Kit (QIAGEN). DNA quality was assessed by
726 Novogene Co. LTRI. using three standard methods: (1) DNA purity was determined by
727 measuring OD260/280 and OD260/230 ratios with a NanoDrop spectrophotometer, (2)
728 agarose gel electrophoresis was used to evaluate DNA degradation and detect RNA
729 contamination, and (3) DNA concentration was quantified with the Qubit® DNA Assay Kit
730 and Qubit® 3.0 Fluorometer (Life Technologies, CA, USA).
731

732 For library preparation, 273 samples (84%) were processed using the SQK-LSK109 kit,
733 while 52 samples (16%) were prepared with the SQK-LSK114 kit (Oxford Nanopore
734 Technologies, UK), following the manufacturer's protocols (Table S1). For each sample,
735 48 μL of gDNA (2 μg) and nuclease-free water (NFW) were combined with 3.5 μL of
736 NEBNext FFPE DNA Repair Buffer, 2 μL of NEBNext FFPE DNA Repair Mix, 3.5 μL of
737 NEBNext Ultra II End Prep Reaction Buffer, and 3 μL of NEBNext Ultra II End Prep
738 Enzyme Mix. The mixture was gently combined in a PCR tube, spun down, and incubated
739 sequentially at 20 °C for 5 minutes and 65 °C for 5 minutes. The prepared sample was
740 transferred to a 1.5-mL tube, and 60 μL of AMPure XP beads was added. After gentle
741 mixing and two washes with 70% ethanol, the purified DNA was combined with 25 μL of
742 Ligation Buffer (LNB), 10 μL of NEBNext Quick T4 DNA Ligase, and 5 μL of Adapter Mix
743 (AMX). This mixture was incubated at 25 °C for 20 minutes. Subsequently, 40 μL of
744 AMPure XP beads was added, and the pellet was resuspended in 25 μL of Elution Buffer
745 (EB).
746

747 For sequencing preparation, 30 μL of flush tether (FLT) was added to PromethION Flush
748 Buffer (PFB), and 500 μL of the Priming Mix was loaded onto a PromethION flow cell and
749 equilibrated for 5 minutes at room temperature. The library preparation involved mixing 75
750 μL of sequencing buffer (SQB), 51 μL of loading buffer (LB), and 24 μL of DNA library.
751 This was followed by loading 500 μL of the Priming Mix and 150 μL of the library onto the
752 flow cells. DNA sequencing was performed on flow cells as per the manufacturer's
753 instructions. Samples processed with SQK-LSK109 were sequenced on R9 flow cells
754 (84%), while those prepared with SQK-LSK114 were sequenced on R10 flow cells (16%).
755 Base calling was performed using Guppy 3.2.8 with default parameters in real time during
756 sequencing.
757

## Short-read RNA-seq Data Analysis

For some NSCLC samples, paired-end reads were aligned to the hg38 reference genome using HISAT2 [99] (version 2.1.0). SAMtools [100] was used to sort and index the bam files. StringTie [101](version 1.3.6) was used to generate the GTF files containing the read count information and calculate the TPM matrix of gene expression for all genes across all samples. The raw count matrix of gene expression for all genes across all samples is calculated by htseq-count [102] (version 2.0.2) based on the gencode annotation file version 42.

## Detection of differential expression gene

In order to defined potential gene associated to somatic REIs, we classified NSCLC samples into two groups based on whether harboring target somatic REI event. Then the gene expression normalized and compared between two groups through DESeq2 [103]. Genes with p-value <= 0.05 and |log2fold-change| > 1 were defined as differential expression genes.

## Oxford Nanopore Sequencing Data Processing

Initially, raw nanopore sequencing data were processed using NanoFilt (v2.8.0, https://github.com/wdecoster/nanofilt.git, parameters: -q 7 -l 1000) to filter the fastq files and generate clean reads. The quality of these filtered reads was then assessed using NanoPlot [104](v1.39.0, parameter: --maxlength 40000). Next, the ONT long reads were aligned to the hg38 reference genome using Minimap2 [44] (v2.2.17) with the following parameters: --MD, -ax map-ont, -L, and -t 30.

## Tumor purity calculation

Somatic variants (SNVs) and copy number alterations (CNVs) were identified in tumor samples using multiple approaches. For LRS WGS sequencing data, SNVs were called using ClairS [105] (version 0.1.2) with parameters set to --platform ont_r10 --min_coverage 3. CNVs were detected using Control-FREEC [106] (version 11.6) with default settings. For the NGS WGS dataset, SNVs were identified in both tumor and matched blood samples using Mutect2 from GATK [107] (version 4.1.2.0). The resulting variants were filtered using GATK's FilterMutectCalls tool. CNVs in the WGS dataset were also detected using Control-FREEC with default parameters. Tumor purity was then estimated using both SNVs and CNVs with ABSOLUTE [108] (version 1.0.6), applying the following parameters: min.ploidy = 0.5, max.ploidy = 8, max.sigma.h = 0.2, max.neg.genome=0.005, sigma.p=0, max.as.seg.count=1500, and max.non.clonal=100.

## Somatic TRI Identification

Considering the high false-positive rate associated with detecting somatic SVs of the tandem repeat type using current algorithms, we have developed a more accurate method for identifying somatic tandem repeat insertions across pan-cancer samples (Extended Figure 2a). Our preliminary screening for potential somatic tandem repeat insertion sites involved the integration of two algorithms: Sniffles2 and Straglr, to pinpoint candidate regions. We initially employed Sniffles2 (version 2.0.7) with parameters --minsvlen 50 --mapq 5 --minsupport 3 to analyze tumor LRS WGS data, detecting INS SV intervals. These intervals were then intersected with tandem repeat regions annotated in the hg38 genome by RepeatMasker, utilizing the bedtools intersect command, with RepeatMasker-annotated tandem repeats as the initial candidate windows. Subsequently, we performed a secondary screen with straglr (version 1.4.1) using parameters --min_str_len 2 --max_str_len 100 --min_ins_size 100 --genotype_in_size --min_support 2 --max_num_clusters 9 on the tumor LRS WGS data, uncovering additional potential tandem repeat regions not identified by the sniffles2 workflow. These regions were added to the pool of candidate somatic tandem repeat intervals using the bedtools 'intersect' command. All candidate windows from each sample were further refined using the TDScope-simple command to efficiently filter for windows likely to contain somatic tandem repeat variations. We then consolidated all candidate somatic tandem repeat windows across samples using bedtools merge -d 200, yielding a comprehensive set of candidate windows for the pan-cancer cohort. Utilizing TDScope (version 1.0) with default parameters, we assessed these windows in both cancer samples and their corresponding normal blood samples. Candidate somatic tandem repeat intervals were identified by filtering for intervals with ≥3 supporting reads in normal blood samples, ≥3 supporting reads in tumor samples, an SV length (SVLEN) of ≥50, and a somatic genotype number = 1.

## Classification of Somatic TRI by Mutation Type

To annotate the consensus sequences of tumor and blood in TRI windows, we employed RepeatMasker, pyTRF (https://github.com/lmdu/pytrf), and mTR[109]. Due to the superior performance of pyTRF in annotating simple repeat sequences, we prioritized its annotation results. The remaining unannotated sequences were subsequently annotated using RepeatMasker and mTR. To eliminate inconsistencies between the different software tools, we normalized motifs of simple repeats, accounting for strand orientation and repeat unit size. This allowed us to obtain a consistent annotation of tumor and blood consensus sequences.

Pairwise alignment of the tumor and blood consensus sequences was performed to identify regions in blood characterized by gaps corresponding to insertions in tumor.

Annotation of the tumor insertion sequence was compared with all annotations of the blood sequence in the corresponding TRI window. If the tumor insertion motif matched the blood motif, the mutation type was classified as "Expansion." If the motifs differed, it was categorized as "Motif-Change." If no motif was annotated for the tumor insertion, the mutation type was classified as "Other." Using this approach, we determined the length of each mutation type in the tumor insertion sequence.

## Detection of Differential Somatic TRI Methylation Levels

To identify somatic TRI windows with significantly altered CpG patterns, we compared the CpG counts in all high-frequency TRI window pairs between somatic and germline sequences. A paired Wilcoxon rank-sum test was employed to assess differences in CG counts between somatic and germline samples. P-values were adjusted for multiple comparisons using the Benjamini-Hochberg method to calculate the false discovery rate (FDR) for each window. The relative change in CpG counts between somatic and germline sequences for each window was expressed as the log2 fold change (log2(FoldChange)), calculated as: log2FolgChange = log2(mean(somaticCG)+1 / mean(germlineCG+1)). High-frequency TRI windows with FDR ≤ 0.05 and |log2(FoldChange)| ≥ 1 were defined as significant CpG alteration windows, and these were used to compare methylation level changes in somatic sequences relative to germline sequences.

Specifically, reads from tumor samples classified by the TRIscope algorithm as 'somatic support' or 'unsupport' reads were aligned to reference genomes where somatic and germline sequences had been replaced, using Minimap2 (version 2.2.17) with the following parameters: --MD -ax map-ont -L. The methylation levels of somatic and germline TRI events were then calculated using the f5c pipeline (version 1.3, parameters: call-methylation -t 20 --iop 40).

To identify CpG insertion sites in somatic TRIs relative to germline TRIs, we used a pairwise alignment comparison algorithm (Python package: pairwise2 from the Biopython library). Based on these comparisons, we calculated methylation level changes for the inserted CpG sequences in somatic TRI events.

## Refinement Of High-confidence Somatic TEI Candidates

Samtools (version 2.8.0) was employed to tag BAM format files with the following command: samtools view -H “$bam_sort”; samtools view “$bam_sort” | awk 'BEGIN{OFS=“\t”}{$1=“tag”$1; print $0}') | samtools view -S -b - -o “$bam_sort_tag”. Following tagging, Samtools was again used to merge the tagged BAM files from tumor

and paired blood samples (samtools merge -h “$blood_bam” “$merge_bam” “$blood_bam” “$tumor_bam”). Subsequently, SVs were identified using Sniffles2 (v2.0.6) based on the merged BAM files, which reported all supporting reads of each SV in the variant call format (VCF) file. Reads with a mapping quality below 10 were excluded from SV detection. To improve the accuracy of SV detection, a tandem repeat file (https://github.com/fritzsedlazeck/Sniffles/blob/master/annotations/human_GRCh38_no_alt_analysis_set.trf.bed) was provided using the --tandem-repeats parameter, enabling Sniffles to better resolve SVs involving repetitive sequences. SVs supported only by reads shorter than 1,000 base pairs (bp) were also discarded. Large-scale insertions were considered up to a maximum length of 10,000 bp. All other parameters were set to default values. Then, an initial set of candidate somatic SVs was filtered using a custom Python script with the following parameters: tumor reads supporting SV ≥ 3, blood reads supporting SV = 0, and SV length ≥ 50 bp.

To improve accuracy, an additional screening of candidate somatic SVs was performed. Specifically, SV positions were restricted to a standard deviation of less than 150 bp. For insertions (INS) and duplications (DUP), comparisons were made between the alignment-supported lengths and those reported by Sniffles, with outliers excluded based on:

A.mean(Lengtha)-2*STRI(Lengtha)<=Lengths <=mean(Lengtha)+2*STRI(Lengtha)
B.Q1-1.5IQR<=Length:<=Q3+1.5IQR.

Additionally, SVs located in repeat regions were assessed for significant length differences between tumor and blood alignments, retaining only those exhibiting clear distinctions.

To further ensure high-confidence somatic SV identification, an error-correction process was conducted. Candidate regions were first identified as those within a ±50 bp range centered around insertion sites reported by Sniffles2. Alignments in these regions were extracted, and the reference genomes were polished using Racon (version 1.4.16, default parameters) over six consecutive iterations, utilizing paired blood reads from each region. Tumor and blood reads were then realigned to these polished reference genomes, and SVs were detected using Sniffles2. SVs supported by more than three tumor reads, with a length exceeding 50 bp, were retained as high-confidence somatic SVs.

## Somatic TEI Annotation

Iris is a software developed to improve the accuracy of insertion calling. We used Iris software (version 1.0.4, parameters: genome_buffer=1000 --keep_long_variants) to correct the positions and sequences of somatic SVs identified as INSs and DUPs. To improve the accuracy of Iris, we pre-processed the data by supplementing missing sequences for variants that Sniffles did not report. Specifically, we extracted the highest-quality reads from those supporting the Sniffles call that also spanned the insertion site. The insertion sequence from these reads was then identified using pysam and used as the input for Iris correction.

After error correction, we used RepeatMasker (version 4.1.2) to annotate the corrected insertion sequences with the following parameters: -species human -engine RMBlast -q, along with 50 bp of reference genome sequences upstream and downstream of the insertion site. If RepeatMasker failed to identify transposon regions, we performed re-annotation using Tandem Repeat Finder (TRF, version 4.10.0, parameters: 2 5 7 80 10 50 2000 -l 10 -ngs -h) and Sdust (version 0.1). Insertion sequences marked as “Unmask” by RepeatMasker were also re-annotated with TRF and Sdust if the annotated length exceeded 5 bp.

## Classification of Somatic TEI

If the 50 bp flanking sequences and the insertion sequence are not annotated as the same type of repeat and the insertion carries a transposable element (TE) sequence, the somatic INS is classified as a somatic TEI. Insertions containing TE sequences are further classified based on their structure and recombination mechanism. A single, complete insertion in the same orientation is classified as Solo, while insertions with multiple or fragmented TEs, or those with different orientations, are classified as Complex.

## Classification of SINE and LINE Truncations in Somatic TEIs

Each SINE and LINE insertion was annotated using RepeatMasker to compare the insertion sequences against template sequences in the RepeatMasker database. Using the positive strand as an example, we classified the insertion sequences based on their alignment characteristics. If the somatic insertion aligned fully with the template, with no gaps at either end, it was classified as Full Length. If the sequence matched entirely to the 3' end (i.e., left = 0), it was categorized as a 5' Truncation. Conversely, if it matched fully to the 5' end (i.e., begin = 1), it was classified as a 3' Truncation. If the sequence partially aligned with both the 3' and 5' ends but was not fully aligned at either end, it was classified as a TE Insertion. Additionally, the alignment coordinates of each SINE and LINE insertion against the template sequence were recorded for subsequent statistical analysis.

## Detection of SINE and LINE PolyA tails in Somatic TEI

For an insertion sequence on the negative strand, if there were 10 or more consecutive 'A' nucleotides within the first 30 bp of the sequence, it was classified as containing a polyA tail. Conversely, for an insertion sequence on the positive strand, the presence of 10 or more consecutive 'T' nucleotides within the last 30 bp was used as the criterion for classifying the sequence as containing a polyA tail.

## Detection of Somatic TEI Homologous Sequences

To identify homologous sequences between the ends of an insertion and the flanking regions of the reference genome, we performed BLAST analysis (version 2.14.1, parameters: -outfmt 6 -word_size 4), This analysis compared each insertion sequence with reference genomic sequences of varying lengths on both sides of the insertion site.

For gradient length determination, the maximum reference length was set equal to the length of the insertion, and the minimum length was defined by the allowable breakpoint deviation. For sequences shorter than 100 bp, we used length increments of 5 bp, while for sequences longer than 100 bp, increments were set to one-twentieth of the insertion length to optimize computational efficiency. Homologous sequences were defined as the longest segments within the breakpoint deviation that closely aligned with the ends of the insertion sequence.

Based on the relative position and length of these homologous sequences, insertions were classified into four categories:

1. Non-allelic homologous recombination (NAHR): Homologous sequences greater than 100 bp in length.
2. Tandem duplications (TSD): Homologous sequences less than 100 bp in length with a distal relative position.
3. Recombination (Else): Homologous sequences that did not meet the criteria for NAHR or TSD.
4. No homologous sequence detected: Insertions were classified as "N" when no homologous sequence was identified.

## Somatic TEIs Motif Clustering

To examine the distribution patterns of somatic TEIs across the genome, we divided the genome into 1 Mb fixed-size windows to create a pan-cancer atlas of somatic TEI distribution. The genome was segmented into consecutive 1 Mb windows, and the presence of somatic TEIs within each window was recorded for each tumor sample. This

resulted in a binary matrix representing the presence or absence of somatic TEIs across the genome for each sample.

Windows containing TEIs in only a single tumor sample were excluded from further analysis. For the remaining data, we calculated the Jaccard similarity coefficient for each pair of tumor samples to assess their genomic TEI similarity. Hierarchical clustering was then performed using the pheatmap package in R (ward.D method) to group tumor samples based on their somatic TEI distribution.

## Analysis of the Association Between L1 Insertion Rate and Genomic Features

The L1 insertion rate is determined by calculating the total number of somatic L1 insertions (10,140) identified in 317 out of 322 samples within each 1Mb window. To investigate the association between L1 insertion rates and multiple predictor variables, we employed a statistical framework based on negative binomial regression1. Specifically, we categorized the genome into four bins (bin0 - bin3) based on the degree of match between the classic L1 endonuclease motif (defined as TTTT|R, where R is A or G or Y|AAAA, where Y is C or T) and the DNA sequence. Bin0 contains DNA motifs that have four or more mismatches (out of five possible), which are highly dissimilar. Bin1, bin2, and bin3 include genomic segments of the hg38 genome assembly with three, two, and at most one mismatch, respectively. The most tightly matching DNA strands at each site are considered.

Regional analysis of genomic features utilized histone mark data (H3K9me3, H3K4me3, H3K36me3, H3K27ac ChIP-Seq data) and DNase hypersensitivity (DHS) data from the Roadmap Epigenomics Consortium. These data were divided into four bins based on the average enrichment scores across eight cell types (E017, E114, E117, E118, E119, E122, E125, E127). Bin0 includes regions with enrichment scores below baseline (Roadmap enrichment scores relative to input <1), while bins 1 to 3 correspond to regions with higher-than-average enrichment scores across the genome. Specifically, the DHS bins 1-3 span 132.1-136.1 Mb, the H3K9me3 bins 1-3 span 192.8-198.7 Mb, the H3K36me3 bins 1-3 span 142.3-146.5 Mb, the H3K4me3 bins 1-3 span 44.3-45.9 Mb, and the H3K27ac bins 1-3 span 77.8-80.1 Mb.

Replication timing (RT) data from ENCODE were processed by averaging smoothed waveform signals from eight cell types (IMR-90, HeLa S3, HUVEC, BJ, HEP G2, NHEK, SK-N-SH, and MCF-7). The data were divided into four equal-sized bins (quartiles), where bin0 represents the latest replication and bin3 represents the earliest replication.

1022

1023 RNA-Seq data, also from the Roadmap, were processed similarly. By averaging data from
1024 eight cell types (E071, E096, E114, E117, E118, E119, E122, E127), the data were
1025 divided into four genomic bins: bin0 contains non-expressed genes (logRPKM=0) and
1026 non-expressed intergenic DNA (totaling 1173.9 Mb), while bin1 (maximum
1027 logRPKM=0.02), bin2 (maximum logRPKM=0.15), and bin3 (logRPKM > 0.15) span 353.7
1028 Mb, 302.2 Mb, and 382.9 Mb, respectively.

1029

1030 In all figures (Supplementary Fig. 3b), the enrichment scores compare differences in
1031 specific features (replication timing, histone marks, gene expression, and L1 motifs)
1032 between bins 1-3 and bin0. The logarithmic enrichment for bin0 is defined as 0, and thus it
1033 is not displayed in the enrichment plots. Regional analysis is limited to genomic regions
1034 with perfect alignment scores, based on the hg38 reference genome.

1035

## 1036 Fresh tissue preparation and dissociation into single cells.

1037 All the freshly resected surgical specimens were immediately washed with
1038 phosphate-buffered saline (PBS) and processed to generate single cell suspensions.
1039 Tissue digestion was performed in 15 ml tube containing 10 ml pre-warmed RPMI 1640
1040 (ThermoFisher Scientific), 2 mg/ml dispase (Roche), 1 mg/ml type IV collagenase (Sigma)
1041 and 10 U/μl DNase I (Roche) for 60 minutes at 37°C. The reaction was deactivated by
1042 10% FBS. Cell suspensions were filtered using a 70 μm filter and then centrifuged at 500
1043 rpm for 6 min at 4°C to pellet dead cells and red blood cells. The cells were washed twice
1044 and suspended in PBS with 0.5% bovine serum albumin (BSA, Sigma), and, used CD45
1045 MicroBeads (Miltenyibiotec ,130-045-801) to distinguish CD45+ cells from the rest of the
1046 cells.

1047

## 1048 Validation of the Somatic TEI&TRI in Tissues and Cell Lines using PCR
## 1049 and Sanger sequencing

1050 DNA extraction and PCR: DNA was extracted from tumor tissues, blood, LUSC cell lines
1051 (H226, H520, HCC95), LUAD cell lines (PC-9, A549, H1299 and H1975), IC cell lines
1052 (LOVO, SW620, RKO, HCT15, HCT116,T84), Breast cancer cell line (HCC1395), OSCC
1053 cell lines (TU212, TU686, SNU46, SNU899), CCA cell lines (H8, C33A, CASK, SiHA),
1054 PAAD cell lines (HPAC, PANC-1,BXPC-1,ASPC-1),HCC cell lines (sun387, Huh7, hcclm9,
1055 snu499,sk-hep1), the H1581 cell line, Normal cell line BEAS-2B, HCC1395BL using the
1056 DNeasy Blood & Tissue Kit (QIAGEN) according to the manufacturer's instructions. The
1057 forward and reverse primers were designed according to the 100 bp upstream and
1058 downstream sequences of the insertion position and synthesized by Sangon Biotech
1059 (Table TEI,Table TRI). DNA was extracted from clinical samples or cell lines and

1060 subjected to PCR using primers. Each 50-μL reaction consisted of 2× PremSTAR Max
1061 Premix 25 μL, 0.8 μL 10 μM anchor primer, 0.8 μL 10 μM test primer, and 50-150 ng DNA.
1062 PCR was performed as follows: an initial step of 2 minutes at 98 °C, 30 cycles of 10
1063 seconds at 98 °C, 15 seconds at 60 °C and 60-300 seconds at 72 °C (size of PCR
1064 products between 200 bp and 6 kb). Finally, the samples were incubated at 72 °C for 5
1065 minutes. The PCR products were sent to Sangon for Sanger sequencing.
1066

1067 **Cell Culture**

1068 Human cell lines provided by the American Type Culture Collection (Manassas, VA, USA)
1069 were used in the present study. The cells were incubated in Dulbecco's modified Eagle's
1070 medium (DMEM) or RPMI 1640 medium (Gibco; Thermo Fisher Scientific, Inc.). The
1071 medium was supplemented with 10% fetal bovine serum (FBS) (Gibco; Thermo Fisher
1072 Scientific, Inc.), and the cells were all maintained at 37 °C in 5% CO2. Upon reaching
1073 70-80% confluence, the cells were washed with PBS and detached with 0.25%
1074 trypsin/0.2% ethylenediaminetetraacetic acid (EDTA). Cell morphology was viewed under
1075 a light microscope, and the cells were suspended to a concentration of 1 x 106 cells/ml.
1076

1077 **Immunofluorescence and FISH**

1078 Briefly, 10 μm thick frozen sections were treated with 0.2% Triton X-100 in PBS for 15
1079 minutes and blocked with 10% serum in PBST for 40 minutes at room temperature. They
1080 were then incubated with primary antibodies, Anti CK7/serpinB3/NES/GFAP (1:500,
1081 Huabio, ET1609-62, ET7111-21, ET1610-96, ET1601-23) overnight at 4°C and washed
1082 with PBS for 15 min, Alexa-conjugated secondary antibodies (1:1000, iFluor 488 Goat
1083 anti-rabbit IgG antibody, Huabio, HA720120F; iFluor 594 Conjugated Goat anti-rabbit IgG
1084 polyclonal Antibody, Huabio, HA1122) were incubated for 1 hour at room temperature. For
1085 FISH, probes specific to the target sequences (Table S20) were hybridized to the cells
1086 overnight at 37°C. After washing, the cells were counterstained with DAPI and analyzed
1087 under a fluorescence microscope.
1088

1089 **Cell co-culture**

1090 Splenic cells were collected from 6-8 weeks old C57/BL6 mice and resuspended in
1091 RoboSep™ Buffer (STEMCELL Technologies, 20104). Total splenocytes were enriched
1092 for CD4+ T cells with EasySep™ Mouse CD4+ T Cell Isolation Kit (STEMCELL
1093 Technologies, 19852). The selected CD4+ T cells and tumor cells were directly
1094 co-cultured with RPMI 1640 medium (containing 10% FBS) at a ratio of 10:1 for 24 hours,
1095 and suspended T cells were collected at 24 hours for subsequent flow experiments.
1096

1097 **Flow cytometry**

1098 For cell surface staining, cells were stained with the corresponding antibodies, APC
1099 anti-mouse CD45 (1:100, Biolegend, 103112); FITC anti-mouse CD4 (1:100, Biolegend,
1100 100405); PE anti-mouse CD154 (CD40L) (1:50, Biolegend, 157003); eBioscience™
1101 Fixable Viability Dye eFluor™ 506 (1:100, Thermo Fisher Scientific, 65-0866-18), in the
1102 presence of Fc receptor blocker on ice for 30 min, and then fixed with 4%
1103 paraformaldehyde for 10 minutes, washed with 1x PBS solution, and resuspended after
1104 centrifugation. Samples were assessed by a spectral flow cytometer (Cytoflex, BECKMAN
1105 Biosciences,) equipped with SpectroFlo software (Cytek Biosciences). Acquired flow
1106 cytometry results were analyzed by FlowJo software.
1107

1108 **Data availability**

1109 All sequencing data are available at the Genome Sequence Archive for Human:
1110 https://ngdc.cncb.ac.cn/gsa-human/s/QfWwWORT,
1111 https://ngdc.cncb.ac.cn/gsa-human/s/d97GuJ2L and
1112 https://ngdc.cncb.ac.cn/gsa-human/s/MnWe8KbN.
1113

1114 **Code availability**

1115 All code used is open source and available at github:
1116 https://github.com/LinXialab/TGS_somatic_TEI/ and
1117 https://github.com/Goatofmountain/TDScope.
1118

1119 # Reference


1120 1. Swanton, C., Bernard, E., Abbosh, C., Andre, F., Auwerx, J., Balmain, A., Bar-Sagi, D., Bernards,
1121 R., Bullman, S., DeGregori, J., et al. (2024). Embracing cancer complexity: Hallmarks of
1122 systemic disease. Cell *187*, 1589-1616. 10.1016/j.cell.2024.02.009.
1123 2. Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer Discov *12*, 31-46.
1124 10.1158/2159-8290.CD-21-1059.
1125 3. International Cancer Genome, C., Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C.,
1126 Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I., et al. (2010). International network of cancer
1127 genome projects. Nature *464*, 993-998. 10.1038/nature08987.
1128 4. Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.,
1129 Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer
1130 Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113-1120. 10.1038/ng.2764.
1131 5. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael,
1132 J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major
1133 cancer types. Nature *502*, 333-339. 10.1038/nature12634.
1134 6. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R.,

1135 Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation
1136 analysis of cancer genes across 21 tumour types. Nature *505*, 495-501.
1137 10.1038/nature12912.
1138 7. Hutter, C., and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value
1139 beyond Its Data. Cell *173*, 283-285. 10.1016/j.cell.2018.03.042.
1140 8. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A.,
1141 Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization
1142 of Cancer Driver Genes and Mutations. Cell *173*, 371-385 e318. 10.1016/j.cell.2018.02.060.
1143 9. Hornshoj, H., Nielsen, M.M., Sinnott-Armstrong, N.A., Switnicki, M.P., Juul, M., Madsen, T.,
1144 Sallari, R., Kellis, M., Orntoft, T., Hobolth, A., and Pedersen, J.S. (2018). Pan-cancer screen for
1145 mutations in non-coding elements with conservation and cancer specificity reveals
1146 correlations with expression and survival. NPJ Genom Med *3*, 1. 10.1038/s41525-017-0040-5.
1147 10. Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., Hornshoj, H., Hess,
1148 J.M., Juul, R.I., Lin, Z., et al. (2020). Analyses of non-coding somatic drivers in 2,658 cancer
1149 whole genomes. Nature *578*, 102-111. 10.1038/s41586-020-1965-x.
1150 11. Hakobyan, A., Meyenberg, M., Vardazaryan, N., Hancock, J., Vulliard, L., Loizou, J.I., and
1151 Menche, J. (2024). Pan-cancer analysis of the interplay between mutational signatures and
1152 cellular signaling. iScience *27*, 109873. 10.1016/j.isci.2024.109873.
1153 12. Hu, Z., Li, Z., Ma, Z., and Curtis, C. (2020). Multi-cancer analysis of clonality and the timing of
1154 systemic spread in paired primary tumors and metastases. Nat Genet *52*, 701-708.
1155 10.1038/s41588-020-0628-z.
1156 13. Fontana, D., Crespiatico, I., Crippa, V., Malighetti, F., Villa, M., Angaroni, F., De Sano, L., Aroldi,
1157 A., Antoniotti, M., Caravagna, G., et al. (2023). Evolutionary signatures of human cancers
1158 revealed via genomic analysis of over 35,000 patients. Nat Commun *14*, 5982.
1159 10.1038/s41467-023-41670-3.
1160 14. Shah, Y., Verma, A., Marderstein, A.R., White, J., Bhinder, B., Garcia Medina, J.S., and
1161 Elemento, O. (2021). Pan-cancer analysis reveals molecular patterns associated with age. Cell
1162 Rep *37*, 110100. 10.1016/j.celrep.2021.110100.
1163 15. Chen, H., Li, C., Peng, X., Zhou, Z., Weinstein, J.N., Cancer Genome Atlas Research, N., and
1164 Liang, H. (2018). A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient
1165 Samples. Cell *173*, 386-399 e312. 10.1016/j.cell.2018.03.027.
1166 16. Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M.,
1167 Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular
1168 Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291-304 e296.
1169 10.1016/j.cell.2018.03.022.
1170 17. Drews, R.M., Hernando, B., Tarabichi, M., Haase, K., Lesluyes, T., Smith, P.S., Morrill Gavarro,
1171 L., Couturier, D.L., Liu, L., Schneider, M., et al. (2022). A pan-cancer compendium of
1172 chromosomal instability. Nature *606*, 976-983. 10.1038/s41586-022-04789-9.
1173 18. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y.,
1174 Ye, K., Jun, G., Fritz, M.H., et al. (2015). An integrated map of structural variation in 2,504
1175 human genomes. Nature *526*, 75-81. 10.1038/nature15394.
1176 19. Zhang, Y., Yang, L., Kucherlapati, M., Chen, F., Hadjipanayis, A., Pantazi, A., Bristow, C.A., Lee,
1177 E.A., Mahadeshwar, H.S., Tang, J., et al. (2018). A Pan-Cancer Compendium of Genes
1178 Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. Cell Rep *24*,

1179 515-527. 10.1016/j.celrep.2018.06.025.
1180 20. Bielski, C.M., Zehir, A., Penson, A.V., Donoghue, M.T.A., Chatila, W., Armenia, J., Chang, M.T.,
1181 Schram, A.M., Jonsson, P., Bandlamudi, C., et al. (2018). Genome doubling shapes the
1182 evolution and prognosis of advanced cancers. Nat Genet *50*, 1189-1195.
1183 10.1038/s41588-018-0165-1.
1184 21. Consortium, I.T.P.-C.A.o.W.G. (2020). Pan-cancer analysis of whole genomes. Nature *578*,
1185 82-93. 10.1038/s41586-020-1969-6.
1186 22. Hadi, K., Yao, X., Behr, J.M., Deshpande, A., Xanthopoulakis, C., Tian, H., Kudman, S., Rosiene,
1187 J., Darmofal, M., DeRose, J., et al. (2020). Distinct Classes of Complex Structural Variation
1188 Uncovered across Thousands of Cancer Genome Graphs. Cell *183*, 197-210 e132.
1189 10.1016/j.cell.2020.08.006.
1190 23. Ho, S.S., Urban, A.E., and Mills, R.E. (2020). Structural variation in the sequencing era. Nat
1191 Rev Genet *21*, 171-189. 10.1038/s41576-019-0180-9.
1192 24. Xu, L., Wang, X., Lu, X., Liang, F., Liu, Z., Zhang, H., Li, X., Tian, S., Wang, L., and Wang, Z.
1193 (2023). Long-read sequencing identifies novel structural variations in colorectal cancer. PLoS
1194 Genet *19*, e1010514. 10.1371/journal.pgen.1010514.
1195 25. Dan Xie, L.X., Zhoufeng Wang,Tianfu Zeng,Xuenan Pi,Huan Wang,Guonian Zhu,Xinyue
1196 Wu,Yangqian Li,Yan Deng,Yawen Qi,Xuyan liu,Fengmei Zhang,Weimin Li (2023). Landscape of
1197 Large-Scale Somatic Genomic Insertions in Non-Small Cell Lung Carcinoma Revealed by
1198 Nanopore Sequencing. bioRxiv. https://doi.org/10.1101/2023.10.07.561149.
1199 26. Reis, A.L.M., Rapadas, M., Hammond, J.M., Gamaarachchi, H., Stevanovski, I., Ayuputeri
1200 Kumaheri, M., Chintalaphani, S.R., Dissanayake, D.S.B., Siggs, O.M., Hewitt, A.W., et al. (2023).
1201 The landscape of genomic structural variation in Indigenous Australians. Nature *624*, 602-610.
1202 10.1038/s41586-023-06842-7.
1203 27. Cortes-Ciriano, I., Lee, S., Park, W.Y., Kim, T.M., and Park, P.J. (2017). A molecular portrait of
1204 microsatellite instability across multiple cancers. Nat Commun *8*, 15180.
1205 10.1038/ncomms15180.
1206 28. Hause, R.J., Pritchard, C.C., Shendure, J., and Salipante, S.J. (2016). Classification and
1207 characterization of microsatellite instability across 18 cancer types. Nat Med *22*, 1342-1350.
1208 10.1038/nm.4191.
1209 29. Shah, N.M., Jang, H.J., Liang, Y., Maeng, J.H., Tzeng, S.C., Wu, A., Basri, N.L., Qu, X., Fan, C., Li,
1210 A., et al. (2023). Pan-cancer analysis identifies tumor-specific antigens derived from
1211 transposable elements. Nat Genet *55*, 631-639. 10.1038/s41588-023-01349-3.
1212 30. Mehdipour, P., Marhon, S.A., Ettayebi, I., Chakravarthy, A., Hosseini, A., Wang, Y., de Castro,
1213 F.A., Loo Yau, H., Ishak, C., Abelson, S., et al. (2020). Epigenetic therapy induces transcription
1214 of inverted SINEs and ADAR1 dependency. Nature *588*, 169-173.
1215 10.1038/s41586-020-2844-1.
1216 31. Jang, H.S., Shah, N.M., Du, A.Y., Dailey, Z.Z., Pehrsson, E.C., Godoy, P.M., Zhang, D., Li, D., Xing,
1217 X., Kim, S., et al. (2019). Transposable elements drive widespread expression of oncogenes in
1218 human cancers. Nat Genet *51*, 611-617. 10.1038/s41588-019-0373-3.
1219 32. Sun, Z., Zhang, R., Zhang, X., Sun, Y., Liu, P., Francoeur, N., Han, L., Lam, W.Y., Yi, Z., Sebra, R.,
1220 et al. (2022). LINE-1 promotes tumorigenicity and exacerbates tumor progression via
1221 stimulating metabolism reprogramming in non-small cell lung cancer. Mol Cancer *21*, 147.
1222 10.1186/s12943-022-01618-5.

1223 33. Nam, C.H., Youk, J., Kim, J.Y., Lim, J., Park, J.W., Oh, S.A., Lee, H.J., Park, J.W., Won, H., Lee, Y.,
1224 et al. (2023). Widespread somatic L1 retrotransposition in normal colorectal epithelium.
1225 Nature *617*, 540-547. 10.1038/s41586-023-06046-z.
1226 34. Erwin, G.S., Gursoy, G., Al-Abri, R., Suriyaprakash, A., Dolzhenko, E., Zhu, K., Hoerner, C.R.,
1227 White, S.M., Ramirez, L., Vadlakonda, A., et al. (2023). Recurrent repeat expansions in human
1228 cancer genomes. Nature *613*, 96-102. 10.1038/s41586-022-05515-1.
1229 35. Grunewald, T.G., Bernard, V., Gilardi-Hebenstreit, P., Raynal, V., Surdez, D., Aynaud, M.M.,
1230 Mirabeau, O., Cidre-Aranaz, F., Tirode, F., Zaidi, S., et al. (2015). Chimeric EWSR1-FLI1
1231 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. Nat Genet
1232 *47*, 1073-1078. 10.1038/ng.3363.
1233 36. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H.,
1234 Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). Long-read
1235 sequencing of 3,622 Icelanders provides insight into the role of structural variants in human
1236 diseases and other traits. Nat Genet *53*, 779-786. 10.1038/s41588-021-00865-4.
1237 37. Wu, Z., Jiang, Z., Li, T., Xie, C., Zhao, L., Yang, J., Ouyang, S., Liu, Y., Li, T., and Xie, Z. (2021).
1238 Structural variants in the Chinese population and their impact on phenotypes, diseases and
1239 population adaptation. Nat Commun *12*, 6501. 10.1038/s41467-021-26856-x.
1240 38. Gustafson, J.A., Gibson, S.B., Damaraju, N., Zalusky, M.P.G., Hoekzema, K., Twesigomwe, D.,
1241 Yang, L., Snead, A.A., Richmond, P.A., De Coster, W., et al. (2024). High-coverage nanopore
1242 sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of
1243 human genetic variation. Genome Res *34*, 2061-2073. 10.1101/gr.279273.124.
1244 39. Ichikawa, K., Kawahara, R., Asano, T., and Morishita, S. (2023). A landscape of complex
1245 tandem repeats within individual human genomes. Nat Commun *14*, 5530.
1246 10.1038/s41467-023-41262-1.
1247 40. Dolzhenko, E., English, A., Dashnow, H., De Sena Brandine, G., Mokveld, T., Rowell, W.J.,
1248 Karniski, C., Kronenberg, Z., Danzi, M.C., Cheung, W.A., et al. (2024). Characterization and
1249 visualization of tandem repeats at genome scale. Nat Biotechnol *42*, 1606-1614.
1250 10.1038/s41587-023-02057-3.
1251 41. Liu, L., Zhang, J., Wood, S., Newell, F., Leonard, C., Koufariotis, L.T., Nones, K., Dalley, A.J.,
1252 Chittoory, H., Bashirzadeh, F., et al. (2024). Performance of somatic structural variant calling
1253 in lung cancer using Oxford Nanopore sequencing technology. BMC Genomics *25*, 898.
1254 10.1186/s12864-024-10792-3.
1255 42. Sakamoto, Y., Xu, L., Seki, M., Yokoyama, T.T., Kasahara, M., Kashima, Y., Ohashi, A., Shimada,
1256 Y., Motoi, N., Tsuchihara, K., et al. (2020). Long-read sequencing for non-small-cell lung
1257 cancer genomes. Genome Res *30*, 1243-1257. 10.1101/gr.261941.120.
1258 43. Fujimoto, A., Wong, J.H., Yoshii, Y., Akiyama, S., Tanaka, A., Yagi, H., Shigemizu, D., Nakagawa,
1259 H., Mizokami, M., and Shimada, M. (2021). Whole-genome sequencing with long reads
1260 reveals complex structure and origin of structural variation in human genetic variations and
1261 somatic mutations in cancer. Genome Med *13*, 65. 10.1186/s13073-021-00883-1.
1262 44. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*,
1263 3094-3100. 10.1093/bioinformatics/bty191.
1264 45. Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra,
1265 E., Gandhi, M., Hong, K., Pehlivan, D., et al. (2024). Detection of mosaic and population-level
1266 structural variants with Sniffles2. Nat Biotechnol *42*, 1571-1580.

1267 10.1038/s41587-023-02024-y.
1268 46. Chiu, R., Rajan-Babu, I.S., Friedman, J.M., and Birol, I. (2021). Straglr: discovering and
1269 genotyping tandem repeat expansions using whole genome long-read sequences. Genome
1270 Biol *22*, 224. 10.1186/s13059-021-02447-3.
1271 47. Kailing Tu, Q.Z., Yang Li,Yucong Li,Lanfang Yuan,Jie Tang,Lin Xia,Jing Wang,Wei Huang,Dan Xie
1272 (2025). Accurate Somatic SV detection via sequence graph model-based local pan-genome
1273 optimization. bioRxiv. https://doi.org/10.1101/2025.02.11.636543.
1274 48. Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Battle, A., Aganezov, S., and Schatz, M.C. (2023).
1275 Jasmine and Iris: population-scale structural variant comparison and analysis. Nat Methods
1276 *20*, 408-417. 10.1038/s41592-022-01753-3.
1277 49. Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate de novo genome
1278 assembly from long uncorrected reads. Genome Res *27*, 737-746. 10.1101/gr.214270.116.
1279 50. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements
1280 in genomic sequences. Curr Protoc Bioinformatics *Chapter 4*, Unit 4 10.
1281 10.1002/0471250953.bi0410s25.
1282 51. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic
1283 Acids Res *27*, 573-580. 10.1093/nar/27.2.573.
1284 52. Morgulis, A., Gertz, E.M., Schaffer, A.A., and Agarwala, R. (2006). A fast and symmetric DUST
1285 implementation to mask low-complexity DNA sequences. J Comput Biol *13*, 1028-1040.
1286 10.1089/cmb.2006.13.1028.
1287 53. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and
1288 Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat
1289 Biotechnol *28*, 495-501. 10.1038/nbt.1630.
1290 54. Gao, Y., Ferguson, D.O., Xie, W., Manis, J.P., Sekiguchi, J., Frank, K.M., Chaudhuri, J., Horner, J.,
1291 DePinho, R.A., and Alt, F.W. (2000). Interplay of p53 and DNA-repair protein XRCC4 in
1292 tumorigenesis, genomic stability and development. Nature *404*, 897-900. 10.1038/35009138.
1293 55. Bhattacharya, S., Srinivasan, K., Abdisalaam, S., Su, F., Raj, P., Dozmorov, I., Mishra, R.,
1294 Wakeland, E.K., Ghose, S., Mukherjee, S., and Asaithamby, A. (2017). RAD51 interconnects
1295 between DNA replication, DNA repair and immunity. Nucleic Acids Res *45*, 4590-4605.
1296 10.1093/nar/gkx126.
1297 56. Coin, F., Auriol, J., Tapias, A., Clivio, P., Vermeulen, W., and Egly, J.M. (2004). Phosphorylation
1298 of XPB helicase regulates TFIIH nucleotide excision repair activity. EMBO J *23*, 4835-4846.
1299 10.1038/sj.emboj.7600480.
1300 57. Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E.,
1301 Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019). Characterizing the Major
1302 Structural Variant Alleles of the Human Genome. Cell *176*, 663-675 e619.
1303 10.1016/j.cell.2018.12.019.
1304 58. Niu, Y., Teng, X., Zhou, H., Shi, Y., Li, Y., Tang, Y., Zhang, P., Luo, H., Kang, Q., Xu, T., and He, S.
1305 (2022). Characterizing mobile element insertions in 5675 genomes. Nucleic Acids Res *50*,
1306 2493-2508. 10.1093/nar/gkac128.
1307 59. Pascarella, G., Hon, C.C., Hashimoto, K., Busch, A., Luginbuhl, J., Parr, C., Hin Yip, W., Abe, K.,
1308 Kratz, A., Bonetti, A., et al. (2022). Recombination of repeat elements generates somatic
1309 complexity in human genomes. Cell *185*, 3025-3040 e3026. 10.1016/j.cell.2022.06.032.
1310 60. Liu, T., Wu, L., Liu, J., Chen, H., Zhu, B., Qiao, D., Zhu, Y., Liu, T., Chen, Q., and Hu, A. (2023).

1311 Comprehensive characterization of MUC16 mutations in lung adenocarcinoma for
1312 immunotherapies and prognosis: An observational study. Medicine (Baltimore) *102*, e35481.
1313 10.1097/MD.0000000000035481.
1314 61. Shah, A., Chaudhary, S., Lakshmanan, I., Aithal, A., Kisling, S.G., Sorrell, C., Marimuthu, S.,
1315 Gautam, S.K., Rauth, S., Kshirsagar, P., et al. (2023). Chimeric antibody targeting unique
1316 epitope on onco-mucin16 reduces tumor burden in pancreatic and lung malignancies. NPJ
1317 Precis Oncol *7*, 74. 10.1038/s41698-023-00423-7.
1318 62. Marimuthu, S., Lakshmanan, I., Muniyan, S., Gautam, S.K., Nimmakayala, R.K., Rauth, S., Atri,
1319 P., Shah, A., Bhyravbhatla, N., Mallya, K., et al. (2022). MUC16 Promotes Liver Metastasis of
1320 Pancreatic Ductal Adenocarcinoma by Upregulating NRP2-Associated Cell Adhesion. Mol
1321 Cancer Res *20*, 1208-1221. 10.1158/1541-7786.MCR-21-0888.
1322 63. Skates, S.J., Greene, M.H., Buys, S.S., Mai, P.L., Brown, P., Piedmonte, M., Rodriguez, G.,
1323 Schorge, J.O., Sherman, M., Daly, M.B., et al. (2017). Early Detection of Ovarian Cancer using
1324 the Risk of Ovarian Cancer Algorithm with Frequent CA125 Testing in Women at Increased
1325 Familial Risk - Combined Results from Two Screening Trials. Clin Cancer Res *23*, 3628-3637.
1326 10.1158/1078-0432.CCR-15-2750.
1327 64. Caputo, M., Xia, Y., Anand, S.K., Cansby, E., Andersson, E., Marschall, H.U., Konigsrainer, A.,
1328 Peter, A., and Mahlapuu, M. (2023). STE20-type kinases MST3 and MST4 promote the
1329 progression of hepatocellular carcinoma: Evidence from human cell culture and expression
1330 profiling of liver biopsies. FASEB J *37*, e23105. 10.1096/fj.202300397RR.
1331 65. Huang, N., Lin, W., Shi, X., and Tao, T. (2018). STK24 expression is modulated by DNA copy
1332 number/methylation in lung adenocarcinoma and predicts poor survival. Future Oncol *14*,
1333 2253-2263. 10.2217/fon-2018-0126.
1334 66. Li, L., Chen, S., Tang, Y., Wu, J., He, Y., and Qiu, L. (2022). Oncogene or tumor suppressor gene:
1335 An integrated pan-cancer analysis of NBPF1. Front Endocrinol (Lausanne) *13*, 950326.
1336 10.3389/fendo.2022.950326.
1337 67. Zhang, J., Zhao, Y., Wang, J., Sneh, T., Yu, Q., Zhou, X., and Gong, C. (2020). NBPF1
1338 independently determine the risk stratification and prognosis of patients with neuroblastoma.
1339 Genomics *112*, 3951-3957. 10.1016/j.ygeno.2020.06.038.
1340 68. Li, J., Xian, L., Zhu, Z., Wang, Y., Zhang, W., Zheng, R., Xue, W., and Li, J. (2023). Role of CELF2
1341 in ferroptosis: Potential targets for cancer therapy (Review). Int J Mol Med *52*.
1342 10.3892/ijmm.2023.5291.
1343 69. Jen, J., and Wang, Y.C. (2016). Zinc finger proteins in cancer progression. J Biomed Sci *23*, 53.
1344 10.1186/s12929-016-0269-9.
1345 70. Frankson, R., Yu, Z.H., Bai, Y., Li, Q., Zhang, R.Y., and Zhang, Z.Y. (2017). Therapeutic Targeting
1346 of Oncogenic Tyrosine Phosphatases. Cancer Res *77*, 5701-5705.
1347 10.1158/0008-5472.CAN-17-1510.
1348 71. Stanford, S.M., and Bottini, N. (2023). Targeting protein phosphatases in cancer
1349 immunotherapy and autoimmune disorders. Nat Rev Drug Discov *22*, 273-294.
1350 10.1038/s41573-022-00618-w.
1351 72. Fang, L., Liu, Q., Monteys, A.M., Gonzalez-Alegre, P., Davidson, B.L., and Wang, K. (2022).
1352 DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore
1353 sequencing. Genome Biol *23*, 108. 10.1186/s13059-022-02670-6.
1354 73. Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J.,

1355 Santamarina, M., Ju, Y.S., Temes, J., Garcia-Souto, D., et al. (2020). Pan-cancer analysis of
1356 whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat
1357 Genet *52*, 306-319. 10.1038/s41588-019-0562-0.
1358 74. Helman, E., Lawrence, M.S., Stewart, C., Sougnez, C., Getz, G., and Meyerson, M. (2014).
1359 Somatic retrotransposition in human cancer revealed by whole-genome and exome
1360 sequencing. Genome Res *24*, 1053-1063. 10.1101/gr.163659.113.
1361 75. Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R.,
1362 Wilson, R.K., and Eichler, E.E. (2010). A human genome structural variation sequencing
1363 resource reveals insights into mutational mechanisms. Cell *143*, 837-847.
1364 10.1016/j.cell.2010.10.027.
1365 76. Horton, C.A., Alexandari, A.M., Hayes, M.G.B., Marklund, E., Schaepe, J.M., Aditham, A.K.,
1366 Shah, N., Suzuki, P.H., Shrikumar, A., Afek, A., et al. (2023). Short tandem repeats bind
1367 transcription factors to tune eukaryotic gene expression. Science *381*, eadd1250.
1368 10.1126/science.add1250.
1369 77. Depienne, C., and Mandel, J.L. (2021). 30 years of repeat expansion disorders: What have we
1370 learned and what are the remaining challenges? Am J Hum Genet *108*, 764-785.
1371 10.1016/j.ajhg.2021.03.011.
1372 78. Dubois, F., Sidiropoulos, N., Weischenfeldt, J., and Beroukhim, R. (2022). Structural variations
1373 in cancer and the 3D genome. Nat Rev Cancer *22*, 533-546. 10.1038/s41568-022-00488-9.
1374 79. Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., and Timp, W. (2017).
1375 Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods *14*, 407-410.
1376 10.1038/nmeth.4184.
1377 80. Guo, C.T., Jardin, B.D., Lin, J.S., Ambroise, R.L., Wang, Z., Yang, L.Z., Mazumdar, N., Lu, F.J., Ma,
1378 Q., Cao, Y.P., et al. (2024). In vivo proximity proteomics uncovers palmdelphin (PALMD) as a
1379 Z-disc-associated mitigator of isoproterenol-induced cardiac injury. Acta Pharmacol Sin *45*,
1380 2540-2552. 10.1038/s41401-024-01348-y.
1381 81. Sainz-Jaspeado, M., Smith, R.O., Plunde, O., Pawelzik, S.C., Jin, Y., Nordling, S., Ding, Y.,
1382 Aspenstrom, P., Hedlund, M., Bastianello, G., et al. (2021). Palmdelphin Regulates Nuclear
1383 Resilience to Mechanical Stress in the Endothelium. Circulation *144*, 1629-1645.
1384 10.1161/CIRCULATIONAHA.121.054182.
1385 82. Yuan, J., Chen, L., Xiao, J., Qi, X.K., Zhang, J., Li, X., Wang, Z., Lian, Y.F., Xiang, T., Zhang, Y., et al.
1386 (2019). SHROOM2 inhibits tumor metastasis through RhoA-ROCK pathway-dependent and
1387 -independent mechanisms in nasopharyngeal carcinoma. Cell Death Dis *10*, 58.
1388 10.1038/s41419-019-1325-7.
1389 83. Farber, M.J., Rizaldy, R., and Hildebrand, J.D. (2011). Shroom2 regulates contractility to
1390 control endothelial morphogenesis. Mol Biol Cell *22*, 795-805. 10.1091/mbc.E10-06-0505.
1391 84. Yu, J.E., Han, S.Y., Wolfson, B., and Zhou, Q. (2018). The role of endothelial lipase in lipid
1392 metabolism, inflammation, and cancer. Histol Histopathol *33*, 1-10. 10.14670/HH-11-905.
1393 85. Wang, J.W., Gamsby, J.J., Highfill, S.L., Mora, L.B., Bloom, G.C., Yeatman, T.J., Pan, T.C., Ramne,
1394 A.L., Chodosh, L.A., Cress, W.D., et al. (2004). Deregulated expression of LRBA facilitates
1395 cancer cell growth. Oncogene *23*, 4089-4097. 10.1038/sj.onc.1207567.
1396 86. Liu, Y., Zheng, X., Yu, Q., Wang, H., Tan, F., Zhu, Q., Yuan, L., Jiang, H., Yu, L., and Zeng, S.
1397 (2016). Epigenetic activation of the drug transporter OCT2 sensitizes renal cell carcinoma to
1398 oxaliplatin. Sci Transl Med *8*, 348ra397. 10.1126/scitranslmed.aaf3124.

87. Man, K.F., Zhou, L., Yu, H., Lam, K.H., Cheng, W., Yu, J., Lee, T.K., Yun, J.P., Guan, X.Y., Liu, M., and Ma, S. (2023). SPINK1-induced tumor plasticity provides a therapeutic window for chemotherapy in hepatocellular carcinoma. Nat Commun *14*, 7863. 10.1038/s41467-023-43670-9.

88. Reith, W., LeibundGut-Landmann, S., and Waldburger, J.M. (2005). Regulation of MHC class II gene expression by the class II transactivator. Nat Rev Immunol *5*, 793-806. 10.1038/nri1708.

89. Lin, J., Jia, P., Wang, S., Kosters, W., and Ye, K. (2023). Comparison and benchmark of structural variants detected from long read and long-read assembly. Brief Bioinform *24*. 10.1093/bib/bbad188.

90. Sahlin, K., Baudeau, T., Cazaux, B., and Marchet, C. (2023). A survey of mapping algorithms in the long-reads era. Genome Biol *24*, 133. 10.1186/s13059-023-02972-3.

91. Hickey, G., Monlong, J., Ebler, J., Novak, A.M., Eizenga, J.M., Gao, Y., Human Pangenome Reference, C., Marschall, T., Li, H., and Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol *42*, 663-673. 10.1038/s41587-023-01793-w.

92. Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. Genome Biol *21*, 265. 10.1186/s13059-020-02168-z.

93. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44-53. 10.1126/science.abj6987.

94. Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., et al. (2022). Segmental duplications and their variation in a complete human genome. Science *376*, eabj6965. 10.1126/science.abj6965.

95. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol *36*, 338-345. 10.1038/nbt.4060.

96. Li, H., Li, K., Shu, D., Shen, M., Tan, Z., Zhang, W., Pu, D., Tan, W., Tang, Z., Jin, A., and Liu, S. (2022). MED16 Promotes Tumour Progression and Tamoxifen Sensitivity by Modulating Autophagy through the mTOR Signalling Pathway in ER-Positive Breast Cancer. Life (Basel) *12*. 10.3390/life12101461.

97. Brahmer, J., Reckamp, K.L., Baas, P., Crino, L., Eberhardt, W.E., Poddubskaya, E., Antonia, S., Pluzanski, A., Vokes, E.E., Holgado, E., et al. (2015). Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. N Engl J Med *373*, 123-135. 10.1056/NEJMoa1504627.

98. Battaglia, S., Dong, K., Wu, J., Chen, Z., Najm, F.J., Zhang, Y., Moore, M.M., Hecht, V., Shoresh, N., and Bernstein, B.E. (2022). Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. Nat Genet *54*, 1504-1513. 10.1038/s41588-022-01188-8.

99. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol *37*, 907-915. 10.1038/s41587-019-0201-4.

100. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079. 10.1093/bioinformatics/btp352.

101. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015).

1443 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat
1444 Biotechnol *33*, 290-295. 10.1038/nbt.3122.
1445 102. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with
1446 high-throughput sequencing data. Bioinformatics *31*, 166-169.
1447 10.1093/bioinformatics/btu638.
1448 103. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and
1449 dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550. 10.1186/s13059-014-0550-8.
1450 104. De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack:
1451 visualizing and processing long-read sequencing data. Bioinformatics *34*, 2666-2669.
1452 10.1093/bioinformatics/bty149.
1453 105. Su, J., Zheng, Z., Ahmed, S.S., Lam, T.W., and Luo, R. (2022). Clair3-trio: high-performance
1454 Nanopore long-read variant calling in family trios with trio-to-trio deep neural networks. Brief
1455 Bioinform *23*. 10.1093/bib/bbac301.
1456 106. Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey,
1457 I., Delattre, O., and Barillot, E. (2012). Control-FREEC: a tool for assessing copy number and
1458 allelic content using next-generation sequencing data. Bioinformatics *28*, 423-425.
1459 10.1093/bioinformatics/btr670.
1460 107. do Valle, I.F., Giampieri, E., Simonetti, G., Padella, A., Manfrini, M., Ferrari, A., Papayannidis,
1461 C., Zironi, I., Garonzi, M., Bernardi, S., et al. (2016). Optimized pipeline of MuTect and GATK
1462 tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome
1463 sequencing data. BMC Bioinformatics *17*, 341. 10.1186/s12859-016-1190-7.
1464 108. Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C.,
1465 Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in
1466 human cancer. Nat Biotechnol *30*, 413-421. 10.1038/nbt.2203.
1467 109. Morishita, S., Ichikawa, K., and Myers, E.W. (2021). Finding long tandem repeats in long noisy
1468 reads. Bioinformatics *37*, 612-621. 10.1093/bioinformatics/btaa865.

1469

1470 ## Acknowledgments

1471 We thank members of the Dan Xie lab for useful discussions. This work was supported by
1472 the National Natural Science Foundation of China 82173383 and 82241238 (to D.X.),
1473 32200508 (to L.X.), 32300522 (to K.L.T), 32300003 (to H.W.), 82371883 (to J.F.L.),
1474 82472365 (to H.T.L.), the Sichuan Province Science and Technology Program
1475 2025ZNSFSC1098 (to K.L.T.), and the 1·3·5 project for disciplines of excellence, West
1476 China Hospital, Sichuan University ZYYC23024 (to D.X.).

1477

1478 ## Author contributions:

1479 Conceptualization: D.X.; Data Curation: W.L., Z.F.W., Z.W., C.L., J.F.L., Y.X.Q., Q.W.Z.,
1480 H.T.L., S.Z.H., Z.Y.Z., A.L., A.L.W.; Software, Formal Analysis and Methodology: L.X.,

K.L.T., X.Y.L., Q.L.Z., R.L.W., Z.T.F., X.N.P.; Validation: H.W., M.L.L., Y.H., J.W.; Supervision: D.X.; Writing – Original Draft: D.X., L.X., K.L.T., H.W., X.Y.L., Y.H.L., C.Y.Y., Z.T.F.; Writing – Review & Editing: all authors.

## Declaration of interests

The authors declare no competing interests.

# Figure Legends

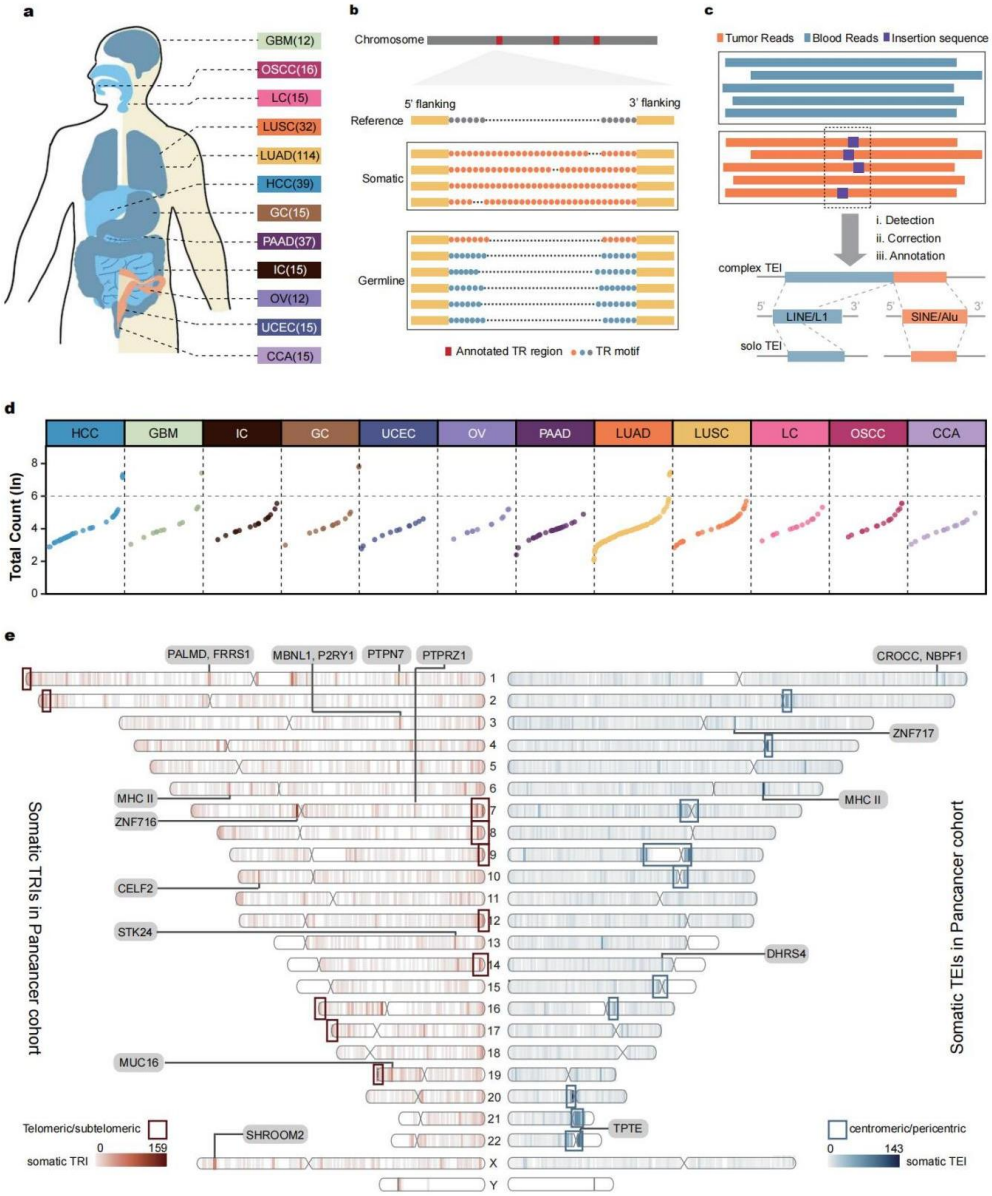



**Figure 1. Genome-wide detection of somatic insertions in pan-cancer genomes.**

a. Overview of a pan-cancer cohort comprising 325 paired tumor and blood samples from 299 patients across 12 cancer types: 1) glioblastoma (GBM); 2) oral squamous cell carcinoma (OSCC); 3) laryngeal cancer (LC); 4) lung adenocarcinoma (LUAD); 5) lung squamous cell carcinoma (LUSC); 6) hepatocellular carcinoma (HCC); 7) gastric carcinoma (GC); 8) pancreatic adenocarcinoma (PAAD); 9) intestinal carcinoma (IC);

1498 10) ovarian cancer (OV); 11) uterine corpus endometrial carcinoma (UCEC); 12)
1499 cervical squamous cell carcinoma (CCA).

1500 b. Scheme of the bioinformatics pipeline used to identify somatic TRIs. The top track
1501 ('reference') shows the hg38 genome. the second track ('somatic') shows reads
1502 alignments supporting somatic SVs mapped to the reference genome, and the third
1503 track ('germline') shows reads alignments supporting germline SVs mapped to the
1504 reference genome. Circles represent repeat motifs, with different colors indicating
1505 different repeat motifs.

1506 c. Scheme of the bioinformatics pipeline for identifying somatic TEI. The first blue track
1507 represents alignments from the tumor sample mapped to the reference genome,
1508 while the second orange track shows alignments from the paired blood sample. After
1509 initial detection (Method), inserted sequences are corrected. Following annotation
1510 (Method), somatic TEIs are classified into two categories: complex TEIs (upper track)
1511 or solo TEIs (lower track).

1512 d. Somatic insertion frequency (ln) across cancer types. Somatic insertion frequencies
1513 were calculated based on the total number of somatic TRIs and TEIs per sample.

1514 e. Ideogram depicting the distribution of somatic TEI (blue, left) and somatic TRI (red,
1515 right). The ideogram shows the genomic distribution of somatic TEIs (blue) and
1516 somatic TRIs (red), with darker colors indicating higher frequencies of TEIs or TRIs
1517 within each 1Mb window. The top 30 high-frequency 1Mb windows are highlighted
1518 according to the following features: regions with clustered TEs near characteristic
1519 genes are marked with a gray background; high-frequency windows located in
1520 telomeric, sub-telomeric, centromeric regions are enclosed in a box.

1521

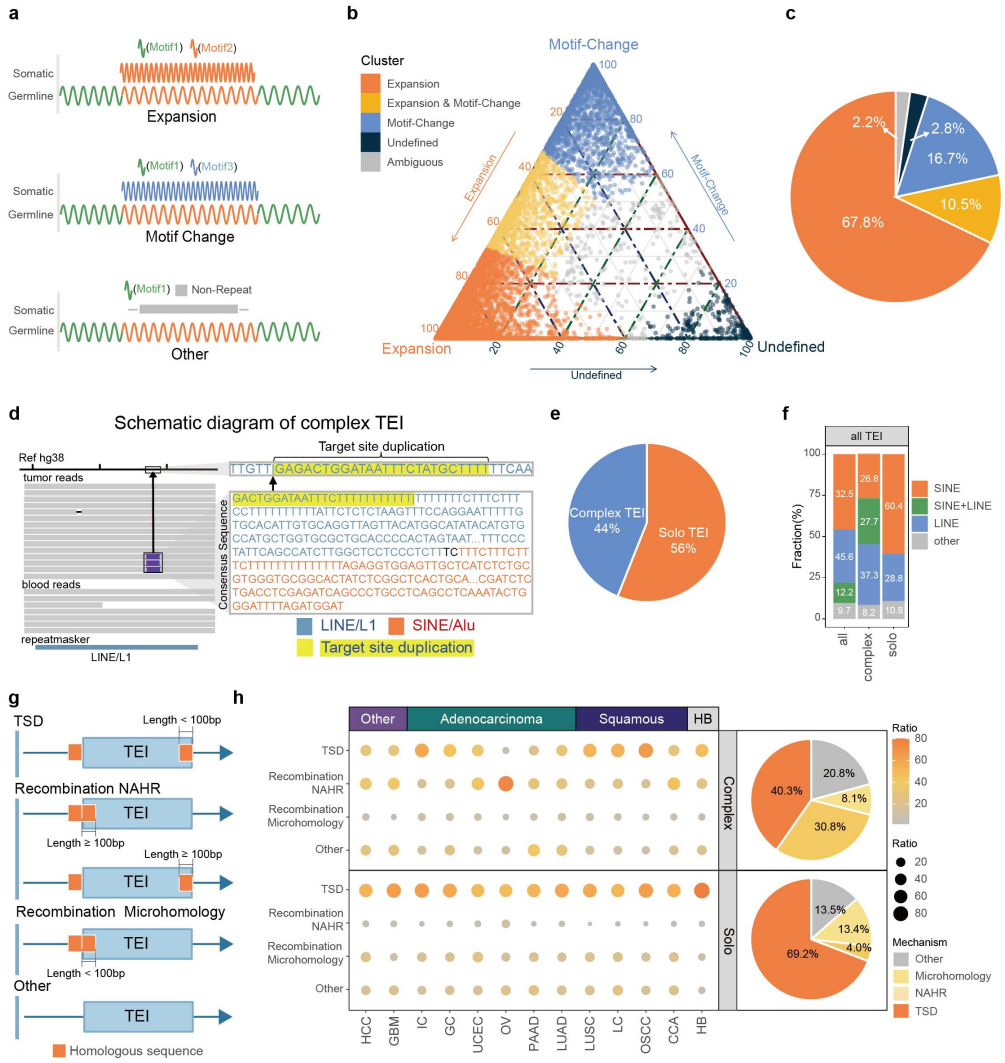



**Figure 2. Sequence features and mechanisms of somatic TRIs and somatic TEIs.**

a. Schematic representation of three somatic TRI motif mutation patterns, namely expansion, motif change and other, arranged from top to bottom. In each subpanel, the first row illustrates the somatic repeat patterns, whereas the second row presents the corresponding germline repeat patterns. Different repeat motifs are color-coded for clarity.

b. Ternary phase diagram illustrating the relative proportions of expansion, motif-change, and other events in somatic TRIs. Categories include: Expansion (orange), Expansion & Motif-Change (yellow), Motif-Change (blue), Undefined (dark blue), and Ambiguous (gray).

1534 c. Proportion of the five categories of somatic TRI: Expansion, Expansion &
1535 Motif-Change, Motif-Change, Undefined, and Ambiguous.
1536 d. Schematic representation of a somatic complex TEI sequence. The color coding of
1537 the bases indicates their origin, with yellow-highlighted sequences representing
1538 target site duplication (TSD) sequences. For LINE/L1 sequences, 5755 base pairs
1539 were omitted, and for SINE/Alu sequences, 141 base pairs were omitted.
1540 e. Pie charts shows the proportion of somatic complex (blue) and solo (orange) TEI.
1541 f. Proportion of different TE families within all somatic TEIs. 'SINE' indicates TEIs
1542 containing SINE elements but lacking LINE elements; 'LINE' indicates TEIs
1543 containing LINE elements but lacking SINE elements; 'SINE+LINE' denotes TEIs
1544 containing both SINE and LINE elements.
1545 g. Schematic illustration of somatic TEI insertion mechanisms.
1546 h. Summary of somatic TEI insertion mechanisms. The bubble chart shows the relative
1547 proportions of insertion mechanisms for complex and solo TEIs across various
1548 cancer types, with bubble size and color intensity corresponding to the magnitude of
1549 these proportions. A pie chart summarizes the overall distribution of insertion
1550 mechanisms for complex and solo somatic TEIs. TSD: target site duplication; NAHR:
1551 non-allelic homologous recombination; microhomology: microhomology
1552 recombination.
1553

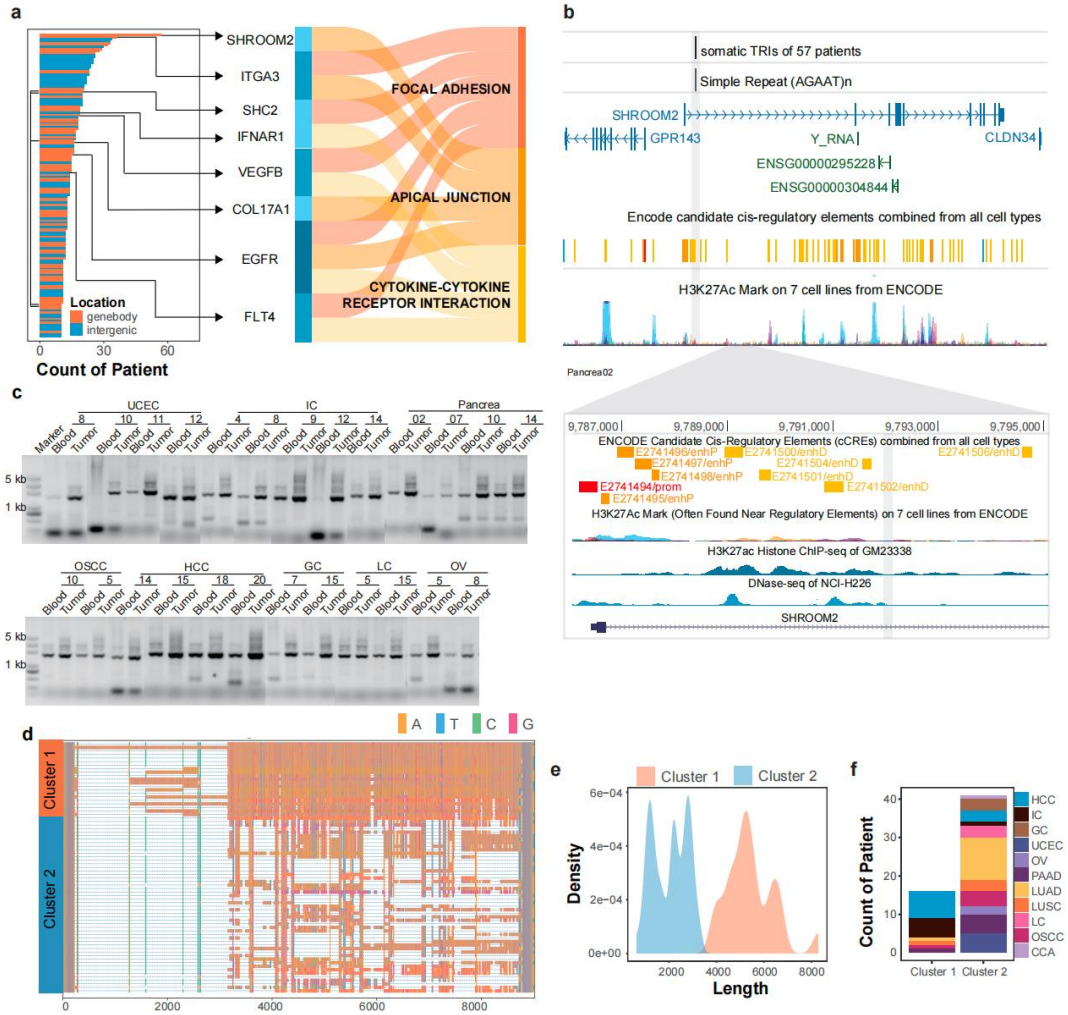



**Figure 3. High-frequency somatic TRIs located in the intron of *SHROOM2*.**

a. Bar plot showing the number of patients harbored somatic TRIs in each high-frequency TRI window (>= 10 patients), along with the pathways enrichment analysis of genes located within or near these high-frequency somatic TRI windows.

b. Genome browser tracks displaying the ENCODE-annotated genes, and ENCODE-predicted cis-regulatory elements (cCREs) surrounding somatic TRIs in the SHROOM2 gene locus. Lower panel: Epigenomic signals around somatic TRIs in SHROOM2, as visualized by chromatin marks.

c. Detecting somatic TRI events harbored by *SHROOM2* through PCR in clinical samples, blood DNA as a control.

d. The detailed sequence overview of somatic TRIs harbored by *SHROOM2* in different

1567 clusters. The ATCG bases are colored differently, with gaps represented as short
1568 dashes.
1569 e. Distribution of somatic TRI lengths in *SHROOM2* across different clusters.
1570 f. Bar chart showing the distribution of cancer types in patients harboring somatic TRIs
1571 within SHROOM2 across the different clusters.
1572

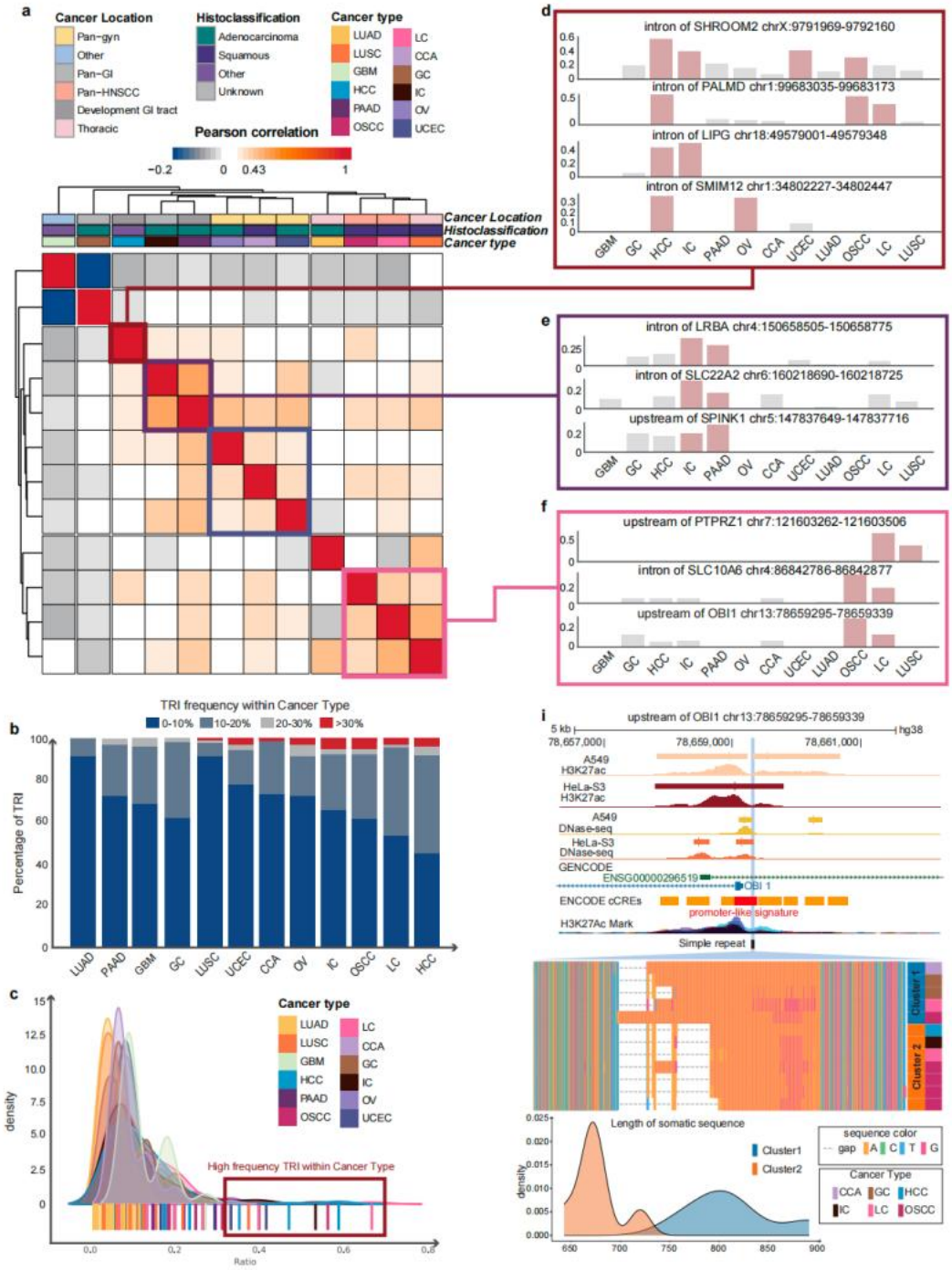



**Figure 4. The characteristic of somatic TRI clusters**

a. Clustering heatmap of somatic TRI frequencies across cancer types.

b. Proportion of tumor samples with somatic TRIs. The number represents the proportion of mutations within the cancer type. Different colors represent various percentage categories.

1580 c. Frequency distribution of somatic TRIs across cancer types. The bars at the bottom
1581 represent the mutation frequency of the mutation within individual cancer types, with
1582 colors indicating different cancer types.
1583 d. Bar plots showing the frequency of HCC associated high-frequency somatic TRI
1584 across cancer types. The high frequency cancer types (over 30%) are filled in red.
1585 e-f. Bar plots showing the frequency distribution of high-frequency somatic TRI in
1586 IC&PAAD (e), as well as OSCC, LC and LUSC (f) clusters compared to other cancer types.
1587 Bars of significant cancer types are filled in red.
1588 g. Browser plot displaying the epigenomic signal, ENCODE-annotated gene transcript,
1589 and cCREs (Candidate Cis-Regulatory Elements) predicted by ENCODE around the
1590 upstream somatic TRIs of the OBI1 gene. The detailed sequence overview of different
1591 clusters surrounding the upstream somatic TRIs of OBI1 is presented in the lower
1592 enlarged panel, with the length distribution of these 2 clusters color-coded.
1593

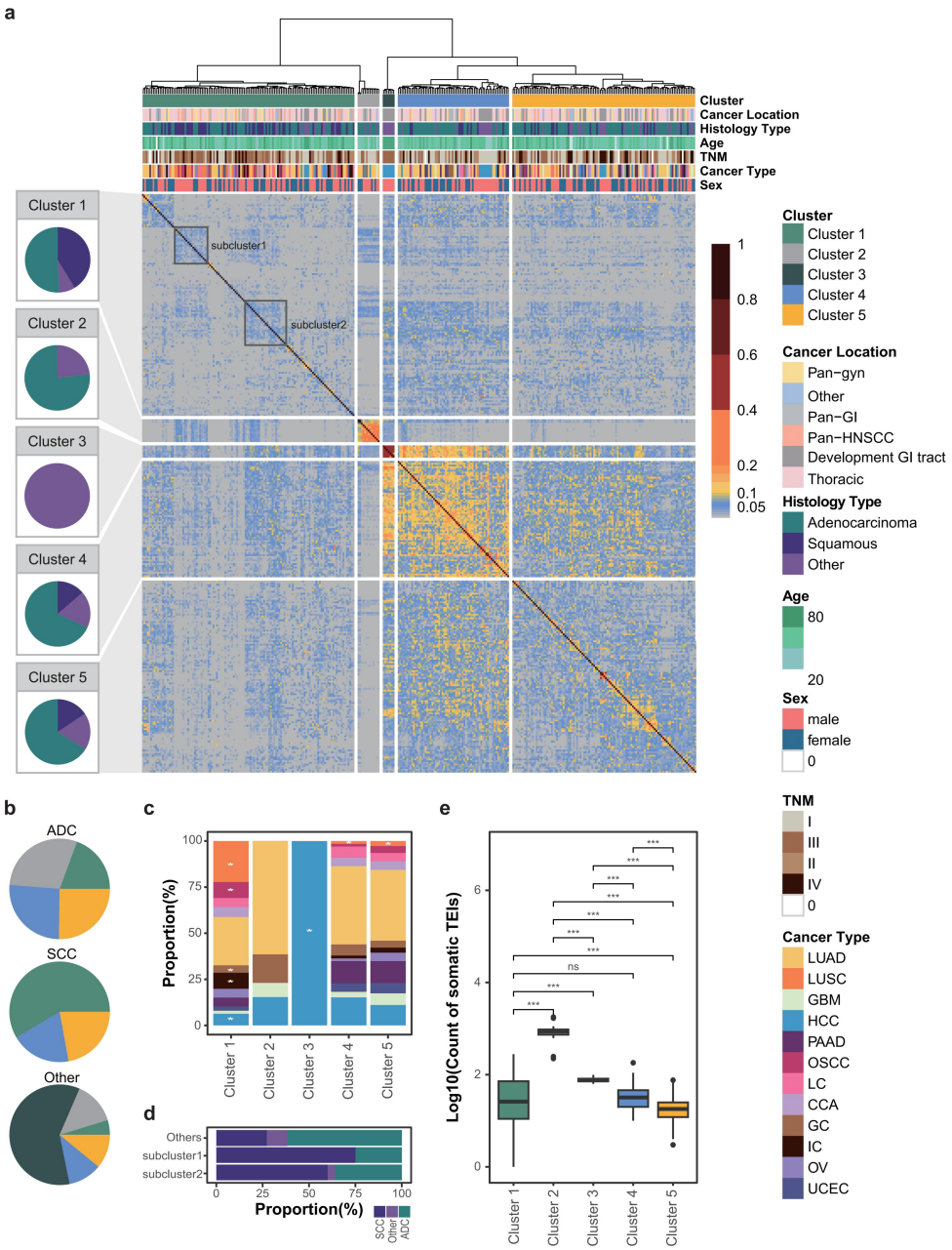



**Figure 5. Clusters of somatic TEIs.**

a. Heatmap showing the clustering results of somatic TEIs. Each column and each row represent Jaccard similarities of a sample compared to others. Hierarchical clustering was performed according to Jaccard similarity with Ward.D linkage. Each cluster has a pie chart to indicate the proportions of adenocarcinoma, squamous cell carcinoma, and others. Two SCC-subclusters of cluster 1 are distinguished.

b. Pies showing the proportional of 5 clusters within the three major histology categories:

1603 SCC (squamous cell carcinoma), ADC (adenocarcinoma), and Others.

1604 c. A stacked bar chart displays the distribution percentages of different cancer types
1605 across five clusters.

1606 d. Proportional distribution of ADC (adenocarcinoma), SCC (squamous cell carcinoma),
1607 and others within subclusters of Cluster 1.

1608 e. A boxplot depicting the number of somatic TEIs (log10) within 5 clusters, with the
1609 p-value calculated using t-test. ns $p>0.05$, *$p < 0.05$, **$p < 0.005$, ***$p < 0.0005$

1610

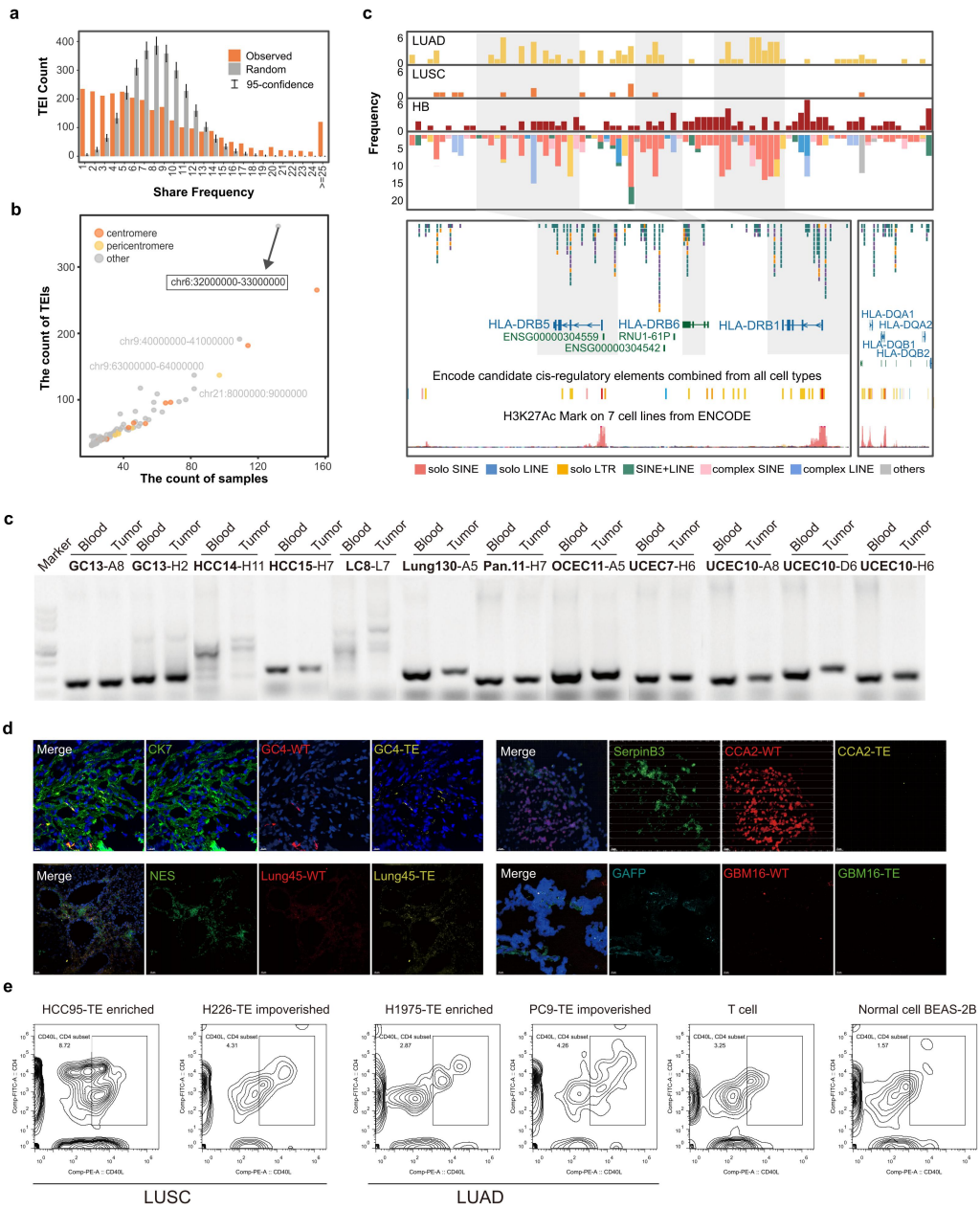



**Figure 6. The MHC II gene family harbored high frequency somatic TRIs.**

a. Distribution of the number of shared TEI samples. Gray: random distribution (Methods); Orange: observed distribution. The x-axis represents the number of samples sharing a somatic TEI, and the y-axis represents the count of somatic TEIs.

b. Scatter plot illustrating the number of samples sharing somatic TEIs and the count of somatic TEIs. Each dot represents a 1Mb genomic window, with the color indicating the source of its chromosomal coordinates. The genomic window containing the highest number of TEIs, chr6:32,000,000-33,000,000, is highlighted.

1621 c. Bar plots showing the number of somatic transposable element insertions (TEIs)
1622 within 500 bp windows in LUAD (first track), LUSC (second track), and high-burden
1623 samples (third track). The fourth track displays the cumulative number of somatic
1624 TEIs within 500 bp windows, with bars color-coded according to the type of TEI. Each
1625 bar represents a 500 bp window. The browser plot includes epigenomic signals,
1626 ENCODE-annotated gene transcripts, and candidate cis-regulatory elements (cCREs)
1627 predicted by ENCODE around the somatic TEIs within the MHC II gene family.
1628 d. Detecting somatic TEIs events through PCR in some clinical samples, blood DNA as
1629 a control (A: ALL; H: HCC; L: LC; D: DRB5).
1630 e. Detecting somatic TEIs events through FISH in in some sample tissue slices,
1631 CK7/serpinB3/NES/GAFP is the marker for GC/CCA/LUNG/GBM respectively. WT
1632 refers to the sequence without somatic TEIs events, TE contains the sequence of
1633 somatic TEIs events. Flow cytometric analyses of CD40L mobilization in CD4+ T cells
1634 after 24 hours of in vitro co-culture with tumor cells.
1635

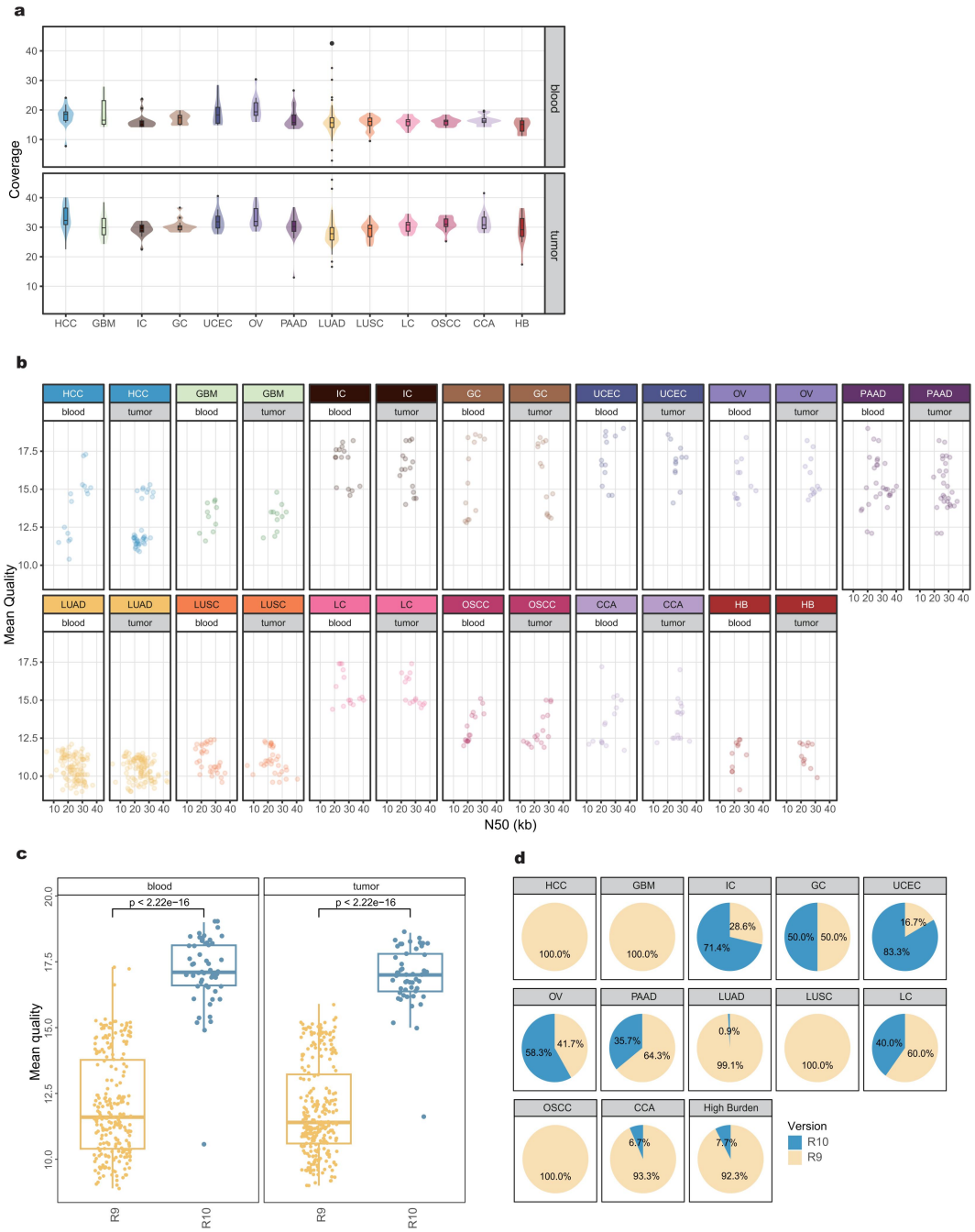



**Figure S1. Summary of long-read sequencing data.**

a. Sequencing depth of long-read data from 325 paired tumor (below) and paired blood (above) samples across 12 cancer types and high-burden samples.

b. Scatter plots showing the N50 and mean quality of LRS reads derived from paired tumor (gray) and blood (white) samples, spanning 12 cancer types and high-burden samples. Each dot represents a sample.

1644 c. Box plots comparing sequencing quality between R9 (blue) and R10 (yellow) chips,
1645 with statistical significance assessed by the Wilcoxon rank test. Each dot represents
1646 a sample.
1647 d. Pie chart illustrating the distribution of sequencing chip types used across various
1648 cancer types and high-burden samples. Each color represents a different chip
1649 version.

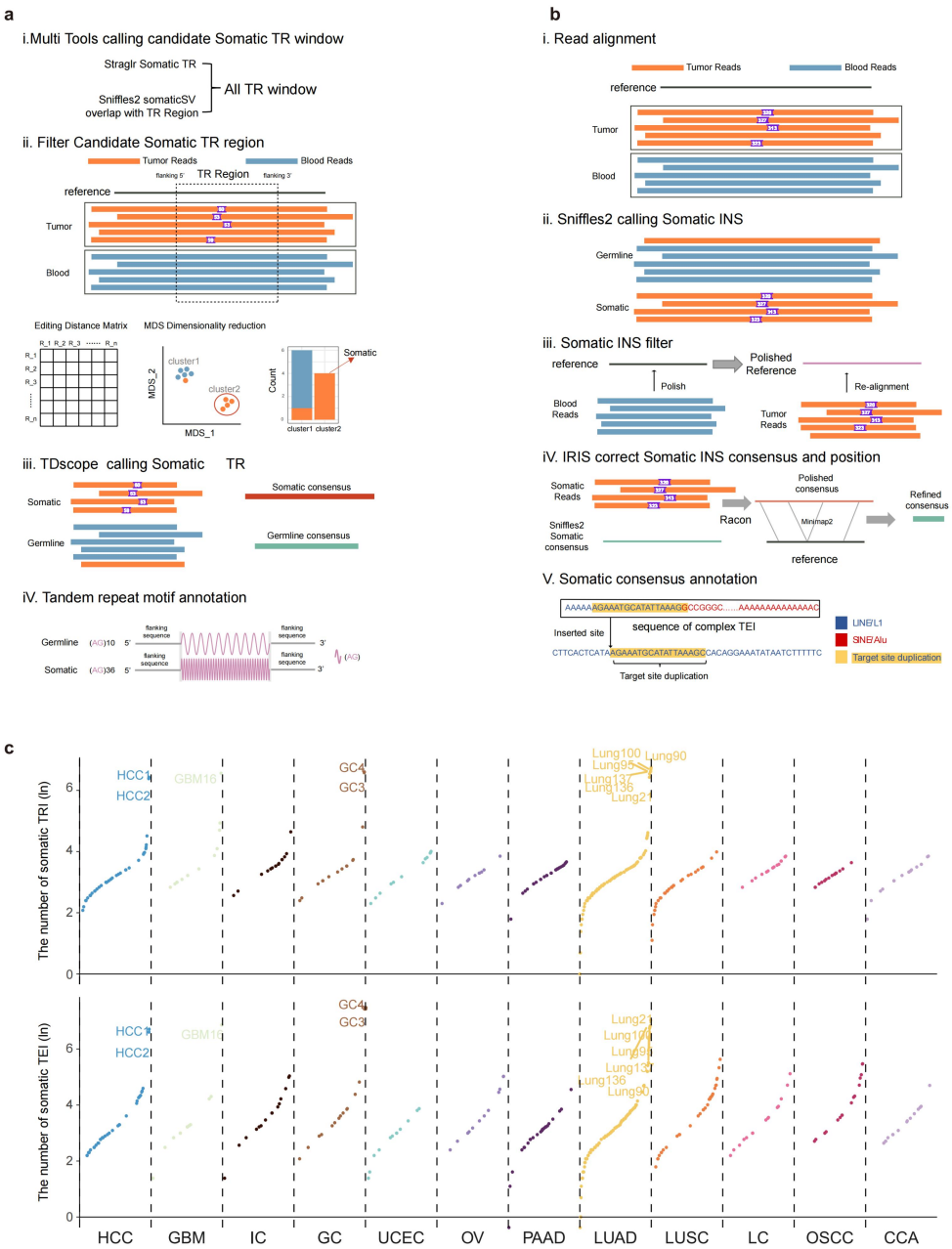



**Figure S2. Overview of Methods for Identifying Somatic Genomic Alterations and The Frequency of Somatic TRIs/TEIs.**

a. Schematic representation of the workflow for identifying somatic TRI. i. Candidate somatic TRI regions are identified using Sniffles2 and Straglr; ii. Reads spanned these regions are extracted, and edit distances are calculated between sequences spanned potential somatic TRI region, followed by clustering to initial screen the presence of somatic TRIs; iii. Somatic TRIs are confirmed using TDscope; iv. Repeat

1659 motif of somatic TRIs are annotated (see Methods for details).

1660 b. Schematic representation of the workflow for identifying somatic TEI. i. Reads from
1661 tumor and blood samples are aligned to the reference genome; ii. Candidate somatic
1662 SVs are identified using Sniffles2.; iii. After polishing the reference genome, reads are
1663 realigned to the refined reference to validate the presence of candidate SVs; iv.
1664 Somatic SVs are further refined using IRIS; v. Somatic SVs are annotated (see
1665 Methods for details).

1666 c. The upper panel shows the frequency (log10 scale) of somatic TRIs, while the lower
1667 panel displays the frequency of somatic TEIs across various cancer types. Each dot
1668 represents an individual sample, with distinct colors indicating different cancer types.
1669 Samples with high burden of TRIs or TEIs are labeled accordingly.

1670

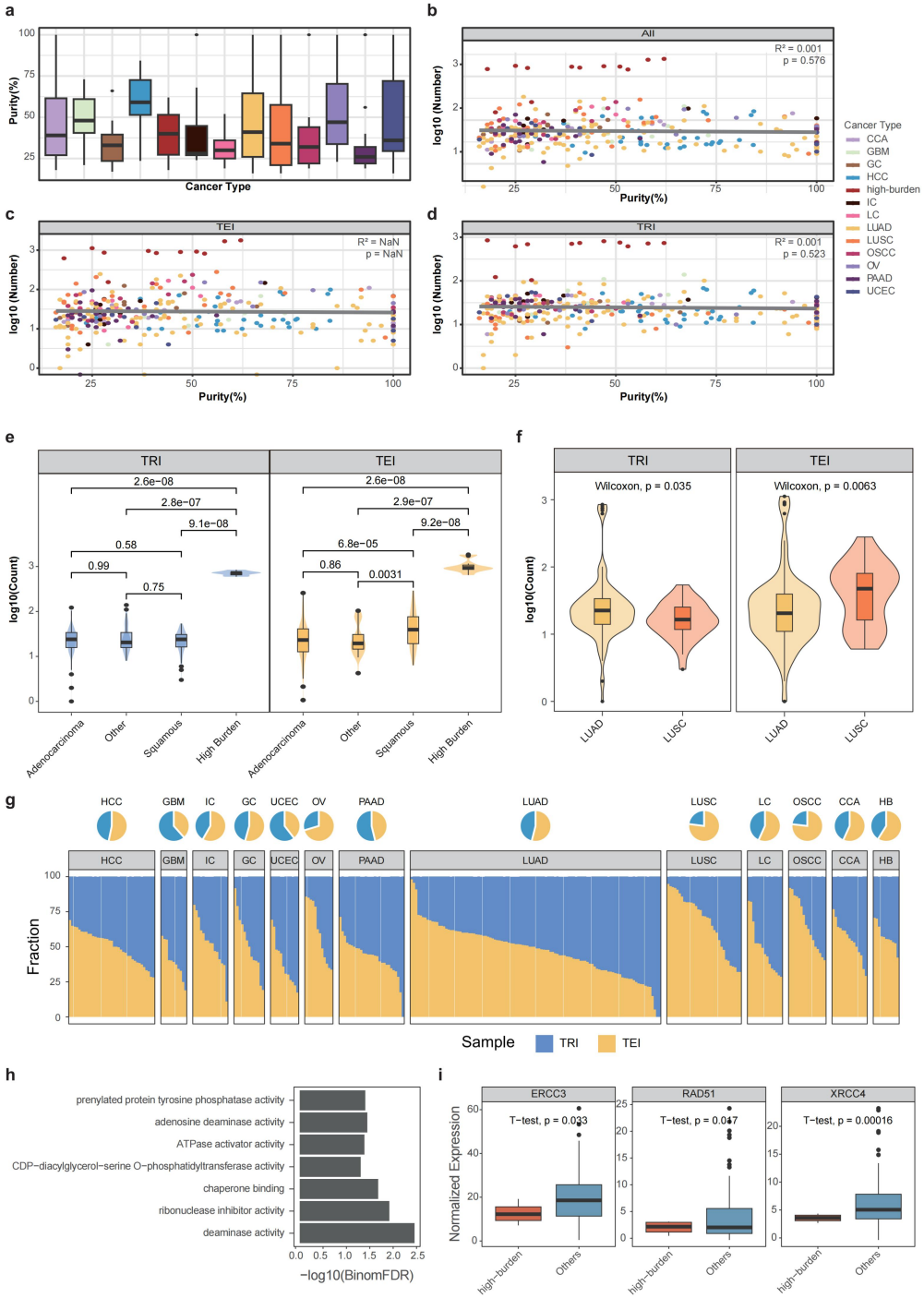



**Figure S3. Tumor Purity and Characteristics of Somatic REIs.**

a. Box plot of tumor purity across different cancer types. Each box represents a specific cancer type, with the y-axis indicating tumor purity percentage.

b-d. Scatter plots showing the correlation between tumor purity and the count of somatic

1677 TRIs (e), somatic TEIs (f), and the combined count of somatic TRIs and TEIs (d). Data
1678 points represent individual samples, colored by cancer type. Pearson correlation
1679 coefficients ($R^2$) and p-values are displayed in the upper-right corner of each plot. Boxplot
1680 showing the log-transformed number of somatic TRIs (blue) and somatic TEIs (yellow)
1681 across different cancer types: "Other" (including HCC, GBM), Squamous (LUSC, CCA,
1682 OSCC, LC), and Adenocarcinoma (LUAD, PAAD, GC, IC, OV, UCEC). High-burden
1683 samples are indicated separately. Statistical significance is assessed using a two-tailed
1684 t-test.
1685 e. Boxplot illustrating the log-transformed number of somatic TRIs (blue) and somatic
1686 TEIs (yellow) in Adenocarcinoma, Squamous, Other and high-burden samples. p-values
1687 were calculated using the Wilcoxon rank-sum test.
1688 f. Boxplot illustrating the log-transformed number of somatic TRIs and somatic TEIs in
1689 LUSC (yellow) and LUAD (orange) samples. p-values were calculated using the Wilcoxon
1690 rank-sum test.
1691 g. A bar chart depicting the proportion of somatic TEIs (yellow) and somatic TRIs (blue)
1692 across cancer types and high-burden samples. Each bar represents a distinct sample. A
1693 pie chart above summarizes the overall percentage of somatic TRIs and TEIs within each
1694 cancer type.
1695 h. Bar plot showing the GO Molecular Function enrichment analysis of genes associated
1696 with high-burden specific TRIs, as predicted by the GREAT. The x-axis represents the
1697 -log10 transformed FDR from the binomial test.
1698 i. Boxplots comparing the expression levels of *ERCC3*, *RAD51*, and *XRCC4* in
1699 high-burden LUAD samples (red) and low-burden LUAD samples (blue). Statistical
1700 significance was assessed using a two-tailed t-test.
1701

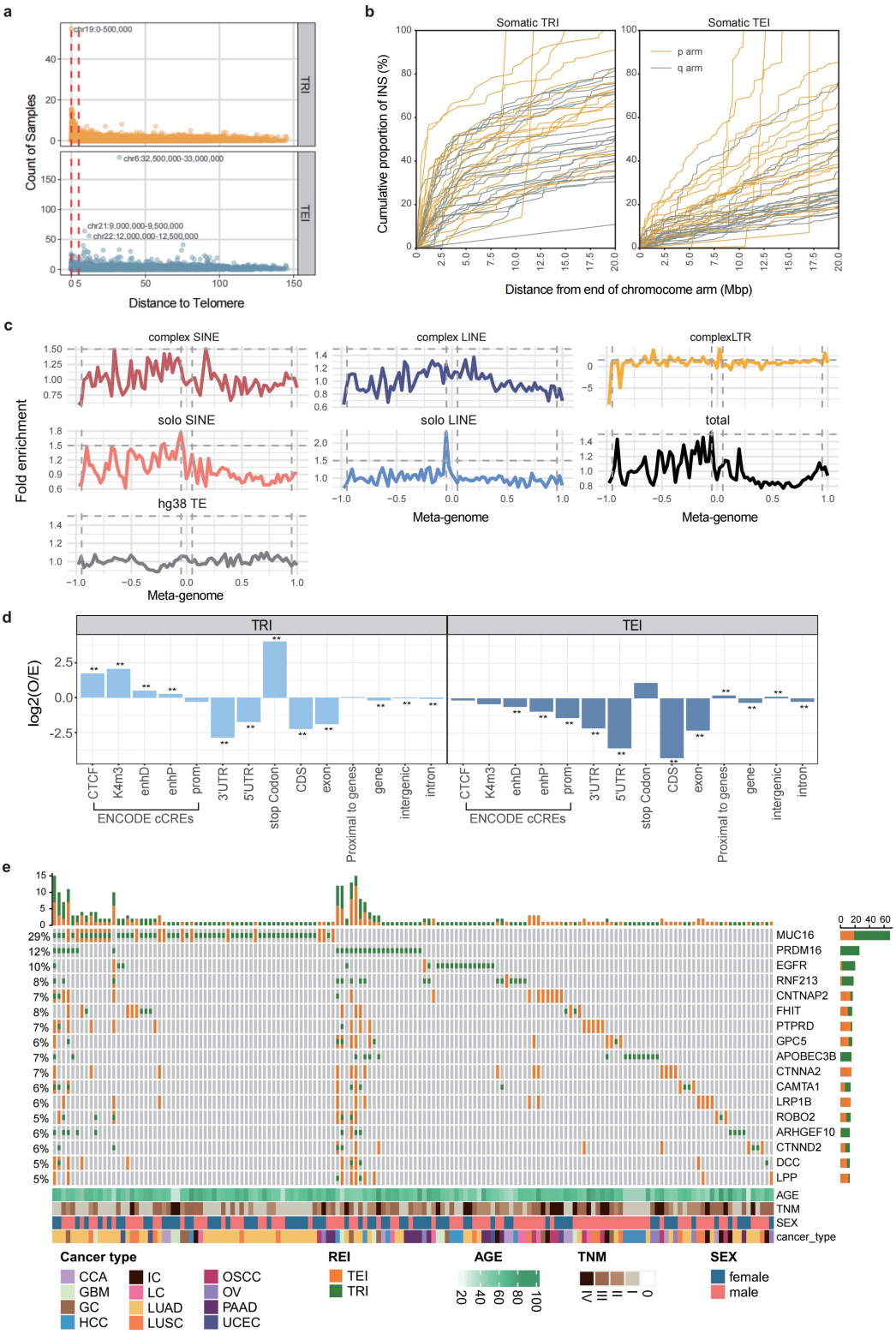



**Figure S4. The genome distribution of somatic TRIs and TEIs.**

a. Distribution of the distance from somatic TRIs (yellow, above) and somatic TEIs (blue, below) to telomeres. The x-axis represents the distance from somatic insertions to

1707 the telomeric ends, while the y-axis indicates the number of samples. Two red dashed
1708 lines mark the positions of 0 bp (telomere ends) and 5 Mbp from the telomere ends.
1709 b. Cumulative distribution of somatic TRI (left) and somatic TEI (right). The plot reveals
1710 a more rapid saturation of somatic TRIs at the ends of chromosome arms. A sliding
1711 window of 500,000bp, starting from the telomeric regions and extending toward the
1712 centromere, was used to cumulatively count somatic TRIs and TEIs. The x-axis is
1713 truncated at 20 Mbp.
1714 c. Chromosome-level distribution of various somatic TEI types ("observed") compared
1715 to TEs in the hg38 reference genome ("expected"). The fold enrichment of observed
1716 TEIs relative to random expectations is calculated refer to the previous study [58]. Each
1717 density plot corresponds to a distinct type of TEI, represented by different colors.
1718 d. Enrichment of somatic TRIs (left) and somatic TEIs (right) on genomic elements. The
1719 ratio of observed insertions to random expectation (O/E) is calculated using a
1720 hypergeometric test. *$p < 0.05$. The y-axis is presented on a log scale for O/E values.
1721 Somatic TRIs and TEIs are shown in different colors to distinguish between the two
1722 insertion types.
1723 e. Oncoplot shows high frequency (Sample frequency > 5% ) COSMIC cancer genes
1724 with somatic REI events located within gene body and 2kbp flanking regions. The bar
1725 plot on the top and right show the number of somatic REI events located within
1726 samples and gene flank regions with color coded in the bottom,

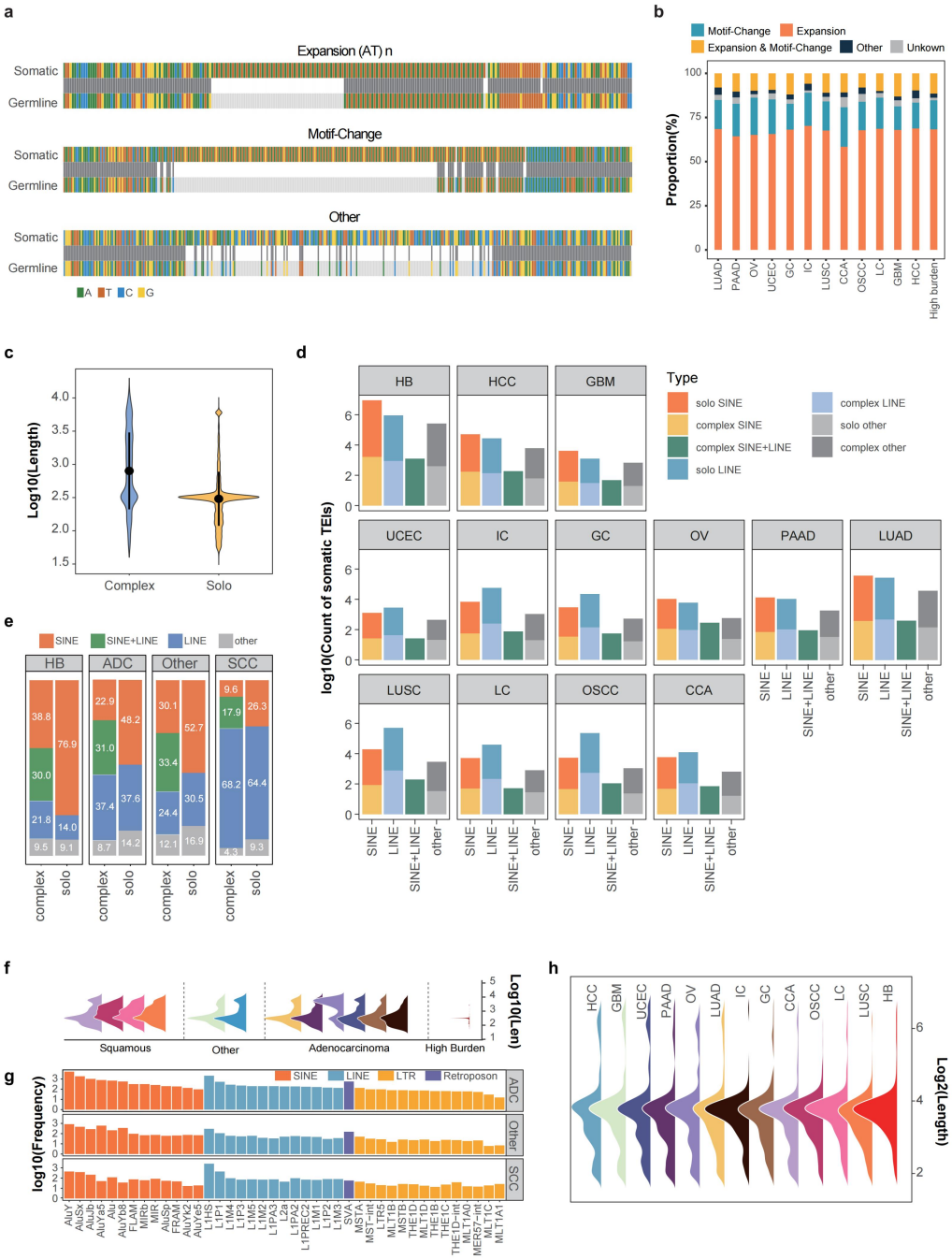



**Figure S5. Sequence charateristic of somatic TRIs and somatic TEIs.**

a. Schematic representation of somatic TRI patterns for Expansion, Motif Change, and Other. The top row represents somatic sequences, the middle row indicates aligned/unaligned, and the bottom row shows germline sequences.

b. Proportions of each somatic TRI category in cancer types: Expansion (orange), Expansion & Motif-Change (yellow), Motif-Change (blue), Ambiguous (dark blue) and

1735 Unknown (gray).

1736 c. Violin plot displaying the log-transformed length distribution for somatic complex (blue)
1737 and solo (yellow) TEI.

1738 d. Proportions of somatic complex and solo TEI across various cancer types. ‘complex
1739 SINE’ indicates complex TEIs containing SINE elements but lacking LINE elements;
1740 ‘complex LINE’ indicates complex TEIs containing LINE elements but lacking SINE
1741 elements; ‘complex SINE+LINE’ denotes TEIs containing both SINE and LINE
1742 elements.

1743 e. Frequency (log10) of SINE (orange), LINE (blue), LTR (yellow) and Retroposon
1744 (purple) subfamilies in adenocarcinoma, squamous cell carcinoma and other cancer
1745 type. Each column represents a subfamily. ’ADC’: adenocarcinoma, ‘SCC’:
1746 squamous cell carcinoma, ‘other’: HCC and GBM.

1747 f. Length density distribution (log10 scale) of somatic TEIs in squamous cell carcinoma,
1748 other cancer types, adenocarcinoma, and high-burden samples. Different colors
1749 represent different cancer types, similar to panel h.

1750 g. Log-transformed counts of different subfamilies of somatic TEIs across cancer types.
1751 "ADC" represents adenocarcinoma, "SCC" stands for squamous cell carcinoma, and
1752 "other" includes HCC and GBM.

1753 h. Density plot of log-transformed length distribution of TSD across different cancer
1754 types.

1755

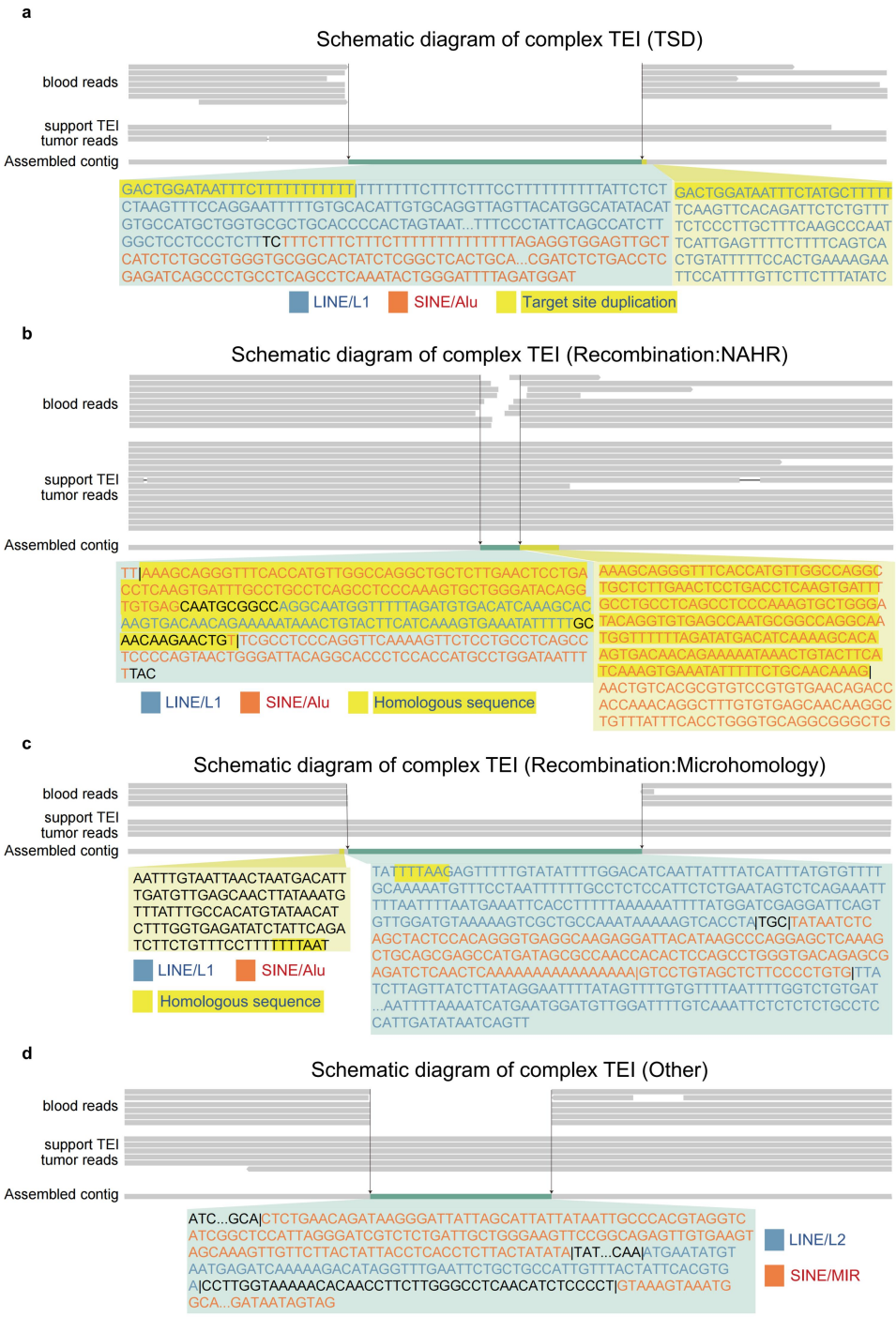



**Figure S6. Examples of somatic TEI insertion mechanisms**

a. The schematic illustrates example of Target Site Duplication. In IGV, the reference genome comprises assembled sequences containing TEI-supporting reads, with all blood reads and tumor TEI-supporting reads aligned to this assembly. Complex TEI sequences are highlighted in light green, and homologous sequences are marked in

1763 yellow-green. LINE elements are colored in blue, while SINE elements are colored in
1764 orange-red. Target site duplications are marked in bright yellow.
1765 b. The schematic illustrates example of Recombination (NAHR). In IGV, the reference
1766 genome comprises assembled sequences containing TEI-supporting reads, with all
1767 blood reads and tumor TEI-supporting reads aligned to this assembly. Complex TEI
1768 sequences are highlighted in light green, and homologous sequences are marked in
1769 yellow-green. LINE elements are colored in blue, while SINE elements are colored in
1770 orange-red. Homologous sequences are marked in bright yellow.
1771 c. The schematic illustrates example of Recombination (Microhomology). In IGV, the
1772 reference genome comprises assembled sequences containing TEI-supporting reads,
1773 with all blood reads and tumor TEI-supporting reads aligned to this assembly.
1774 Complex TEI sequences are highlighted in light green, and homologous sequences
1775 are marked in yellow-green. LINE elements are colored in blue, while SINE elements
1776 are colored in orange-red. Homologous sequences are marked in bright yellow.
1777 d. The schematic illustrates example of Other. In IGV, the reference genome comprises
1778 assembled sequences containing TEI-supporting reads, with all blood reads and
1779 tumor TEI-supporting reads aligned to this assembly. Complex TEI sequences are
1780 highlighted in light green, and homologous sequences are marked in yellow-green.
1781 LINE elements are colored in blue, while SINE elements are colored in orange-red.
1782

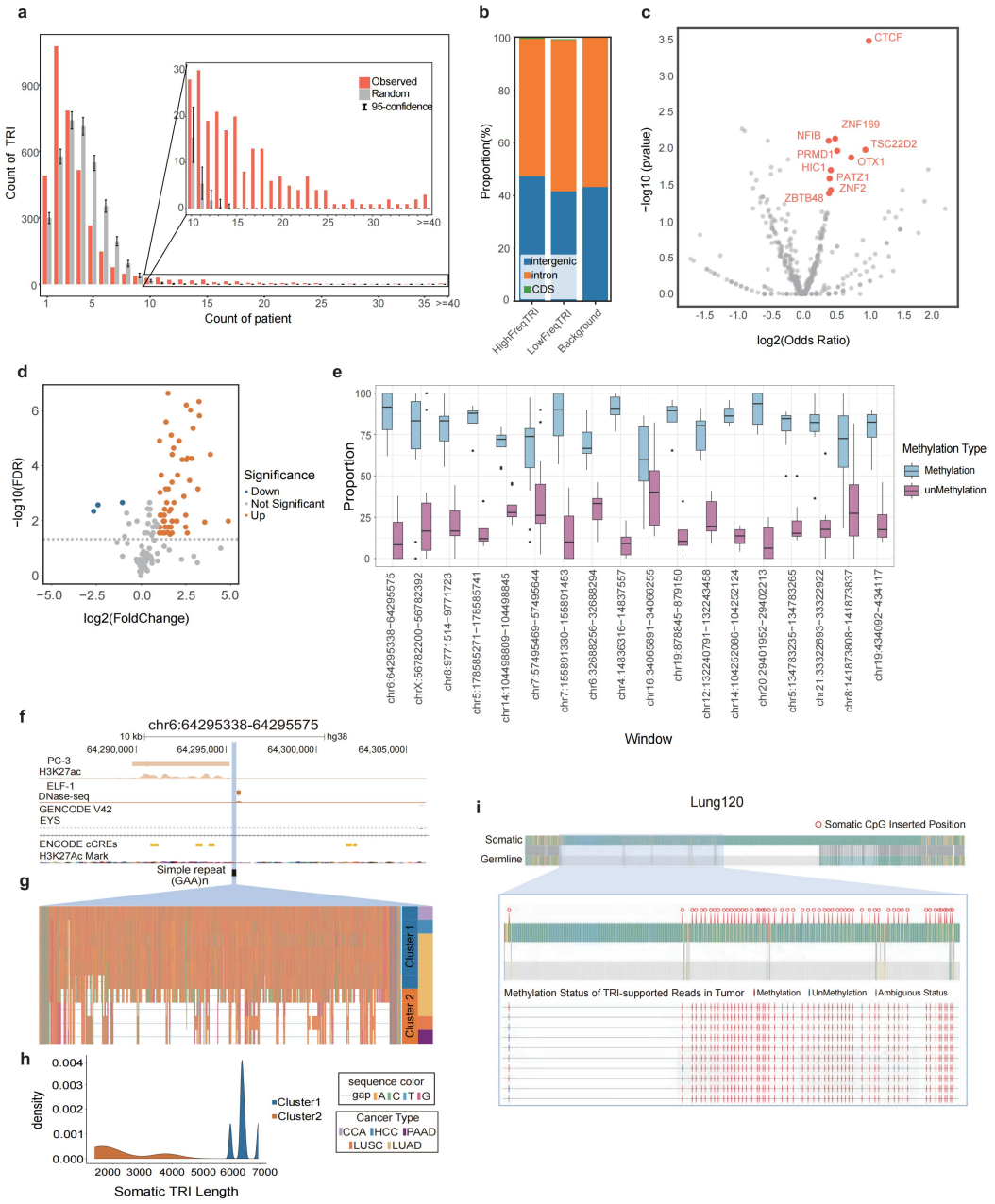



**Figure S7. Overview of TRI distribution**

a. Bar plot depicting the number of patients sharing the same somatic TRI events. Gray bars indicate the random distribution (Method), while orange bars represent the observed distribution. The x-axis represents the number of patients sharing somatic TRIs, and the y-axis represents the count of TRIs.

b. Bar plot illustrating the percentage distribution of low-frequency, high-frequency, and background TRIs located within intergenic, intronic, and coding sequence (CDS)

1791 regions.

1792 c. Volcano plot depicting the differential enrichment between high and low frequency
1793 somatic TRIs on known 490 TF binding regions from ENCODE project. Elements with
1794 p-value < 0.05 (Chi-square test) and |log2(Odds Ratio)| > 1 are highlighted in red.

1795 d. Volcano plot illustrating high-frequency TRIs with significant changes in CpG counts
1796 between somatic and matched germline TRI sequences. Orange: TRIs with FDR <
1797 0.05 and log2(FoldChange) > 1 (Wilcoxon test); Blue:FDR < 0.05 and
1798 log2(FoldChange) < 1 (Wilcoxon test).

1799 e. Box plot displaying the proportions of methylated and unmethylated samples at 18
1800 high-frequency somatic TRIs introduced methylation marks at generated CpG
1801 dinucleotides (Wilcoxon test, p < 0.05). Each dot representing the proportion of
1802 methylated (blue) and unmethylated (pink) states for each sample within each TRI
1803 window.

1804 f. Browser plot displaying the epigenomic signal, ENCODE-annotated gene transcript,
1805 and cCREs (Candidate Cis-Regulatory Elements) predicted by ENCODE around the
1806 somatic TRI located in the intron of the EYS gene.

1807 g. An enlarged panel providing a detailed sequence overview of different clusters
1808 surrounding the somatic TRIs in the EYS intron.

1809 h. Density plot depicting the length distribution of two clusters, color-coded for clarity.

1810 i. Methylation analysis of the TRI sequence from a representative sample: the top panel
1811 displays overall sequences of somatic and corresponding germline with pairwise
1812 alignment details; the middle panel highlights the specific somatic CpG insertion site;
1813 the bottom panel shows all tumor reads supporting the TRI event, along with
1814 methylation status at the somatic CpG insertion site. Methylated CpG sites are
1815 colored in red and unmethylated CpG sites are colored in blue.

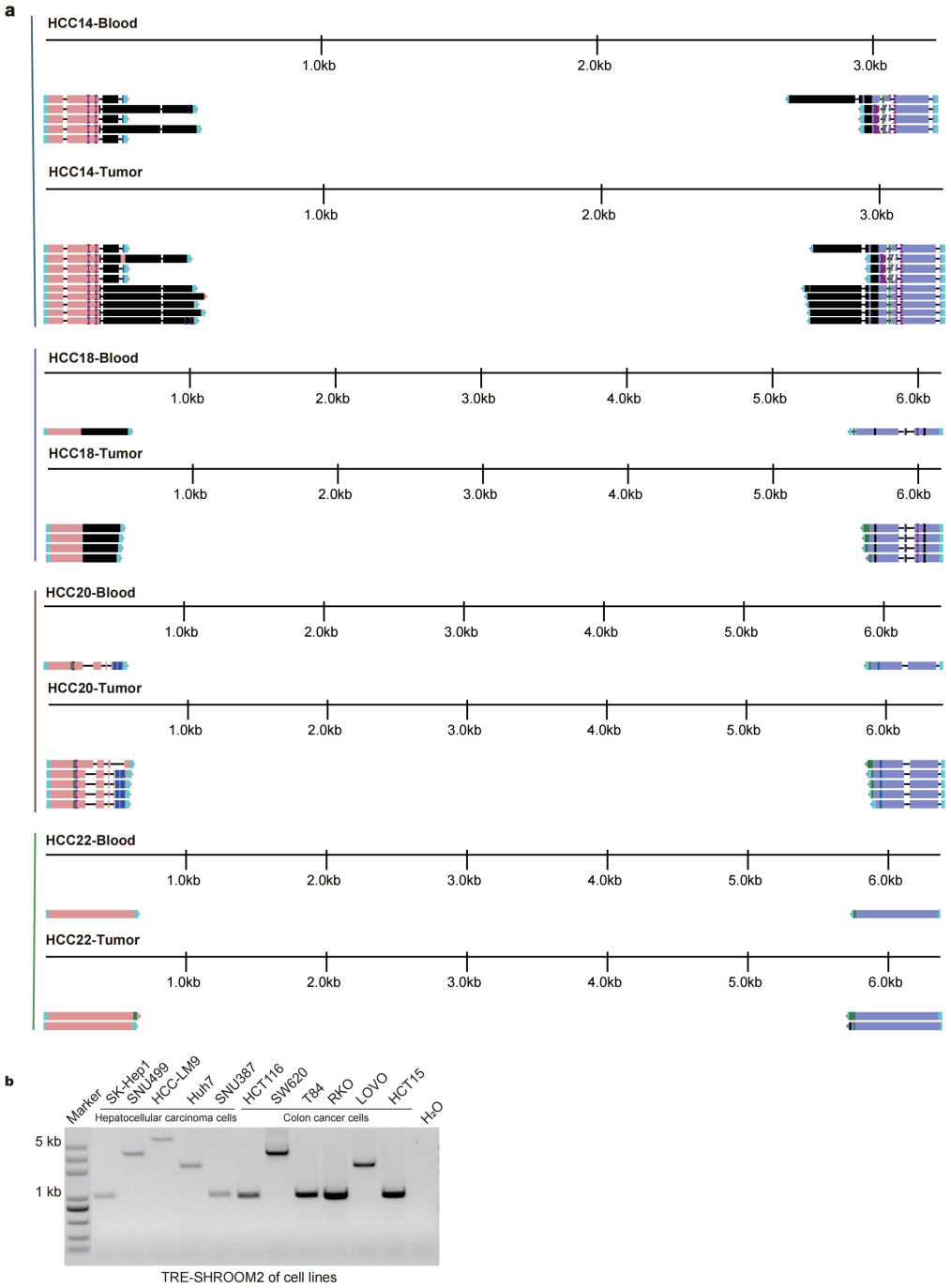



**Figure S8. Sanger sequencing of Shroom2-related SRE in clinical samples and detection in cell lines**

a. Visual display of sanger sequencing of Shroom2-related SRE in clinical samples.

b. Detection of Shroom2-related SRE in liver cancer and intestinal cancer cell lines.

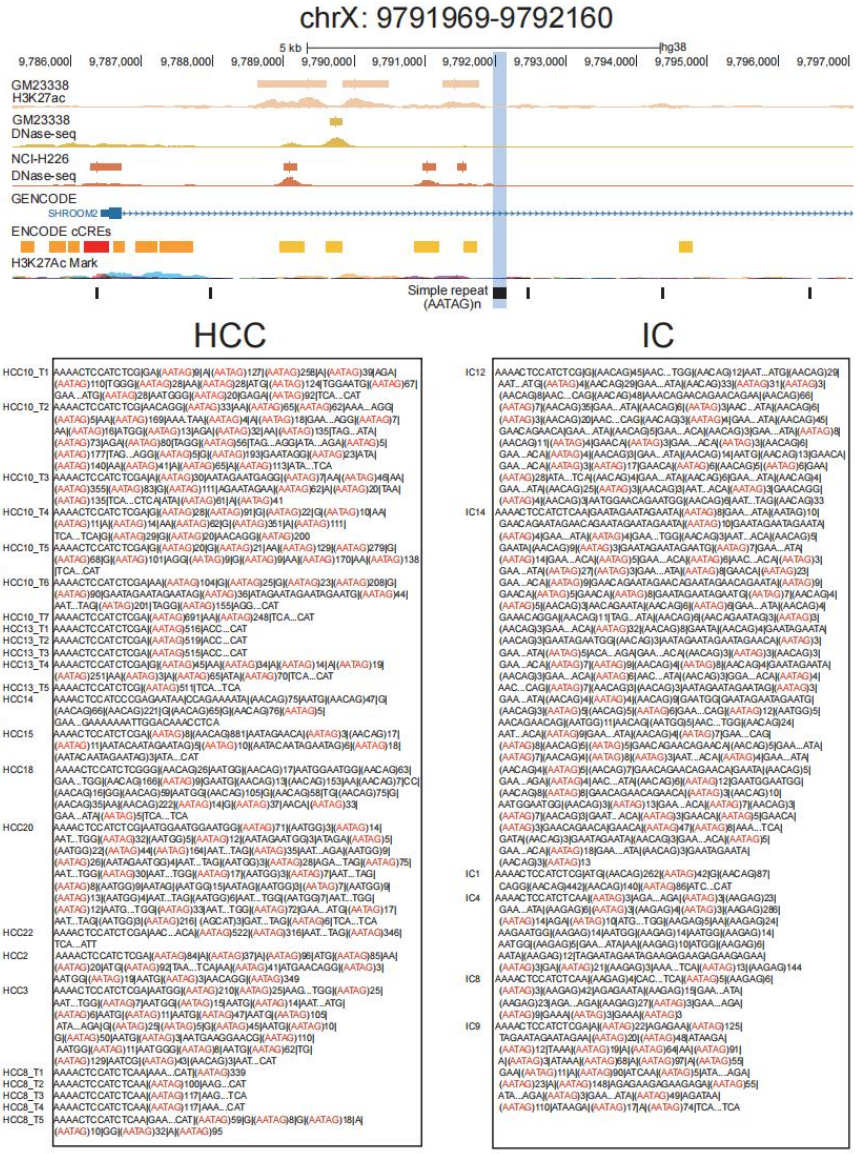



**Figure S9. The browser plot of somatic TRIs located in the intron of *SHROOM2*, along with the somatic TRIs sequence displays for HCC and IC.**

The browser plot illustrates epigenomic signals from ENCODE project, including cell line H3K27ac chip-seq and DNase signal, gene transcript Candidate Cis-Regulatory Elements (cCREs), and H3K27ac marks identified by ENCODE.The somatic TRI (chrX: 9791969-9792160) is highlighted in blue. The bottom panel illustrates detailed somatic sequences of HCC and IC samples harbored the somatic TRI event with the motif AATAG abbreviated and highlighted in red.

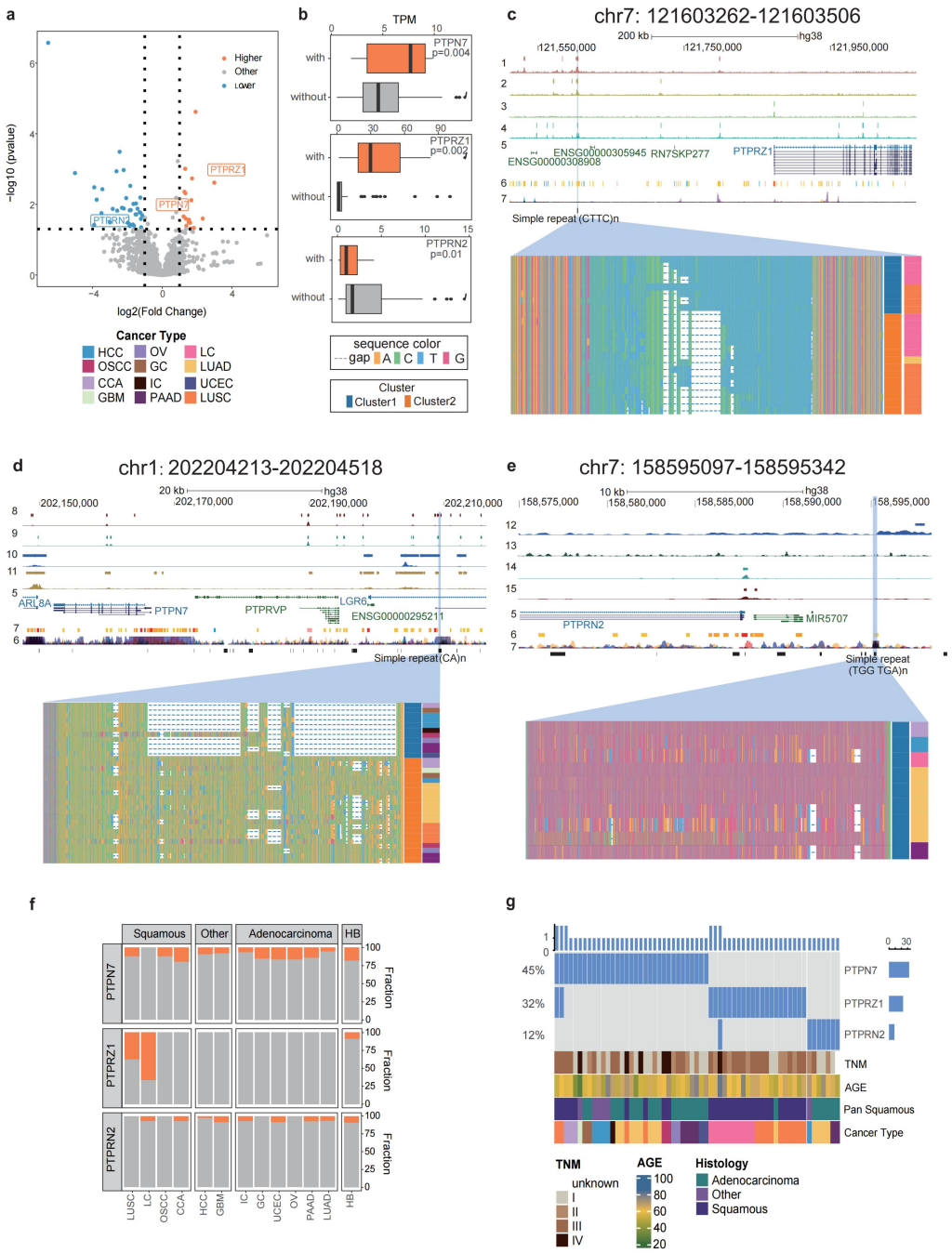



**Figure S10. Three somatic TRIs that associated with the expression of PTP family genes in NSCLC.**

a. Volcano plot showing differentially expressed genes between NSCLC samples with and without high-frequency somatic TRIs. Genes from the PTP family are highlighted with text. Dotted lines parallel to the X-axis indicates where p = 0.05, while two vertical dashed lines represent log2(Fold change) values of -1 and 1, respectively.

1841 b. Box plot of the expression levels of differentially expressed PTP family genes with and
1842 without somatic TRI. Orange: with TRI; Gray: without TRI. P values were calculated by
1843 t-test.
1844 c. The browser plot of somatic TRIs (chr7:121603262-121603506) illustrates the
1845 following epigenomic signals: (1) H3K27ac in the A673 cell line, (2) H3K27ac in the
1846 SK-N-MC cell line, (3) DNase-seq in the A673 cell line, (4) DNase-seq in the SK-N-MC
1847 cell line. Additionally,it displays the ENCODE-annotated gene transcript (5), Candidate
1848 Cis-Regulatory Elements (cCREs) predicted by ENCODE (6), and H3K27ac marks
1849 identified by ENCODE (7). The plot concludes with the CTTC motif of the repeat. An
1850 enlarged panel below outlines a detailed sequence overview of various clusters, with
1851 vertical labels indicating cluster and cancer type information.
1852 d. The browser plot of somatic TRIs (chr1:202204213-202204518) illustrates the
1853 following epigenomic signals: (8) DNase-seq in the H7 cell line, (9) DNase-seq in the
1854 A549 cell line, (10) H3K27ac in the RWPE1 cell line, (11) H3K27ac in the A549 cell
1855 line. Additionally,it displays the ENCODE-annotated gene transcript (5), Candidate
1856 Cis-Regulatory Elements (cCREs) predicted by ENCODE (6), and H3K27ac marks
1857 identified by ENCODE (7). The plot concludes with the CA motif of the repeat. An
1858 enlarged panel below outlines a detailed sequence overview of various clusters, with
1859 vertical labels indicating cluster and cancer type information.
1860 e. The browser plot of somatic TRIs (chr1:202204213-202204518) illustrates the
1861 following epigenomic signals: (12) H3K27ac in the RWPE1 cell line, (13) H3K27ac in
1862 the H9 cell line, (14) DNase-seq in the RWPE1 cell line, (15) DNase-seq in the H9 cell
1863 line. Additionally,it displays the ENCODE-annotated gene transcript (5), Candidate
1864 Cis-Regulatory Elements (cCREs) predicted by ENCODE (6), and H3K27ac marks
1865 identified by ENCODE (7). The plot concludes with the TGGTGA motif of the repeat.
1866 An enlarged panel below outlines a detailed sequence overview of various clusters,
1867 with vertical labels indicating cluster and cancer type information.
1868 f. Frequency of somatic TRI occurrence around PTP genes across cancer types.
1869 Orange: with TRI; Gray: without TRI.
1870 g. Oncoplot of samples associated with the PTP Family somatic TRI windows. The bar

charts on the top and right show the counts, while the percentages on the left indicate the proportion among all samples.

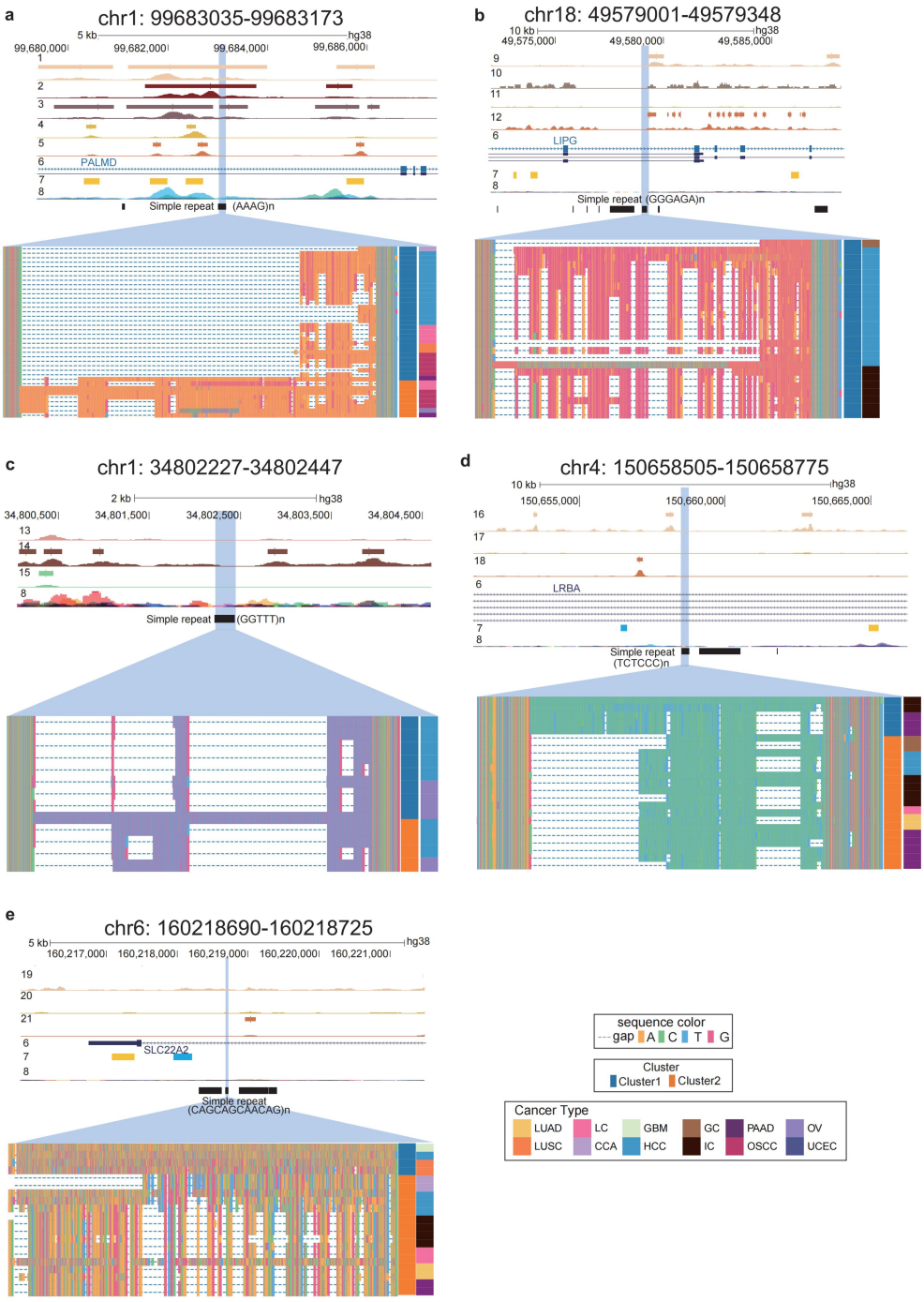



**Figure S11. Display of representative somatic TRI events**

a. The browser plot of somatic TRIs (chr1:99683035-99683173) with significantly high frequency in HCC, OSCC and LC illustrates the following epigenomic signals: (1) H3K27ac in the HeLa-S3 cell line, (2) H3K27ac in the SJSA1 cell line, (3) H3K27ac in the ACC112 cell line, (4) DNase-seq in the HeLa-S3 cell line, and (5) DNase-seq in

1881 the SJSA1 cell line. Additionally, it displays the ENCODE-annotated gene transcript
1882 (6), Candidate Cis-Regulatory Elements (cCREs) predicted by ENCODE (7), and
1883 H3K27ac marks identified by ENCODE (8). The plot concludes with the AAAG motif of
1884 the repeat. An enlarged panel below outlines a detailed sequence overview of various
1885 clusters, with vertical labels indicating cluster and cancer type information.

1886 b. The browser plot of somatic TRIs (chr18:49579001-49579348) with significantly high
1887 frequency in HCC and IC illustrates the following epigenomic signals: (9) H3K27ac in
1888 the Caco-2 cell line, (10) H3K27ac in the Karpas-422 cell line, (11) DNase-seq in the
1889 Caco-2 cell line, (12) DNase-seq in the Karpas-422 cell line, it displays the
1890 ENCODE-annotated gene transcript (6), Candidate Cis-Regulatory Elements (cCREs)
1891 predicted by ENCODE (7), and H3K27ac marks identified by ENCODE (8). The plot
1892 concludes with the GGGAGA motif of the repeat. An enlarged panel below outlines a
1893 detailed sequence overview of various clusters, with vertical labels indicating cluster
1894 and cancer type information.

1895 c. The browser plot of somatic TRIs (chr1:34802227-34802447) with significantly high
1896 frequency in HCC and OV illustrates the following epigenomic signals: (13) H3K27ac
1897 in the HFFc6 cell line, (14) DNase-seq in the K562 cell line, (15) DNase-seq in the
1898 HFFc6 cell line, and H3K27ac marks identified by ENCODE (8). The plot concludes
1899 with the GGTTT motif of the repeat. An enlarged panel below outlines a detailed
1900 sequence overview of various clusters, with vertical labels indicating cluster and
1901 cancer type information.

1902 d. The browser plot of somatic TRIs (chr4:150658505-150658775) with significantly high
1903 frequency in IC and PAAD illustrates the following epigenomic signals: (16) H3K27ac
1904 in the H9 cell line, (17) DNase-seq in the H9 cell line, (18) DNase-seq in the KBM-7
1905 cell line, it displays the ENCODE-annotated gene transcript (6), Candidate
1906 Cis-Regulatory Elements (cCREs) predicted by ENCODE (7), and H3K27ac marks
1907 identified by ENCODE (8). The plot concludes with the GGTTT motif of the repeat. An
1908 enlarged panel below outlines a detailed sequence overview of various clusters, with
1909 vertical labels indicating cluster and cancer type information.

1910 e. The browser plot of somatic TRIs (chr6:160218690-160218725) with significantly high

1911 frequency in IC and PAAD illustrates the following epigenomic signals: (19) H3K27ac
1912 in the WTC11 cell line, (20) DNase-seq in the WTC11 cell line, (21) DNase-seq in the
1913 RCC cell line. Additionally, it displays the ENCODE-annotated gene transcript (6),
1914 Candidate Cis-Regulatory Elements (cCREs) predicted by ENCODE (7), and
1915 H3K27ac marks identified by ENCODE (8). The plot concludes with the
1916 CAGCAGCAACAG motif of the repeat. An enlarged panel below outlines a detailed
1917 sequence overview of various clusters, with vertical labels indicating cluster and
1918 cancer type information.
1919

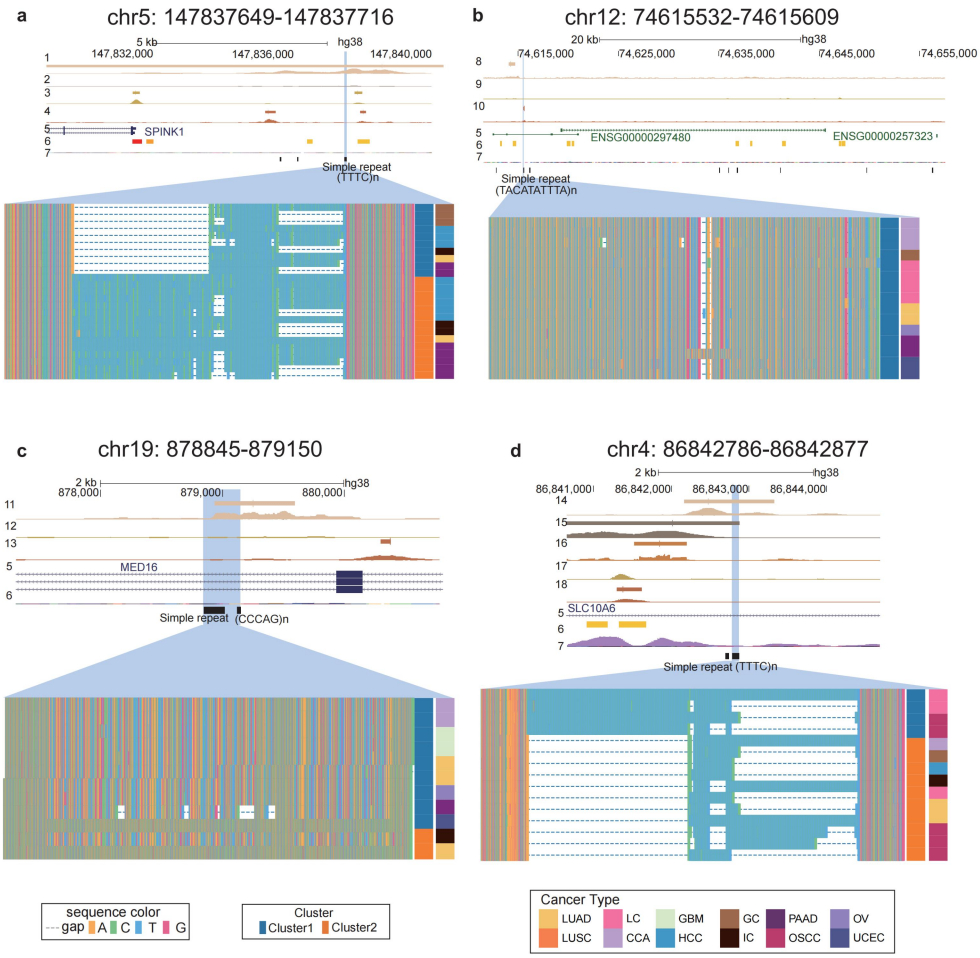



**Figure S12. Display of representative somatic TRI events**

a. The browser plot of somatic TRIs (chr5:147837649-147837716) with significantly high frequency in IC and PAAD illustrates the following epigenomic signals: (1) H3K27ac in the 22Rv1 cell line, (2) H3K27ac in the A549 cell line, (3) DNase-seq in the Caco-2 cell line, (4) DNase-seq in the A549 cell line. Additionally, it displays the ENCODE-annotated gene transcript (5), Candidate Cis-Regulatory Elements (cCREs) predicted by ENCODE (6), and H3K27ac marks identified by ENCODE (7). The plot concludes with the TTTC motif of the repeat. An enlarged panel below outlines a detailed sequence overview of various clusters, with vertical labels indicating cluster and cancer type information.

b. The browser plot of somatic TRIs (chr12:74615532-74615609) illustrates the following

1933 epigenomic signals: (8) H3K27ac in the MCF-7 cell line, (9) DNase-seq in the MCF-7
1934 cell line, (10) DNase-seq in the RCC 7860 cell line. Additionally, it displays the
1935 ENCODE-annotated gene transcript (5), Candidate Cis-Regulatory Elements (cCREs)
1936 predicted by ENCODE (6), and H3K27ac marks identified by ENCODE (7). The plot
1937 concludes with the TACATATTTA motif of the repeat. An enlarged panel below outlines
1938 a detailed sequence overview of various clusters, with vertical labels indicating cluster
1939 and cancer type information.

1940 c. The browser plot of somatic TRIs (chr19:878845-879150) illustrates the following
1941 epigenomic signals: (11) H3K27ac in the Caco-2 cell line, (12) DNase-seq in the
1942 Caco-2 cell line, (13) DNase-seq in the L1-S8R cell line, Additionally, it displays the
1943 ENCODE-annotated gene transcript (5), Candidate Cis-Regulatory Elements (cCREs)
1944 predicted by ENCODE (6). The plot concludes with the CCCAG motif of the repeat. An
1945 enlarged panel below outlines a detailed sequence overview of various clusters, with
1946 vertical labels indicating cluster and cancer type information.

1947 d. The browser plot of somatic TRIs (chr4:86842786-86842877) illustrates the following
1948 epigenomic signals: (14) H3K27ac in the RWPE1 cell line, (15) H3K27ac in the
1949 RWPE2 cell line, (16) H3K27ac in the MCF 10A cell line, (17) DNase-seq in the
1950 RWPE2 cell line, and (18) DNase-seq in the MCF 10A cell line. Additionally, it displays
1951 the ENCODE-annotated gene transcript (5), Candidate Cis-Regulatory Elements
1952 (cCREs) predicted by ENCODE (6), and H3K27ac marks identified by ENCODE (7).
1953 The plot concludes with the TTTC motif of the repeat. An enlarged panel below
1954 outlines a detailed sequence overview of various clusters, with vertical labels
1955 indicating cluster and cancer type information.

1956

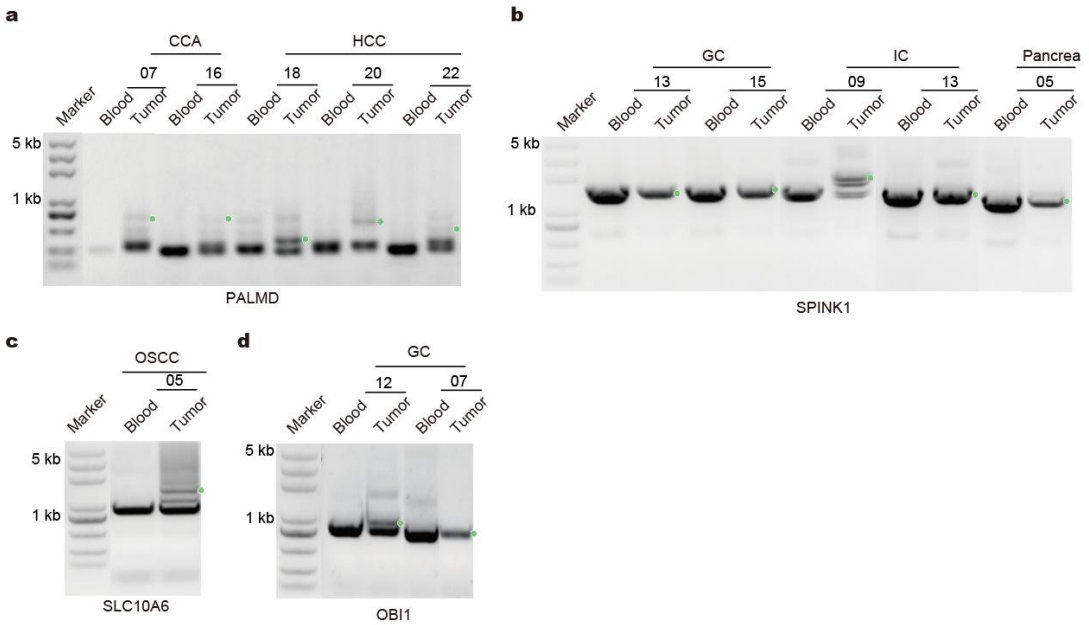



**Figure S13. Validation of clinical samples for TRI events.**

a. Detecting somatic TRI events harbored by *PALMD* through PCR in clinical samples, blood DNA as a control.

b. Detecting somatic TRI events harbored by *SPINK1* through PCR in clinical samples, blood DNA as a control.

c. Detecting somatic TRI events harbored by *SLC10A6* through PCR in clinical samples, blood DNA as a control.

d. Detecting somatic TRI events harbored by *OBI1* through PCR in clinical samples, blood DNA as a control.

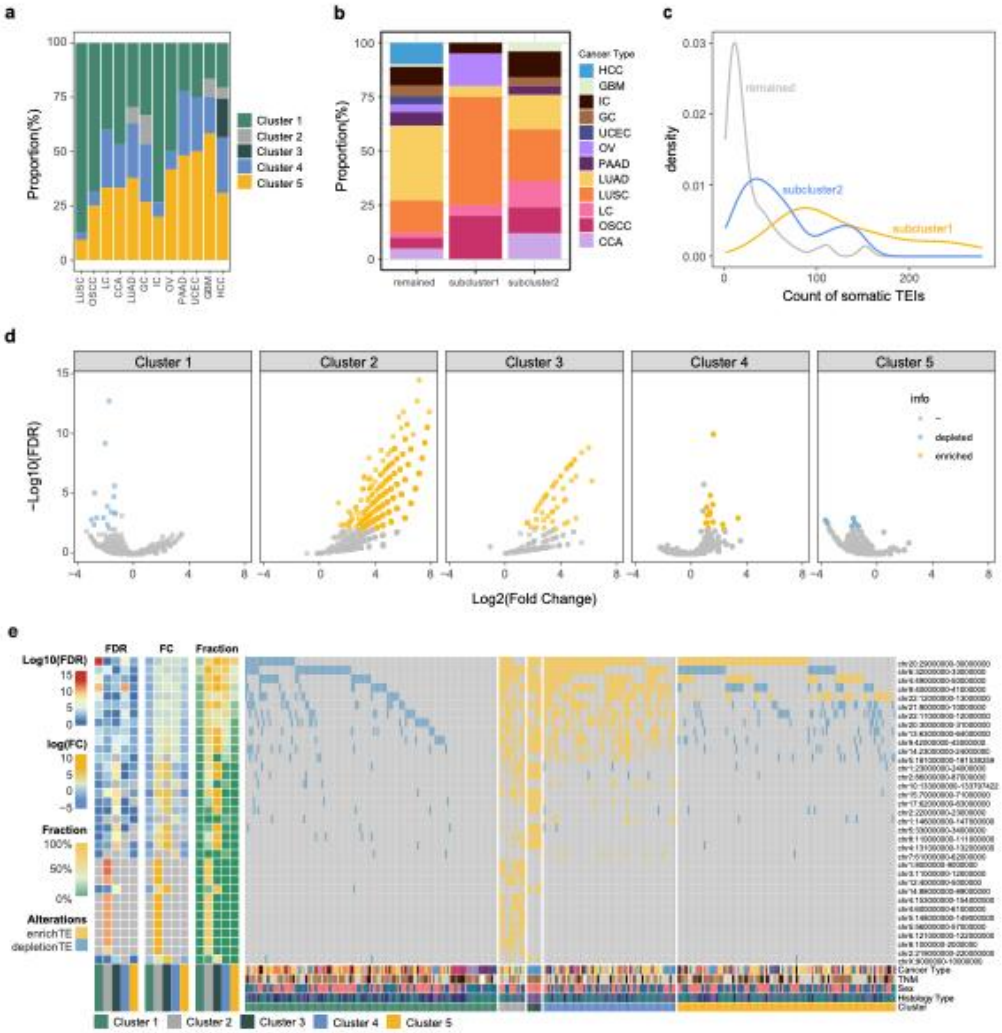



**Figure S14. The characteristic of somatic TEI clusters.**

a. Distribution of the 5 clusters of somatic TEI patterns in 12 cancer types.

b. Distribution of the 12 cancer types in subclusters of Cluster 1. 'other' presents samples out of subcluster1 and subcluster2.

c. Density plot showing the distribution of somatic transposable element insertions (TEIs) in Cluster 1, separated into subcluster1 (yellow), subcluster2 (blue), and the remaining Cluster 1 samples (gray). The x-axis represents the number of somatic TEIs.

d. Volcano plot of significant enriched/depleted somatic TEI windows (1Mbp) between samples in a cluster and samples not in the cluster (Fisher's exact test, p value < 0.05).

e. Oncoplot of top 10 enriched/depleted somatic TEI windows of each cluster (FDR <

0.05). The color of the left heatmap represents the significance level (Fisher's exact test, log10(FDR)) of enrichment or depletion of a specific cluster in a specific 1Mb window. The color of the middle heatmap represents the log2(Fold Change) of enrichment or depletion of a specific cluster in a specific 1Mb window. The color of the middle heatmap represents the log2(Fold Change) of enrichment or depletion of a specific cluster in a specific 1Mb window. The color of the right heatmap represents the fraction of samples harbored somatic TEIs in the specific 1Mb window.

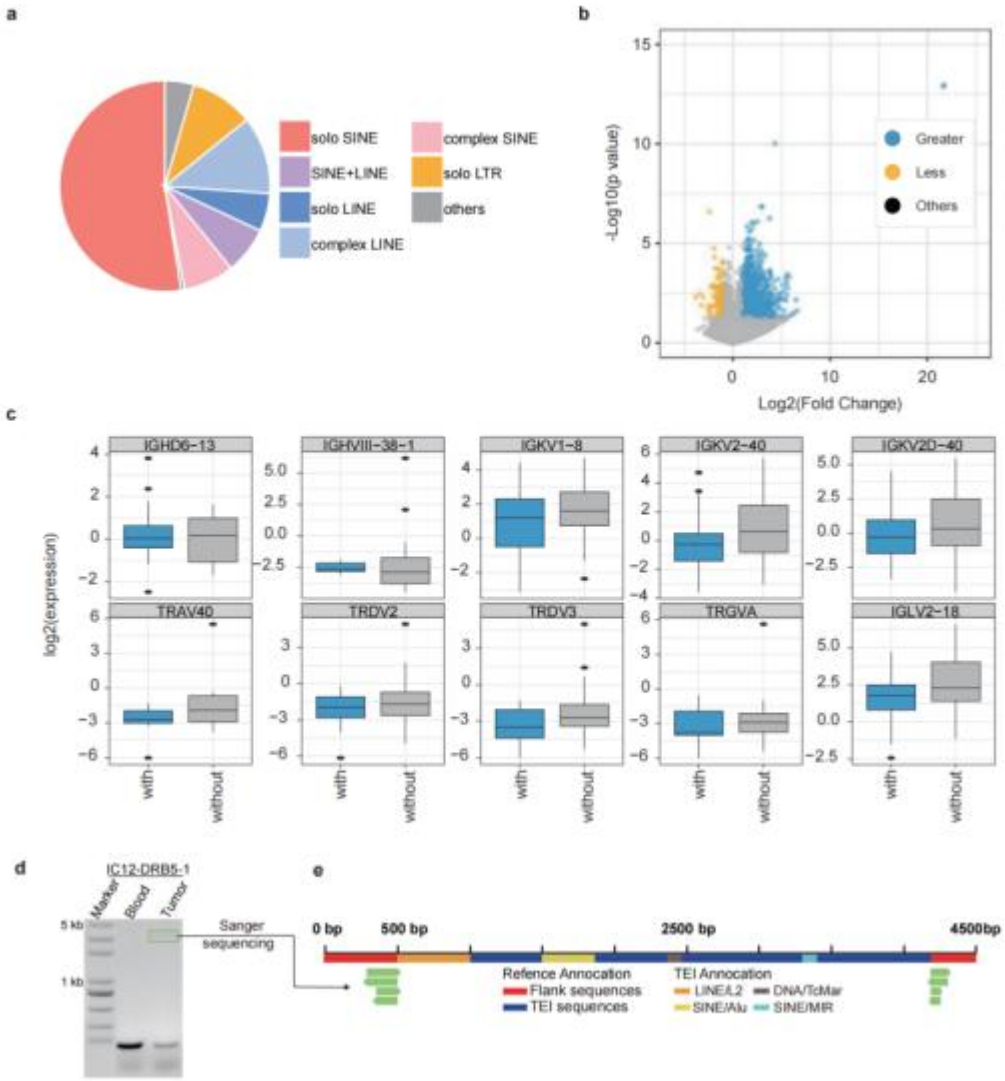



**Figure S15. The proportion of different types of TEIs inserted in MHC II gene family, and differentially expressed genes related to these somatic TEIs.**

a. Pie chart of somatic TEI type distribution. 'solo SINE': TEIs contained solo SINE segments. 'complex SINE': TEIs contained multi SINE segments and not contained other segments. 'solo LINE': TEIs contained solo LINE segments. 'complex LINE': TEIs contained multi LINE segments and not contained other segments. 'SINE+LINE': TEIs both contained SINE segments and LINE segments. 'solo LTR': TEIs consist of solo LTR segments. 'other': the rest type of TEI.

1999 b. Volcano plot showing the differential gene expression between LUAD with somatic
2000 TEIs and LUAD without somatic TEIs.
2001 c. Box plot of gene expression (log2) in samples with (blue) and without (gray) somatic
2002 TEIs.
2003 d. Detection of somatic TEI events via PCR: PCR analysis was performed in IC12, with
2004 blood DNA used as a control to distinguish somatic TEI events.
2005 e. Sanger sequencing visualization of TEI: The reference genome sequence
2006 incorporates both the flanking region (highlighted in red) and the TEI sequence
2007 (highlighted in blue). The TEI annotation is displayed above the reference genome for
2008 precise localization and identification.
2009

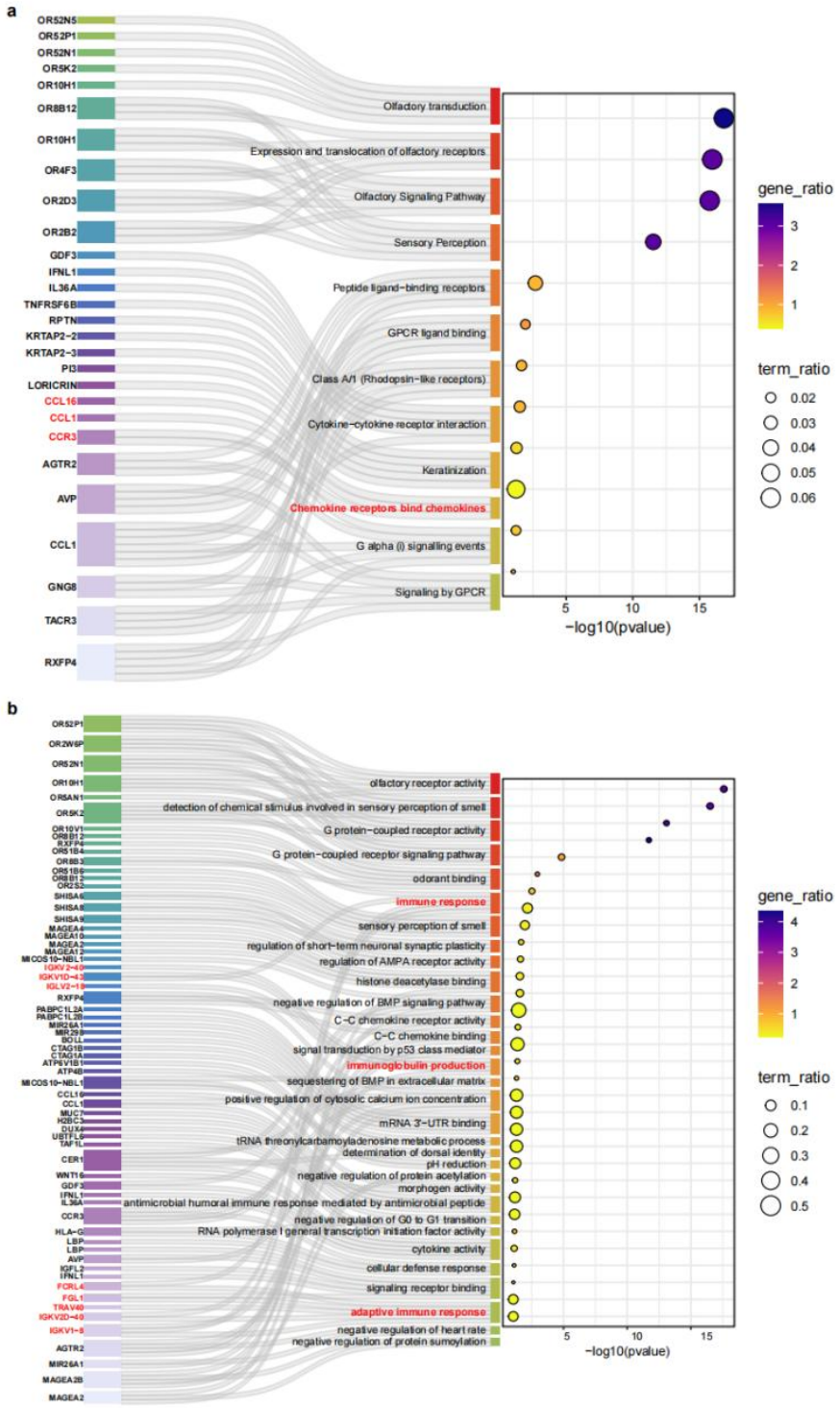



**Figure S16. Bubble plot of pathway and GO enrichment.**

a. Sankey plots of KEGG pathway and corresponding genes are shown on the left, and bubble plots of enrichment results are shown on the right. The x-axis represents the -log10(p-value), y-axis lists the pathways, with the color gradient from purple to yellow

representing the gene ratio. The size of the dots corresponds to the term ratio, with larger dots representing a greater magnitude of change. The connecting lines represent relationships or pathways between the genes.

b. Sankey plots of GO biological process and corresponding genes are shown on the left, and bubble plots of enrichment results are shown on the right. The x-axis represents the -log10(p-value), y-axis lists the pathways, with the color gradient from purple to yellow representing the gene ratio. The size of the dots corresponds to the term ratio, with larger dots representing a greater magnitude of change. The connecting lines represent relationships or pathways between the genes.

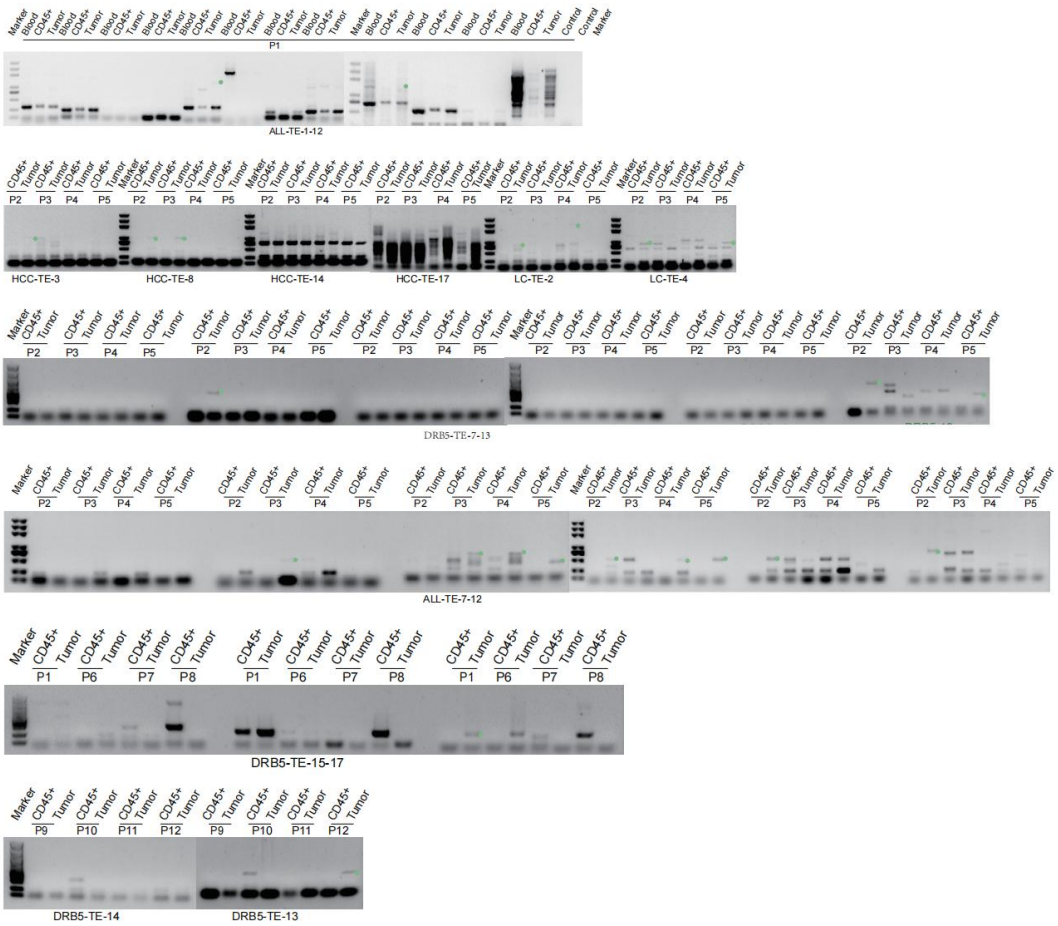
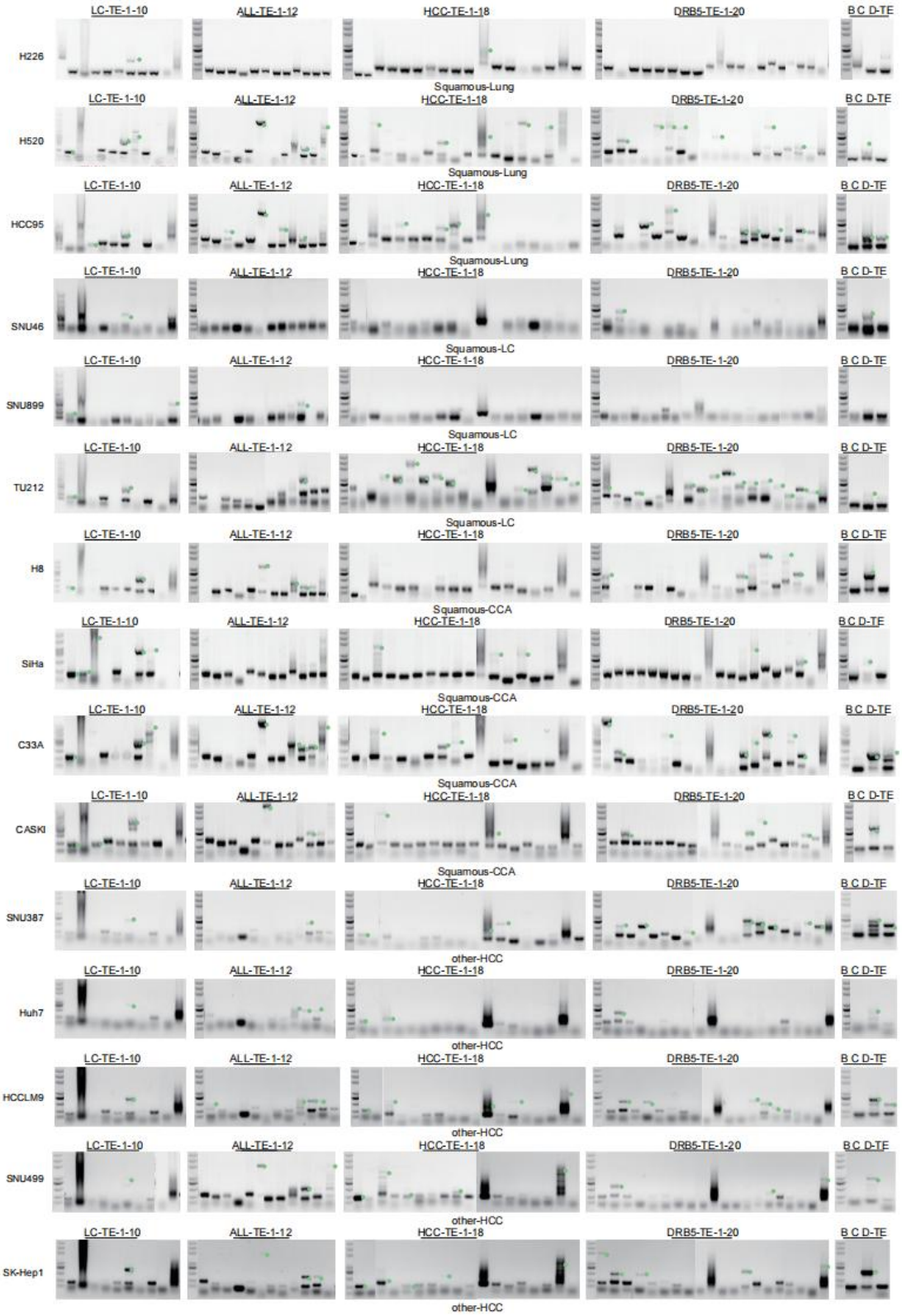



**Figure S17. DetectingTEI events through PCR in additional fresh clinical samples after CD45+ cell sorting.**

The green dots mark the TEI events that were only amplified in the tumor.




2031

2032

2033 **Figure S18. Detecting TEI events through PCR in tumor cell lines.**

2034 The green dots mark the potential TEI events.

**Figure S19. Detecting TEI events through PCR in tumor cell lines.**

The green dots mark the potential TEI events.

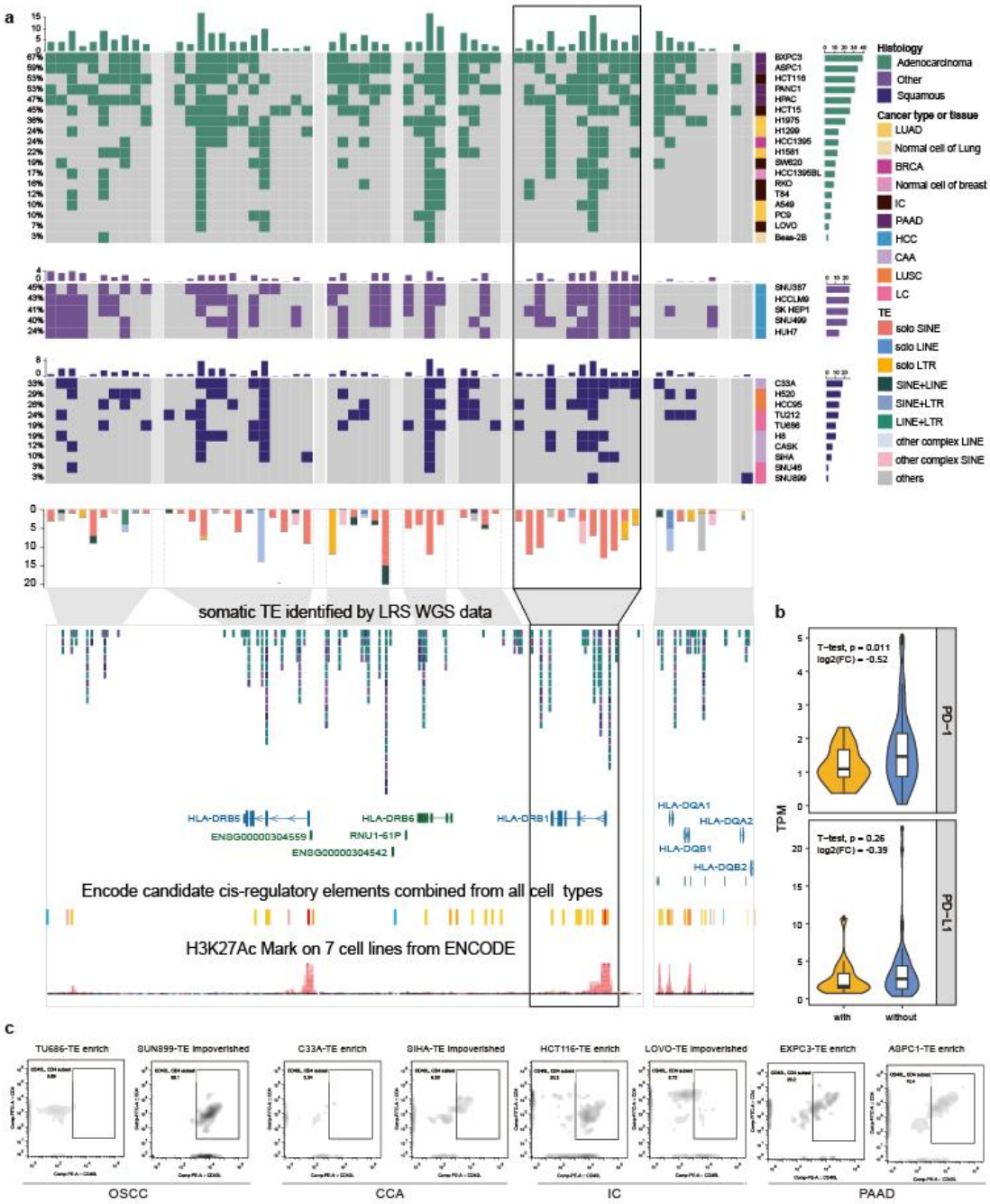



**Figure S20. Statistics of TEI events in tumor cell lines and Flow sytometric analysis.**

a. Statistics of TEI events in tumor cell lines. Each column represents a TEI site, and each row represents a cell line. The first three tracks correspond to cell lines from adenocarcinoma, other types, and squamous, respectively. The fourth track summarizes TEI events from 33 cell lines in the region chr6:32,000,000-33,000,000, with different colors indicating the categories of these TEI events. The lower enlarged panel displays a browser plot showing ENCODE-annotated gene transcripts, cCREs

2048 predicted by ENCODE, and epigenomic signals around the somatic TEIs within the
2049 MHC II gene family.
2050 b. Violin plot showing the expression level of PD-1 and PD-L1 in LUAD samples
2051 harboring somatic TEIs within the HLA-DRB1 genebody and its ±5kb flanking regions
2052 ('with'), compare to LUAD samples not harboring somatic TEIs within these regions
2053 ('without').
2054 c. Flow cytometric analyses of CD40L mobilization in CD4+ T cells after 24 hours of in
2055 vitro co-culture with tumor cells.
2056
2057