

RetroTest unravels LINE-1 retrotransposition in Head and Neck Squamous Cell Carcinoma

Jenifer Brea-Iglesias^{1,2,3#}, Ana Oitabén^{1,2,3#}, Sonia Zumalave², Bernardo Rodriguez-Martin^{2,4}, María Gallardo-Gómez¹, Martín Santamarina², Ana Pequeño-Valtierra², Laura Juaneda-Magdalena¹, Ramón García-Escudero⁵, José Luis López-Cedrún⁶, Máximo Fraga⁷, José MC Tubio², Mónica Martínez-Fernández^{*1,2}

#Equal contribution

1. Translational Oncology Research Group; Galicia Sur Health Research Institute (IIS Galicia Sur). SERGAS-UVIGO. Estrada de Clara Campoamor, 341, 36213 Vigo, (Spain).

2. Mobile Genomes Lab. Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS). (Universidad de Santiago de Compostela). Avda. Barcelona 31, 15706 Santiago de Compostela (Spain).

3. Equal contribution.

4. Present address: Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona (Spain). Universitat Pompeu Fabra (UPF), Barcelona (Spain).

5. Molecular and Translational Oncology Division, CIEMAT (ed 70A), Ave Complutense 40, 28040 Madrid (Spain). Research Institute Hospital 12 de Octubre (imas12), 28041 Madrid (Spain). Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid (Spain).

6. Department of Maxillofacial Surgery, University Hospital of A Coruña, As Xubias, 84, 15006 A Coruña (Spain).

26 7. Pathological Anatomy. Faculty of Medicine. University Clinical Hospital & Health
 27 Research Institute of Santiago de Compostela (IDIS). Travesa da Choupana s/n,
 28 15706 Santiago de Compostela (Spain).

29

30 ***Corresponding author:** Monica Martínez Fernández.
 31 monica.martinez@iisgaliciasur.es Translational Oncology. Galicia Sur Health
 32 Research Institute (IIS Galicia Sur). Hospital Álvaro Cunqueiro. Estrada de Clara
 33 Campoamor, 341, 36213 Vigo, (Spain).

34

35

36 **Highlights**

- 37
- 38 • RetroTest represents the first method to determine LINE-1
- 39 retrotransposition from tumor biopsies in real clinical settings.
- 40 • RetroTest not only offers global LINE-1 retrotransposition ratios but also
- 41 identifies the active LINE-1 source elements.
- 42 • RetroTest elucidates a really early LINE-1 activation in early tumor stages
- 43 of Head and Neck Squamous Cell Carcinoma.
- 44 • Whole Genome Analysis and LINE-1 retrotransposition demonstrates
- 45 processes of field cancerization in Head and Neck Squamous Cell
- 46 Carcinoma.
- 47 • LINE-1 retrotransposition favors an earlier and more efficient Head and
- 48 Neck Squamous Cell Carcinoma diagnosis.
- 49

50 **Abstract**

51 The relevant role of LINE-1 (L1) retrotransposition in cancer has been recurrently

52 demonstrated in recent years. However, their repetitive nature hampers their

53 identification and detection, hence remaining inaccessible for clinical practice.

54 Also, its clinical relevance for cancer patients is still limited. Here, we develop a

55 new method to quantify L1 activation, called RetroTest, based on targeted

56 sequencing and a sophisticated bioinformatic pipeline, allowing its application in

57 tumor biopsies. First, we performed the benchmarking of the method and

58 confirmed its high specificity and reliability. Then, we unravel the L1 activation in

59 HNSCC according to a more extensive cohort including all the HNSCC tumor stages.

60 Our results confirm that RetroTest is remarkably efficient for L1 detection in

61 tumor biopsies, reaching a high sensitivity and specificity. In addition, L1

62 retrotransposition estimation reveals a surprisingly early activation in HNSCC

63 progression, contrary to its classical association with advanced tumor stages. This

64 early activation together with the genomic mutational profiling of normal adjacent

65 tissues supports field cancerization process in this tumor. These results underline
66 the importance of estimating L1 retrotransposition in clinical practice towards an
67 earlier and more efficient diagnosis in HNSCC.

68

69 **Keywords**

70 Retrotransposition, LINE-1, HNSCC, early diagnosis biomarker, field cancerization

71 **1. Introduction**

72 Approximately half of human genome is composed of transposable elements (TEs),
73 sequences with the ability of moving from one place to another, changing the
74 normal structure of the genome in the places where they are integrated [1,2].
75 Among them, long interspersed nuclear element retrotransposons (LINE-1, L1)
76 represent 17% of the entire DNA content with approximately 500,000 copies, most
77 of them truncated or inactive [3–6]. Specifically, only a small subset of these L1s is
78 active in the human genome, although they stay transcriptionally repressed due to
79 epigenetic mechanisms that prevent the damage that their mobilization would
80 cause [7]. When this repression is lost, L1 activation can cause different diseases,
81 including cancer [8, 9].

82 In the framework of the International Consortium of PanCancer (PCAWG), our
83 previous analyses shown that somatic L1 insertions represent the major
84 restructuring source of cancer genomes, especially important for head-and-neck
85 squamous cell carcinoma (HNSCC)[9]. Aberrant L1 retrotransposition contribute
86 to instability within cancer genomes, promoting cancer-driving rearrangements
87 that involve the loss of tumor-suppressor genes and/or the amplification of
88 oncogenes, and favoring some cancer clones to survive and grow [9].

89 Despite this demonstrated impact of L1 activation in cancer genomes, the
90 repetitive nature of L1 elements and their dispersion along the genome hinder
91 their real activation estimates in tumor samples, preventing its translation into
92 patient diagnosis/prognosis. This repetitive nature compels that the majority of
93 the available detection methods are based on Whole Genome Sequencing (WGS),
94 unfeasible for most hospitals in their diary practice, or require such good quality

95 DNA input quantities that remain unaffordable for a clinical practice mainly based
 96 on small and degraded tissue biopsies.

97 HNSCC represents the second tumor type with highest L1 activation [9], but the
 98 clinical implications of L1 retrotransposition are still to be elucidated. Head and
 99 Neck Cancer is a heterogeneous group of cancers of which more than 90% are
 100 diagnosed as HNSCC, arising in the stratified epithelium of the oral cavity, pharynx,
 101 and larynx [10,11]. Tumor development is often triggered by chronic exposure to
 102 tobacco or alcohol, while infection with high-risk human papillomaviruses (HPVs)
 103 causes a substantial and rising proportion of these tumors [12]. The lack of
 104 symptoms in the early stages together with the non-existent cancer biomarkers
 105 lead towards most diagnoses at advanced stages, where the 5-year survival rate is
 106 less than 50% [11]. Thus, there is an urgent need to find molecular biomarkers
 107 that can facilitate an earlier diagnosis and increase patients' life expectancy.

108 Based on this, we aimed to unravel the impact of possible L1 activation along
 109 HNSCC development in a clinical setting. First, we have developed a new efficient
 110 method based on L1 transductions, RetroTest, to measure L1 activation even in
 111 biopsies with low DNA inputs. Then, we have evaluated RetroTest sensitivity and
 112 specificity, demonstrating its high values. Finally, we have assessed L1
 113 retrotransposition along different tumor stages, unraveling for the first time a really
 114 early activation in HNSCC and field characterization, arising L1 as a promising
 115 early diagnostic biomarker.

116
 117

118 2. Materials and methods

119 2.1. Patients and tumor samples

120 Tumor samples and medical records were analyzed in a series of 96 HNSCC
 121 patients of 3 different cohorts (*Complejo Hospitalario Universitario de A Coruña*
 122 (*CHUAC*), *Biobanco Vasco* and *Fundación Pública Galega de Medicina Xenómica*
 123 (*FPGMX*)). The Ethical Committee for Clinical Research of Santiago-Lugo approved
 124 the study (CEIC 2018/567). Fresh-frozen and FFPE HNSCC tissue biopsies were
 125 collected, and characteristics of each tumor are specified in Table 1. The
 126 histopathologic status was confirmed by each corresponding Pathology
 127 Department following the latest TNM guidelines.

128

129 2.2. DNA isolation

130 Genomic DNA (gDNA) was extracted from fresh-frozen tissue and formalin-fixed
 131 paraffin-embedded (FFPE) samples using AllPrep DNA/RNA and AllPrep FFPE
 132 DNA/RNA Mini Kits (Qiagen), respectively. DNA quantification and integrity were
 133 assessed using Qubit dsDNA BR Assay Kit in Qubit 4.0 (ThermoFisher Scientific)
 134 and a 4200 TapeStation system (Agilent).

135

136 2.3. RetroTest method: design, library construction and target sequencing

137 During L1 transcription, the processing machinery sometimes bypasses the L1
 138 polyadenylation signal until a second 3' downstream polyadenylation site,
 139 mobilizing unique sequences downstream of the element in a process called *L1 3'*
 140 *transduction*. This process has been reported to occur in around 10% of the L1
 141 mobilizations [7] and can be used as an indirect measurement of real active L1
 142 elements. Accordingly, there are three different types of retrotranspositions: solo-

143 L1 (TD0), when a partial or complete L1 is retrotransposed; partnered
144 transductions (TD1), in which a L1 and downstream unique sequence are
145 retrotransposed; and orphan transductions (TD2), in which only the unique
146 sequence downstream of the active L1 is mobilized without the associated L1 [7].
147 RetroTest is based on targeted sequencing, where the capture probes are designed
148 against the unique sequence downstream of the 124 L1 full-length competent
149 elements previously described [7] using SureSelect design from Agilent. RetroTest
150 identifies L1 TD2 orphan insertions through the detection of discordant reads and
151 clipped reads. The main idea behind the method is that discordant read pairs,
152 where one of the mates is mapped to a L1 3' downstream sequence while the other
153 is mapped to the insertion target sequence, support an insertion. In addition,
154 clipped reads, mapped to the target sequence but containing a discordant extreme
155 blotting to a L1 3' downstream sequence, are detected to identify the breakpoint of
156 the insertion (Fig. 1A). Here, a minimum size of 2 reads per cluster and a minimum
157 of 4 supporting reads was specified to call a L1 insertion (Additional file 1).
158 Target libraries were constructed with SureSelect Target Enrichment System for
159 Illumina Paired-End Multiplexed Sequencing (Agilent). 100 ng of gDNA from each
160 sample were sheared using a Covaris M220 Focused-Ultrasonicator (Covaris) and
161 libraries and capture with targeted RNA baits were performed. The multiplexed
162 samples were sequenced with Illumina 150bp paired-end.
163 Sequencing reads were mapped to the hg19 reference genome by Burrows-
164 Wheeler Aligner BWA-mem [13]. Samtools [14] was used to sort the aligned reads
165 and to index the obtained bam file, applying Bammarkduplicates2 from Picard
166 tools [15] to mark duplicated reads.

167

168 2.4. RetroTest benchmarking

169 We generated simulated paired-end read datasets using ART [16] and several
170 commands from MEIGA-MEIsimulator, an in-house bioinformatic tool (Additional
171 file 1).

172 To mimic the capture process, we selected read pairs with at least one of the mates
173 mapping on a given set of target regions with Picard FilterSamReads v2.18.14 [15].

174 We used the resulting bam files to assess sensitivity and specificity under different
175 conditions. To test how our method performs with subclonal events, we simulated
176 transductions at different VAFs (10%, 20%, 40% and 50%). We also studied how
177 coverage affects the detection power of our algorithm. Using Picard
178 DownSampleSam v2.18.14 [15], we subsampled reads from a 150x simulation at
179 50% VAF under different sequencing depths (15x, 30x, 60x, 90x, 120x and 150x).

180 We compared RetroTest performance with Transposon Finder in Cancer (TraFiC),
181 used previously to explore somatic retrotransposition in PCAWG [7]. Both were
182 used to call the simulated events. Precision was calculated dividing the number of
183 true positive calls by the number of total calls performed by the method. Recall
184 was calculated dividing the number of true positive calls by the number of total
185 simulated events. True positives, False positives and False negatives were
186 identified by intersecting the coordinates of simulated events with the coordinates
187 of the calls using BEDTools intersect [17]. To compare RetroTest and TraFiC
188 results, Venn diagrams with the insertions detected by each method were plotted
189 by using *vennDiagram* R package.

190

191 2.5. Whole Genome Sequencing and determination of mutation profile.

192 To obtain WGS data, DNA was sent to an external service (Macrogen). Truseq Nano
193 DNA Libraries (350bp) were constructed and sequenced in a NovaSeq6000
194 Illumina platform (150bp paired-end). For a detailed explanation see Additional
195 File 1.

196

197 *2.6. Statistical analyses*

198 The association between L1 transduction rate (corrected by coverage) and patient
199 clinical features was assessed by multiple linear regressions. Wilcoxon or Fisher
200 test, depending on sample size, were applied to compare the differences in mean
201 values for clinical variables. Overall survival (OS), progression free survival (PFS)
202 and survival probability analyses were performed with the *survminer* and *survival*
203 R packages, using log-rank test to compare different groups. The association
204 between survival and clinical variables was evaluated by Cox regression.

205

206 *2.7. Enrichment analysis*

207 All enrichment analyses were performed using *enrichR* R package [18]. The
208 complete list of pathways databases can be found in Additional File 1.

209

210

211 3. Results

212 3.1. RetroTest Benchmarking

213 During L1 transcription, transductions are reported to occur in around 10% of the
 214 L1 mobilizations [7], moving also unique sequences downstream of the source
 215 element. RetroTest is designed to capture these mobilized and downstream-
 216 transduced unique sequences from orphan transductions (TD2) and use them as
 217 barcodes (Fig. 1A). These barcodes can identify unequivocally the insertions
 218 caused by these 124 L1 source elements active in cancer [7,9]. With this aim, we
 219 focused RetroTest probe design on the first 5000 nucleotides adjacent to the L1 3'
 220 regions for each of the 124 L1 source elements, since these are the regions most
 221 frequently transduced [7]. Since RetroTest uses transductions as an indirect
 222 measure of real L1 activation, we first compared L1 activation *versus* TD2 using the
 223 whole PCAWG data. In this way, we confirmed a high and statistically significant
 224 correlation between both events ($r=0.88$, $p=2.2e^{-16}$), and a relation of 1:10 as
 225 previously described (Fig. 1B).

226 Next, we optimized the lab protocol for both FFPE and fresh-frozen tissues and
 227 developed an associated bioinformatic pipeline. We evaluated the performance
 228 and accuracy of RetroTest in detecting L1 activation by generating an artificial
 229 cancer genome. In this genome, we have randomly distributed a total of 2480 L1
 230 transductions. Using simulations, we assessed the performance of RetroTest for
 231 different sequencing depth: 15x, 30x, 60x, 90x, 120x and 150x. RetroTest obtained
 232 a precision around 0.99 in all cases and a recall around 0.96 (Fig. 1C). We also
 233 evaluated the variation of the performance depending on the variant allele
 234 frequency (VAF) of the integrations, precisely for the following VAFs: 10%, 20%,
 235 40% and 50%. In this case, RetroTest obtained a precision around 1, decreasing

for lower VAFs, and a recall ranged from 0.81 to 0.96, augmenting as increasing the VAF (Fig. 1D).

Then, we compared RetroTest and the classical TraFiC, used for WGS in PCAWG [7]. As TraFiC is designed to work only with standard WGS 30x data, we compared the analysis varying exclusively the VAFs. TraFiC precision ranged from 0.99 to 0.88 and recall ranged from 0.11 to 0.87 as VAF increased (although the maximum recall was obtained for VAFs of 40%, being 0.87) (Fig. 1E) (Additional File 2: Table S1). We compared the performance of both methods by intersecting both calls in the mock tumor genome. Using a VAF of 50%, most of the variants were called by both methods, concretely 2216 variants, while RetroTest exclusively called 167, and TraFiC 33 private events, most of them resulting in false positives according to IGV (Fig. 1F).

3.2. L1 activation in HNSCC

Once confirmed the accuracy and precision of RetroTest, we decided to apply it to a more extensive HNSCC cohort, composed of 96 tumors along the tumor stages, from T1 to T4 (Table 1). We detected L1 activation in 75% of the patients (Fig. 2A), out of which 48.6% showed high activity (beyond the median) (Table 2). When the activation was studied along the different tumor stages of the disease, advanced disease (T3-T4) showed statistically higher L1 activation compared with early stages (T1-T2) ($p=0.0072$) (Fig. 2B). Following tumor staging, L1 activation was detected in 62.5% of T1 tumors, 63.1% of T2 tumors, 94.4% of T3 tumors, and 73.5% of the tumors in T4 (Table 2). The detection of L1 activation in all the tumor stages among the different patients, even in the first stage of the disease (T1), supports an early L1 activation during tumor progression.

261 Then, we analyzed L1 activation considering different patients' clinical
262 characteristics, finding no statistically significant association between the L1
263 activation and alcohol consumption ($p=0.14$) or sex ($p=0.055$). Interestingly,
264 smoker patients showed statistically significantly higher L1 activity than non-
265 smokers ($p=0.0015$) (Fig. 2C). We did not find an association between L1
266 activation and survival probability ($p=0.59$ active vs. inactive, $p=0.49$ high vs. low
267 rate) (Additional File 3: Fig. S1), although when considering those T1 patients who
268 have L1 already active at early stages, Kaplan-Meier curve showed that they tend
269 to have a lower survival probability, not being statistically significant ($p=0.43$)
270 (Fig. 2D).

271 Finally, in addition to localizing the position of each transduction, RetroTest can
272 also identify the source L1 element that has been mobilized. Thus, we have
273 detected L1 transductions in 23 genes throughout 27 patients. As shown in Fig. 2E,
274 most of these insertions are caused by a few source elements that resulted very
275 active in HNSCC, especially the 22q12.1.

276

277 *3.3. HNSCC mutation profile and L1 activation*

278 To further characterize the molecular profile of L1 activation in our HNSCC cohort,
279 we obtained WGS data from 19 tumor samples from patients measured by
280 RetroTest. To detect only somatic variation, their corresponding paired-normal
281 samples were included as germline control. We detected a median of 40 single
282 nucleotide variants (SNVs) and INDELs, identifying a total of 1012 somatic variants
283 throughout the cohort, affecting 918 genes (Additional File 2: Table S2). As shown
284 in Fig. 3A, we did not detect a correlation between L1 activation and the general
285 tumor mutation burden (TMB). Our results showed that the most frequently

286 mutated gene was *TP53* (36.8%), followed by *NOTCH1* (26.3%), *MT-ND5* (26.3%),
 287 *FAT1*, and *GRIN2A* (21.1%). Interestingly, we found that most of the patients with
 288 *TP53* mutations showed also high L1 activity (71.4%). In fact, when we compared
 289 the L1 activation with the *TP53* mutation, we found a clear association ($p=0.056$)
 290 (Fig. 3B). Enrichment analyses with the mutated genes showed involved key
 291 processes for cancer progression such as *Notch signaling*, *TGF β signaling*, and
 292 again *p53 activity regulation* (Fig. 3C) (Additional File 2: Table S3). CHEA analysis
 293 revealed the alteration of transcription factors related to epigenetic mechanisms
 294 including Polycomb members (*EZH2*, *SUZ12*) (Fig. 3D) (Additional File 2: Table S4).

295

296 3.4. Profiling normal adjacent tissues to the HNSCC tumor (NAT)

297 To further understand the early HNSCC features, we decided to evaluate a possible
 298 field cancerization process, in which the normal cell population is replaced by
 299 cancer-primed cells, without anatomical or morphological changes, but being
 300 already premalignant at molecular level. To this, we analyzed the available normal
 301 samples obtained from adjacent tissues to the tumor and the peripheral blood
 302 mononuclear cells (PBMCs), used as germline control (Fig. 4A). We detected a total
 303 of 25 high impact and/or possibly pathogenic somatic variants affecting the NAT
 304 (Additional File 2: Table S5). Most of these specific mutations ($n=20$; 80%) were
 305 exclusive from NAT tissue, including those affecting key genes such as *NOTCH1*
 306 (the most mutated gene), *FAT1* or *PPARD*; while 20% were shared between NAT
 307 and tumor tissue, affecting genes such as *CDKN2A*.

308 To elucidate whether L1 further supports field cancerization, we evaluated the
 309 NAT of a total of 9 patients by RetroTest. We compared the L1 elements active in
 310 the tumor, in the NAT, and in their corresponding paired germline. In this way, we

could confirm that most of the L1 activation was present only in the tumor, 5
 insertions resulted germline, and 1 element appeared active and shared by tumor
 and NAT. Surprisingly, 4 insertions appeared exclusively in NAT (Fig. 4B). Thus, we
 could confirm field cancerization and demonstrate that L1 is already active in NAT,
 supporting once again its early activation in HNSCC.

4. Discussion

Since the recently demonstrated high impact of L1 in cancer genomes [7,9], L1 has
 been evaluated as a cancer biomarker in different studies assessing its activity by
 evaluating L1 methylation, RNA expression, or protein levels [19–22]. However,
 most of L1 genome sequences are truncated and not functional, so results can
 present important biases in their L1 estimations, while its translation into clinical
 routine can be challenging due to RNA/protein instability. Previous proposed
 technologies based on DNA full-length L1 retrotransposon capture sequencing,
 such as RC-seq [23], present the associated possible bias of the real L1 activation,
 besides only offer global L1 estimates, and important DNA requirements (2.5µg
 starting genomic DNA), unaffordable for clinical practice. The most recent
 approaches are based on the identification of L1 insertions as real L1 activation
 measure (TraFiC [7], xTea [24], MELT [25] and Mobster [26]). They use Whole
 Genome or Whole Exome Sequencing from Illumina pair-end short reads,
 nevertheless, short reads hamper the detection of transposable elements
 insertions in highly repetitive or complex rearrangement regions. Additionally,
 these high throughput approaches are not affordable for most of the hospitals.
 More recently, long reads sequencing has arisen as a new possibility (xTea [24],
 PALMER [27]), but again their requirements of huge amounts of high-quality DNA

336 remain unaffordable for biopsies mainly composed of small samples and
337 fragmented DNA. Thus, incorporating L1 activation into the clinical practice
338 requires new methods supported by standardization and rigorous validation to
339 demonstrate its utility.

340 Here, we present RetroTest: a new method to detect real L1 activation in clinical
341 samples with low DNA input requirements, both from fresh/frozen or FFPE
342 biopsies. Its novelty and power lie in the targeted detection of active source L1
343 elements, which showed a highly and direct statistical correlation with total L1
344 activity (as demonstrated by our data) in a more cost-effective than previously
345 proposed approaches. Our method not only offers global L1 estimates but also
346 identifies the L1 elements active in each real sample. Our benchmarking supports
347 the high precision (1) and recall of RetroTest (0.81-0.96), improving according to
348 coverage, even reaching the possibility of detecting subclonal insertions. The high
349 coincidence in the variants called by TraFic and RetroTest, in both simulations and
350 real-world data, supports the high potential for this new methodology.

351 L1 somatic retrotransposition was previously demonstrated as the second most
352 frequent type of structural variants among HNSCC genomes [9]. Accordingly, our
353 results demonstrate that most HNSCC patients (75%) present L1 activation, even
354 near half of them with high levels (49%). The activation was higher in late stages,
355 as found in Barret's esophagus, where lower L1 activity was detected in the early
356 stages increasing with cancer progression [28]. Interestingly, we found a
357 surprisingly early activation already in the first stages of HNSCC, with activation of
358 L1 in 62-63% of the T1 patients. These data pointed towards L1 activation as an
359 early event in the configuration of HNSCC genomes and, thus, in the development

360 of the disease. In fact, those stage T1 patients with L1 activation tended to present
361 a lower OS.

362 We found an association between high L1 activity and smoking habits. Previous
363 studies had reported higher hypomethylation rates of L1 in smokers [29], even in
364 non-cancerous epithelial tissues [30]. Considering that around 75% of HNSCC are
365 associated with tobacco [31], this mechanism could be responsible for the high L1
366 activity levels in HNSCC, since methylation is one of the best-demonstrated
367 mechanisms preventing L1 reactivation [7,32]. In fact, L1 hypomethylation has
368 been associated with worse prognosis in HNSCC, including a higher risk of relapse
369 [33–35]. This hypomethylation has been found as an early event in CRC, gastric
370 and oral cancer [36–39]. Interestingly, it was reported that L1 methylation levels
371 were significantly lower in oral premalignant lesions of patients who then
372 developed oral cancer [39].

373 The source element L1 most active in our cohort was that at 22q12.1, coincident
374 with our previous results [7, 10]. This L1 is located antisense to an intron of the
375 TTC28 gene and has been also recently identified as the intact LINE-1 mRNA most
376 highly expressed in breast, ovarian and colon cancer [40], and the L1 element
377 accounting for most transductions in colorectal cancer [33], supporting the same
378 hottest activity in HNSCC.

379 Our WGS analyses confirmed no correlation between L1 activity and TMB, but an
380 association between *TP53* mutation and L1 activation, in line with previous studies
381 suggesting that *TP53* can repress L1 mobilization [9,40–43]. We also identified an
382 important epigenetic regulation among the transcription factors strongly
383 interacting with the mutated genes, specially related to repressive complex
384 Polycomb. Intriguingly, Mangoni et al. have just deciphered that L1 RNAs can act as

385 long non-coding RNAs and directly interact with the Polycomb during brain
386 development and evolution [44]. We also reported previously that Polycomb could
387 regulate lncRNA *HOTAIR* in bladder cancer [45]. In the same sense, Ishak et al.
388 demonstrated an EZH2-dependent silencing of genomic repeat sequences,
389 including L1 elements [46]. Therefore, additional analyses would be required to
390 further address if this epigenetic network plays a role in cancer genome
391 reorganization.

392 Then, to evaluate possible HNSCC early diagnosis potential biomarkers is
393 fundamental to get insights into the transition from pre-tumoral to cancer disease.
394 Thus, we evaluated the presence of field cancerization and identified genetic
395 features that can eventually lead to cancer development [47], finding exclusive
396 somatic mutations in the NAT and a small proportion shared with the tumor.
397 Several studies have recently described patchworks of different clones in normal
398 tissues, some of them even bearing driver mutations, increasing with age or
399 smoking habits [48–50]. We found *NOTCH1* and *FAT1* as the most mutated genes,
400 as Martincorena's results in normal skin and esophagus [48,51]. These genes
401 showed also high mutation rates in the tumor, indicating the presence of a
402 precancerous or cancer invasion field.

403 Finally, when we evaluated L1 activity, several mobilizations were exclusively
404 found in NAT, and only one appeared shared with the tumor, again supporting a
405 cancerization field. Somatic L1 retrotransposition events have been recently
406 described in normal urothelium and colorectal epithelium, although with much
407 lower rates than in bladder and colorectal cancer [52,53]. L1 activity has been also
408 described in a few pre-tumor samples including Barrett's esophagus and colorectal
409 adenomas [8,28,54]. Therefore, L1 activation arises as a reinforced early event,

even in pre-tumor stages, in the natural history of HNSCC, with the associated potential of L1 as a field cancerization biomarker.

Conclusions

We present the development and benchmarking of RetroTest, which allows an easy measurement of real L1 activation from small tumor biopsies with high efficacy, favoring its implementation in real clinical settings. RetroTest revealed that most of the HNSCC patients present L1 activation, associated with smoking habits and already active in early stages of the disease and NAT, supporting a field cancerization process and its potential as early diagnostic biomarker.

Acknowledgements

Authors thank all the enrolled patients and their families. This research project was made possible through the access granted by the Galician Supercomputing Center (CESGA) to its supercomputing infrastructure. The supercomputer FinisTerrae III and its permanent data storage system have been funded by the Spanish Ministry of Science and Innovation, the Galician Government and the European Regional Development Fund (ERDF).

This work was supported by the Instituto de Salud Carlos III (ISCIII) and the European Social Fund (“Investing in your future”) (PI19/01113, and partially by P121/00208, co-funded by FEDER and the European Union), and the Spanish Association Against Cancer Scientific Foundation (IDEAS19122MART). A.O. was supported by a predoctoral fellowship from the Galician Innovation Agency, Xunta de Galicia (ED481A-2020/214). M.M.-F. and J.B. were previously supported by the Spanish Association Against Cancer Scientific Foundation (INVES207MART and

PRDCR19007BREA_001, respectively). M.M.-F. is currently supported by the Miguel Servet program (CP20/00188) from the Instituto de Salud Carlos III (ISCIII) and the European Social Fund (“Investing in your future”). M.G.-G. is supported by a postdoctoral fellowship from the Galician Innovation Agency, Xunta de Galicia (IN606B-2024/014).

Ethics statement

The Ethical Committee for Clinical Research of Santiago-Lugo gave ethical approval for this work (CEIC 2018/567).

Competing interests

The authors declare no competing interests.

Consent for publication

All authors give consent for the publication of this manuscript.

Data availability statement

Whole Genome Sequencing data from the HNSCC cohort is deposited into the Sequence Read Archive (SRA) repository under the following BioProject ID: PRJNA1053897. RetroTest pipeline is publicly available at <https://gitlab.com/mobilegenomesgroup/RETROTEST>.

CRedit authorship contribution statement

Jenifer Brea-Iglesias and Ana Oitabén: Conceptualization, Methodology, Software, Data curation, Investigation, Validation, Visualization, Writing – review

460 and editing. **Sonia Zumalave:** Conceptualization, Investigation, Methodology,
 461 Software, Validation, Writing – review and editing. **Bernardo Rodríguez-Martin:**
 462 Conceptualization, Methodology, Software. **María Gallardo-Gómez:**
 463 Conceptualization, Visualization, Writing – review and editing. **Martin**
 464 **Santamarina:** Methodology, Validation. **Ana Pequeño-Valtierra:**
 465 Conceptualization, Investigation. **Laura Juaneda-Magdalena:** Methodology,
 466 Conceptualization, Writing - review and editing. **Ramón García-Escudero:**
 467 Methodology, Writing - review and editing. **Jose Luis López-Cedrún:**
 468 Methodology. **Máximo Fraga:** Funding acquisition. **José MC Tubio:**
 469 Conceptualization, Funding acquisition. **Mónica Martínez-Fernández:** Writing –
 470 review & editing, Writing – original draft, Supervision, Resources, Investigation,
 471 Project administration, Funding acquisition, Conceptualization. All authors
 472 approved the final version.

473

474 **References**

- 475 [1] H.H. Kazazian, J. V. Moran, The impact of L1 retrotransposons on the human
 476 genome, *Nat. Genet.* 19 (1998) 19–24. <https://doi.org/10.1038/NG0598-19>.
 477 [2] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K.
 478 Devon, K. Dewar, M. Doyle, W. Fitzhugh, R. Funke, D. Gage, K. Harris, A.
 479 Heaford, J. Howland, L. Kann, J. Lehoczký, R. Levine, P. McEwan, K.
 480 McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C.
 481 Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann,
 482 N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough,
 483 S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P.
 484 Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S.

485 Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray,
 486 L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross,
 487 R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D.
 488 McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin,
 489 W.R. Gish, S.L. Chisoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A.
 490 Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W.
 491 Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T.
 492 Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M.
 493 Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C.
 494 Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati,
 495 D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A.
 496 Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J.
 497 Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E.
 498 Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S.
 499 Taudien, A. Rump, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K.
 500 Weinstock, M.L. Hong, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L.
 501 Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J.
 502 Proctor, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R.
 503 McCombie, M. De La Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G.
 504 Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E.
 505 Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M.
 506 Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan,
 507 J.G.R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K.
 508 Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kasprzyk, S. Kennedy,
 509 W.J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A.

510 McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G.
511 Schuler, J. Schultz, G. Slater, A.F.A. Smit, E. Stupka, J. Szustakowki, D. Thierry-
512 Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf,
513 K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A.
514 Felsenfeld, K.A. Wetterstrand, R.M. Myers, J. Schmutz, M. Dickson, J.
515 Grimwood, D.R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K.
516 Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, A. Patrinos,
517 M.J. Morgan, Initial sequencing and analysis of the human genome, Nature
518 409 (2001) 860–921. <https://doi.org/10.1038/35057062>.
519 [3] C.R. Beck, P. Collier, C. Macfarlane, M. Malig, J.M. Kidd, E.E. Eichler, R.M.
520 Badge, J. V. Moran, LINE-1 retrotransposition activity in human genomes,
521 Cell 141 (2010) 1159–1170.
522 [4] B. Brouha, J. Schustak, R.M. Badge, S. Lutz-Prigge, A.H. Farley, J. V. Morant,
523 H.H. Kazazian, Hot L1s account for the bulk of retrotransposition in the
524 human population, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 5280–5285.
525 [5] S.J. Hoyt, J.M. Storer, G.A. Hartley, P.G.S. Grady, A. Gershman, L.G. de Lima, C.
526 Limouse, R. Halabian, L. Wojenski, M. Rodriguez, N. Altemose, A. Rhie, L.J.
527 Core, J.L. Gerton, W. Makalowski, D. Olson, J. Rosen, A.F.A. Smit, A.F. Straight,
528 M.R. Vollger, T.J. Wheeler, M.C. Schatz, E.E. Eichler, A.M. Phillippy, W. Timp,
529 K.H. Miga, R.J. O'Neill, From telomere to telomere: The transcriptional and
530 epigenetic state of human repeat elements, Science. 376 (2022).
531 <https://doi.org/10.1126/science.abk3112>.
532 [6] D.M. Sassaman, B.A. Dombroski, J. V. Moran, M.L. Kimberland, T.P. Naas, R.J.
533 DeBerardinis, A. Gabriel, G.D. Swergold, H.H. Kazazian, Many human L1
534 elements are capable of retrotransposition, Nat. Genet. 16 (1997) 37–43.

- 535 [7] J.M.C. e. al Tubio, Extensive transduction of nonrepetitive DNA mediated by
536 L1retrotransposition in cancer genomes, *Science*. 345 (2014) 1251343.
537 <https://doi.org/10.1126/science.1251343>.Extensive.
- 538 [8] A.D. Ewing, A. Gacita, L.D. Wood, F. Ma, D. Xing, M.S. Kim, S.S. Manda, G. Abril,
539 G. Pereira, A. Makohon-Moore, L.H.J. Looijenga, A.J.M. Gillis, R.H. Hruban, R.A.
540 Anders, K.E. Romans, A. Pandey, C.A. Iacobuzio-Donahue, B. Vogelstein, K.W.
541 Kinzler, H.H. Kazazian, S. Solyom, Widespread somatic L1 retrotransposition
542 occurs early during gastrointestinal cancer evolution, *Genome Res*. 25
543 (2015) 1536–1545. <https://doi.org/10.1101/gr.196238.115>.
- 544 [9] B. Rodriguez-Martin, E.G. Alvarez, A. Baez-Ortega, J. Zamora, F. Supek, J.
545 Demeulemeester, M. Santamarina, Y.S. Ju, J. Temes, D. Garcia-Souto, H.
546 Detering, Y. Li, J. Rodriguez-Castro, A. Dueso-Barroso, A.L. Bruzos, S.C.
547 Dentro, M.G. Blanco, G. Contino, D. Ardeljan, M. Tojo, N.D. Roberts, S.
548 Zumalave, P.A.W. Edwards, J. Weischenfeldt, M. Puiggròs, Z. Chong, K. Chen,
549 E.A. Lee, J.A. Wala, K. Raine, A. Butler, S.M. Waszak, F.C.P. Navarro, S.E.
550 Schumacher, J. Monlong, F. Maura, N. Bolli, G. Bourque, M. Gerstein, P.J. Park,
551 D.C. Wedge, R. Beroukhim, D. Torrents, J.O. Korbel, I.I. Martincorena, R.C.
552 Fitzgerald, P. Van Loo, H.H. Kazazian, K.H. Burns, K.C. Akdemir, E.G. Alvarez,
553 A. Baez-Ortega, R. Beroukhim, P.C. Boutros, D.D.L. Bowtell, B. Brors, K.H.
554 Burns, P.J. Campbell, K. Chan, I. Cortés-Ciriano, A. Dueso-Barroso, A.J.
555 Dunford, P.A.W. Edwards, X. Estivill, D. Etemadmoghadam, L. Feuerbach, J.L.
556 Fink, M. Frenkel-Morgenstern, D.W. Garsed, M. Gerstein, D.A. Gordenin, D.
557 Haan, J.E. Haber, J.M. Hess, B. Hutter, M. Imielinski, D.T.W. Jones, M.D.
558 Kazanov, L.J. Klimczak, Y. Koh, J.O. Korbel, K. Kumar, E.A. Lee, J.J.K. Lee, Y. Li,
559 A.G. Lynch, G. Macintyre, F. Markowetz, A. Martinez-Fundichely, M.

560 Meyerson, S. Miyano, H. Nakagawa, F.C.P. Navarro, S. Ossowski, J. V. Pearson,
561 J. V. Pearson, K. Rippe, N.D. Roberts, S.A. Roberts, B. Rodriguez-Martin, B.
562 Rodriguez-Martin, S.E. Schumacher, M. Shackleton, N. Sidiropoulos, L.
563 Sieverling, C. Stewart, J.M.C. Tubio, I. Villasante, N. Waddell, J.A. Wala, J.
564 Weischenfeldt, L. Yang, X. Yao, S.S. Yoon, J. Zamora, C.Z. Zhang, P.J. Campbell,
565 J.M.C. Tubio, S.E. Schumacher, R. Scully, B. Rodriguez-Martin, Y.S. Ju, M.D.
566 Kazanov, L.J. Klimczak, Y. Koh, J.O. Korb, K. Kumar, E.A. Lee, J.J.K. Lee, Y. Li,
567 A.G. Lynch, G. Macintyre, F. Markowetz, I.I. Martincorena, A. Martinez-
568 Fundichely, M. Meyerson, S. Miyano, H. Nakagawa, F.C.P. Navarro, S.
569 Ossowski, J. V. Pearson, M. Puiggròs, K. Rippe, N.D. Roberts, S.A. Roberts, B.
570 Rodriguez-Martin, S.E. Schumacher, R. Scully, M. Shackleton, N. Sidiropoulos,
571 L. Sieverling, C. Stewart, J.M.C. Tubio, I. Villasante, N. Waddell, J.A. Wala, J.
572 Weischenfeldt, L. Yang, X. Yao, S.S. Yoon, J. Zamora, C.Z. Zhang, P.J. Campbell,
573 J.M.C. Tubio, Pan-cancer analysis of whole genomes identifies driver
574 rearrangements promoted by LINE-1 retrotransposition, Nat. Genet. 52
575 (2020) 306–319. <https://doi.org/10.1038/s41588-019-0562-0>.
576 [10] A. Jou, J. Hess, Epidemiology and Molecular Biology of Head and Neck
577 Cancer, Oncol. Res. Treat. 40 (2017) 328–332.
578 <https://doi.org/10.1159/000477127>.
579 [11] M. Plath, J. Gass, M. Hlevnjak, Q. Li, B. Feng, X.P. Hostench, M. Bieg, L.
580 Schroeder, D. Holzinger, M. Zapatka, K. Freier, W. Weichert, J. Hess, K. Zaoui,
581 Unraveling most abundant mutational signatures in head and neck cancer,
582 Int. J. Cancer 148 (2021) 115–127. <https://doi.org/10.1002/IJC.33297>.
583 [12] C.R. Leemans, B.J.M. Braakhuis, R.H. Brakenhoff, The molecular biology of
584 head and neck cancer, Nat. Rev. Cancer 11 (2011) 9–22.

585 <https://doi.org/10.1038/nrc2982>.

586 [13] H. Li, Aligning sequence reads, clone sequences and assembly contigs with
587 BWA-MEM, ArXiv: Genomics (2013).
588 <https://doi.org/10.6084/M9.FIGSHARE.963153.V1>.

589 [14] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A.
590 Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, Twelve years of SAMtools
591 and BCFtools, Gigascience 10 (2021) 1–4.
592 <https://doi.org/10.1093/GIGASCIENCE/GIAB008>.

593 [15] Broad Institute, Picard Tools, <Http://Broadinstitute.Github.Io/Picard/> (n.d.).

594 [16] W. Huang, L. Li, J.R. Myers, G.T. Marth, ART: a next-generation sequencing
595 read simulator, Bioinformatics 28 (2012) 593–594.
596 <https://doi.org/10.1093/BIOINFORMATICS/BTR708>.

597 [17] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing
598 genomic features, Bioinformatics 26 (2010) 841–842.
599 <https://doi.org/10.1093/BIOINFORMATICS/BTQ033>.

600 [18] M. V. Kuleshov, M.R. Jones, A.D. Rouillard, N.F. Fernandez, Q. Duan, Z. Wang,
601 S. Koplev, S.L. Jenkins, K.M. Jagodnik, A. Lachmann, M.G. McDermott, C.D.
602 Monteiro, G.W. Gundersen, A. Maayan, Enrichr: a comprehensive gene set
603 enrichment analysis web server 2016 update, Nucleic Acids Res. 44 (2016)
604 W90–W97. <https://doi.org/10.1093/NAR/GKW377>.

605 [19] D. Ardeljan, M.S. Taylor, D.T. Ting, K.H. Burns, The Human Long Interspersed
606 Element-1 Retrotransposon: An Emerging Biomarker of Neoplasia, Clin.
607 Chem. 63 (2017) 816–822.
608 <https://doi.org/10.1373/CLINCHEM.2016.257444>.

609 [20] M.L. Filipenko, U.A. Boyarskikh, L.S. Leskov, K. V. Subbotina, E.A. Khrapov, A.

- 610 V. Sokolov, I.S. Stilidi, N.E. Kushlinskii, The Level of LINE-1 mRNA Is
611 Increased in Extracellular Circulating Plasma RNA in Patients with
612 Colorectal Cancer, Bull. Exp. Biol. Med. 173 (2022) 261–264.
613 <https://doi.org/10.1007/S10517-022-05530-2/METRICS>.
- 614 [21] S. Sato, M. Gillette, P.R. de Santiago, E. Kuhn, M. Burgess, K. Doucette, Y. Feng,
615 C. Mendez-Dorantes, P.J. Ippoliti, S. Hobday, M.A. Mitchell, K. Doberstein, S.M.
616 Gysler, M.S. Hirsch, L. Schwartz, M.J. Birrer, S.J. Skates, K.H. Burns, S.A. Carr,
617 R. Drapkin, LINE-1 ORF1p as a candidate biomarker in high grade serous
618 ovarian carcinoma, Sci. Rep. 13 (2023). [https://doi.org/10.1038/s41598-](https://doi.org/10.1038/s41598-023-28840-5)
619 [023-28840-5](https://doi.org/10.1038/s41598-023-28840-5).
- 620 [22] M.S. Taylor, C. Wu, P.C. Fridy, S.J. Zhang, Y. Senussi, J.C. Wolters, T. Cajuso, W.-
621 C. Cheng, J.D. Heaps, B.D. Miller, K. Mori, L. Cohen, H. Jiang, K.R. Molloy, B.T.
622 Chait, M.G. Goggins, I. Bhan, J.W. Franses, X. Yang, M.-E. Taplin, X. Wang, D.C.
623 Christiani, B.E. Johnson, M. Meyerson, R. Uppaluri, A.M. Egloff, E.N. Denault,
624 L.M. Spring, T.-L. Wang, I.-M. Shih, J.E. Fairman, E. Jung, K.S. Arora, O.H.
625 Yilmaz, S. Cohen, T. Sharova, G. Chi, B.L. Norden, Y. Song, L.T. Nieman, L.
626 Pappas, A.R. Parikh, M.R. Strickland, R.B. Corcoran, T. Mustelin, G. Eng, O.H.
627 Yilmaz, U.A. Matulonis, S.J. Skates, B.R. Rueda, R. Drapkin, S.J. Klempner, V.
628 Deshpande, D.T. Ting, M.P. Rout, J. LaCava, D.R. Walt, K.H. Burns,
629 Ultrasensitive detection of circulating LINE-1 ORF1p as a specific multi-
630 cancer biomarker, Cancer Discov. 13 (2023) OF1–OF16.
631 [https://doi.org/10.1158/2159-8290.CD-23-](https://doi.org/10.1158/2159-8290.CD-23-0313/729035/AM/ULTRASENSITIVE-DETECTION-OF-CIRCULATING-LINE-1)
632 [0313/729035/AM/ULTRASENSITIVE-DETECTION-OF-CIRCULATING-LINE-](https://doi.org/10.1158/2159-8290.CD-23-0313/729035/AM/ULTRASENSITIVE-DETECTION-OF-CIRCULATING-LINE-1)
633 [1](https://doi.org/10.1158/2159-8290.CD-23-0313/729035/AM/ULTRASENSITIVE-DETECTION-OF-CIRCULATING-LINE-1).
- 634 [23] J.K. Baillie, M.W. Barnett, K.R. Upton, D.J. Gerhardt, T.A. Richmond, F. De

635 Sapio, P.M. Brennan, P. Rizzu, S. Smith, M. Fell, R.T. Talbot, S. Gustincich, T.C.
636 Freeman, J.S. Mattick, D.A. Hume, P. Heutink, P. Carninci, J.A. Jeddloh, G.J.
637 Faulkner, Somatic retrotransposition alters the genetic landscape of the
638 human brain, *Nature* 479 (2011) 534–537.
639 <https://doi.org/10.1038/nature10531>.

640 [24] C. Chu, R. Borges-Monroy, V. V. Viswanadham, S. Lee, H. Li, E.A. Lee, P.J. Park,
641 Comprehensive identification of transposable element insertions using
642 multiple sequencing technologies, *Nat. Commun.* 12 (2021).
643 <https://doi.org/10.1038/s41467-021-24041-8>.

644 [25] E.J. Gardner, V.K. Lam, D.N. Harris, N.T. Chuang, E.C. Scott, W. Stephen
645 Pittard, R.E. Mills, S.E. Devine, The mobile element locator tool (MELT):
646 Population-scale mobile element discovery and biology, *Genome Res.* 27
647 (2017) 1916–1929. <https://doi.org/10.1101/GR.218032.116/-/DC1>.

648 [26] D.T. jwa. Thung, J. de Ligt, L.E.M. Vissers, M. Steehouwer, M. Kroon, P. de
649 Vries, E.P. Slagboom, K. Ye, J.A. Veltman, J.Y. Hehir-Kwa, Mobster: accurate
650 detection of mobile element insertions in next generation sequencing data,
651 *Genome Biol.* 15 (2014) 488. [https://doi.org/10.1186/S13059-014-0488-](https://doi.org/10.1186/S13059-014-0488-X/FIGURES/3)
652 [X/FIGURES/3](https://doi.org/10.1186/S13059-014-0488-X/FIGURES/3).

653 [27] W. Zhou, S.B. Emery, D.A. Flasch, Y. Wang, K.Y. Kwan, J.M. Kidd, J. V. Moran,
654 R.E. Mills, Identification and characterization of occult human-specific LINE-
655 1 insertions using long-read sequencing technology, *Nucleic Acids Res.* 48
656 (2020) 1146–1163. <https://doi.org/10.1093/NAR/GKZ1173>.

657 [28] A.C. Katz-Summercorn, S. Jammula, A. Frangou, I. Peneva, M. O'Donovan, M.
658 Tripathi, S. Malhotra, M. di Pietro, S. Abbas, G. Devonshire, W. Januszewicz, A.
659 Blasko, K. Nowicki-Osuch, S. MacRae, A. Northrop, A.M. Redmond, D.C.

- 660 Wedge, R.C. Fitzgerald, Multi-omic cross-sectional cohort study of pre-
661 malignant Barrett's esophagus reveals early structural variation and
662 retrotransposon activity, *Nat. Commun.* 13 (2022).
663 <https://doi.org/10.1038/s41467-022-28237-4>.
- 664 [29] A.W. Caliri, A. Caceres, S. Tommasi, A. Besaratinia, Hypomethylation of LINE-
665 1 repeat elements and global loss of DNA hydroxymethylation in vapers and
666 smokers, *Epigenetics* 15 (2020) 816–829.
667 <https://doi.org/10.1080/15592294.2020.1724401>.
- 668 [30] H. Shigaki, Y. Baba, M. Watanabe, S. Iwagami, K. Miyake, T. Ishimoto, M.
669 Iwatsuki, H. Baba, LINE-1 hypomethylation in noncancerous esophageal
670 mucosae is associated with smoking history, *Ann. Surg. Oncol.* 19 (2012)
671 4238–4243. <https://doi.org/10.1245/s10434-012-2488-y>.
- 672 [31] P. Vineis, M. Alavanja, P. Buffler, E. Fontham, S. Franceschi, Y.T. Gao, P.C.
673 Gupta, A. Hackshaw, E. Matos, J. Samet, F. Sitas, J. Smith, L. Stayner, K. Straif,
674 M.J. Thun, H.E. Wichmann, A.H. Wu, D. Zaridze, R. Peto, R. Doll, Tobacco and
675 cancer: recent epidemiological evidence, *J. Natl. Cancer Inst.* 96 (2004) 99–
676 106. <https://doi.org/10.1093/JNCI/DJH014>.
- 677 [32] K. Hur, P. Cejas, J. Feliu, J. Moreno-Rubio, E. Burgos, C.R. Boland, A. Goel,
678 Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to
679 activation of proto-oncogenes in human colorectal cancer metastasis, *Gut* 63
680 (2014) 635–646. <https://doi.org/10.1136/GUTJNL-2012-304219>.
- 681 [33] M. Casarotto, V. Lupato, G. Giurato, R. Guerrieri, S. Sulfaro, A. Salvati, E.
682 D'Angelo, C. Furlan, A. Menegaldo, L. Baboci, B. Montico, I. Turturici, R.
683 Dolcetti, S. Romeo, V. Baggio, S. Corrado, G. Businello, M. Guido, A. Weisz, V.
684 Giacomarra, G. Franchin, A. Steffan, L. Sigalotti, E. Vaccher, P. Boscolo-Rizzo,

685 P. Jerry, G. Fanetti, E. Fratta, LINE-1 hypomethylation is associated with poor
686 outcomes in locoregionally advanced oropharyngeal cancer, Clin.
687 Epigenetics 14 (2022). <https://doi.org/10.1186/s13148-022-01386-5>.

688 [34] C. Furlan, J. Polesel, L. Barzan, G. Franchin, S. Sulfaro, S. Romeo, F. Colizzi, A.
689 Rizzo, V. Baggio, V. Giacomarra, A.P. Dei Tos, P. Boscolo-Rizzo, E. Vaccher, R.
690 Dolcetti, L. Sigalotti, E. Fratta, Prognostic significance of LINE-1
691 hypomethylation in oropharyngeal squamous cell carcinoma, Clin.
692 Epigenetics 9 (2017). [https://doi.org/10.1186/S13148-017-0357-](https://doi.org/10.1186/S13148-017-0357-Z/FIGURES/4)
693 [Z/FIGURES/4](https://doi.org/10.1186/S13148-017-0357-Z/FIGURES/4).

694 [35] K. Misawa, S. Yamada, M. Mima, T. Nakagawa, T. Kurokawa, A. Imai, D.
695 Mochizuki, D. Shinmura, T. Yamada, J. Kita, R. Ishikawa, Y. Yamaguchi, Y.
696 Misawa, T. Kanazawa, H. Kawasaki, H. Mineta, Long interspersed nuclear
697 element 1 hypomethylation has novel prognostic value and potential utility
698 in liquid biopsy for oral cavity cancer, Biomark. Res. 8 (2020) 1–10.
699 <https://doi.org/10.1186/S40364-020-00235-Y/FIGURES/6>.

700 [36] A. Benard, C.J.H. Van De Velde, L. Lessard, H. Putter, L. Takeshima, P.J.K.
701 Kuppen, D.S.B. Hoon, Epigenetic status of LINE-1 predicts clinical outcome in
702 early-stage rectal cancer, Br. J. Cancer 109 (2013) 3073–3083.
703 <https://doi.org/10.1038/bjc.2013.654>.

704 [37] E.J. Kim, W.C. Chung, D.B. Kim, Y.J. Kim, J.M. Lee, J.H. Jung, Y.K. Lee, Long
705 interspersed nuclear element (LINE)-1 methylation level as a molecular
706 marker of early gastric cancer, Dig. Liver Dis. 48 (2016) 1093–1097.
707 <https://doi.org/10.1016/j.dld.2016.06.002>.

708 [38] E. Sunami, M. de Maat, A. Vu, R.R. Turner, D.S.B. Hoon, LINE-1
709 Hypomethylation During Primary Colon Cancer Progression, PLoS One 6

710 (2011) e18884. <https://doi.org/10.1371/JOURNAL.PONE.0018884>.

711 [39] J.P. Foy, C.R. Pickering, V.A. Papadimitrakopoulou, J. Jelinek, S.H. Lin, W.N.

712 William, M.J. Frederick, J. Wang, W. Lang, L. Feng, L. Zhang, E.S. Kim, Y.H. Fan,

713 W.K. Hong, A.K. El-Naggar, J.J. Lee, J.N. Myers, J.P. Issa, S.M. Lippman, L. Mao,

714 P. Saintigny, New DNA methylation markers and global DNA

715 hypomethylation are associated with oral cancer development, *Cancer Prev.*

716 *Res.* 8 (2015) 1027–1035. [https://doi.org/10.1158/1940-6207.CAPR-14-](https://doi.org/10.1158/1940-6207.CAPR-14-0179)

717 0179.

718 [40] W. McKerrow, X. Wang, C. Mendez-Dorantes, P. Mita, S. Cao, M. Grivainis, L.

719 Ding, J. LaCava, K.H. Burns, J.D. Boeke, D. Fenyő, LINE-1 expression in cancer

720 correlates with p53 mutation, copy number alteration, and S phase

721 checkpoint, *Proc. Natl. Acad. Sci. U. S. A.* 119 (2022).

722 <https://doi.org/10.1073/PNAS.2115999119/-/DCSUPPLEMENTAL>.

723 [41] B. Tiwari, A.E. Jones, C.J. Caillet, S. Das, S.K. Royer, J.M. Abrams, P53 directly

724 represses human LINE1 transposons, *Genes Dev.* 34 (2020) 1439–1451.

725 <https://doi.org/10.1101/GAD.343186.120/-/DC1>.

726 [42] P. Mita, X. Sun, D. Fenyő, D.J. Kahler, D. Li, N. Agmon, A. Wudzinska, S.

727 Keegan, J.S. Bader, C. Yun, J.D. Boeke, BRCA1 and S phase DNA repair

728 pathways restrict LINE-1 retrotransposition in human cells, *Nat. Struct. Mol.*

729 *Biol.* 27 (2020) 179–191. <https://doi.org/10.1038/S41594-020-0374-Z>.

730 [43] N. Rodić, R. Sharma, R. Sharma, J. Zampella, L. Dai, M.S. Taylor, R.H. Hruban,

731 C.A. Iacobuzio-Donahue, A. Maitra, M.S. Torbenson, M. Goggins, I.M. Shih, A.S.

732 Duffield, E.A. Montgomery, E. Gabrielson, G.J. Netto, T.L. Lotan, A.M. De

733 Marzo, W. Westra, Z.A. Binder, B.A. Orr, G.L. Gallia, C.G. Eberhart, J.D. Boeke,

734 C.R. Harris, K.H. Burns, Long interspersed element-1 protein expression is a

735 hallmark of many human cancers, *Am. J. Pathol.* 184 (2014) 1280–1286.
736 <https://doi.org/10.1016/j.ajpath.2014.01.007>.

737 [44] D. Mangoni, A. Simi, P. Lau, A. Armaos, F. Ansaloni, A. Codino, D. Damiani, L.
738 Floreani, V. Di Carlo, D. Vozzi, F. Persichetti, C. Santoro, L. Pandolfini, G.G.
739 Tartaglia, R. Sanges, S. Gustincich, LINE-1 regulates cortical development by
740 acting as long non-coding RNAs, *Nat. Commun.* 14 (2023).
741 <https://doi.org/10.1038/s41467-023-40743-7>.

742 [45] M. Martínez-Fernández, A. Feber, M. Dueñas, C. Segovia, C. Rubio, M.
743 Fernandez, F. Villacampa, J. Duarte, F.F. López-Calderón, M.J. Gómez-
744 Rodriguez, D. Castellano, J.L. Rodriguez-Peralto, F. De La Rosa, S. Beck, J.M.
745 Paramio, Analysis of the polycomb-related lncRNAs HOTAIR and ANRIL in
746 bladder cancer, *Clin. Epigenetics* 7 (2015). [https://doi.org/10.1186/s13148-](https://doi.org/10.1186/s13148-015-0141-x)
747 015-0141-x.

748 [46] C.A. Ishak, A.E. Marshall, D.T. Passos, C.R. White, J. Seung, M.J. Cecchini, S.
749 Ferwati, W.A. Macdonald, J. Christopher, I.D. Welch, S.M. Rubin, M.R.W.
750 Mann, F.A. Dick, An RB-EZH2 Complex Mediates Silencing of Repetitive DNA
751 Sequences, 64 (2017) 1074–1087.
752 <https://doi.org/10.1016/j.molcel.2016.10.021>.

753 [47] P. V. Angadi, J.K. Savitha, S.S. Rao, R.Y. Sivaranjini, Oral field cancerization:
754 Current evidence and future perspectives, *Oral Maxillofac. Surg.* 16 (2012)
755 171–180. <https://doi.org/10.1007/S10006-012-0317-X/TABLES/1>.

756 [48] I. Martincorena, A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D.C.
757 Wedge, A. Fullam, L.B. Alexandrov, J.M. Tubio, L. Stebbings, A. Menzies, S.
758 Widaa, M.R. Stratton, P.H. Jones, P.J. Campbell, High burden and pervasive
759 positive selection of somatic mutations in normal human skin, *Science*. 348

(2015) 880–886.

https://doi.org/10.1126/SCIENCE.AAA6806/SUPPL_FILE/AAA6806-MARTINCORENA-SM.PDF.

[49] K. Yoshida, K.H.C. Gowers, H. Lee-Six, D.P. Chandrasekharan, T. Coorens, E.F. Maughan, K. Beal, A. Menzies, F.R. Millar, E. Anderson, S.E. Clarke, A. Pennycuick, R.M. Thakrar, C.R. Butler, N. Kakiuchi, T. Hirano, R.E. Hynds, M.R. Stratton, I. Martincorena, S.M. Janes, P.J. Campbell, Tobacco smoking and somatic mutations in human bronchial epithelium, *Nat.* 2020 5787794 578 (2020) 266–272. <https://doi.org/10.1038/s41586-020-1961-1>.

[50] T.G. Paulson, P.C. Galipeau, K.M. Oman, C.A. Sanchez, M.K. Kuhner, L.P. Smith, K. Hadi, M. Shah, K. Arora, J. Shelton, M. Johnson, A. Corvelo, C.C. Maley, X. Yao, R. Sanghvi, E. Venturini, A.K. Emde, B. Hubert, M. Imielinski, N. Robine, B.J. Reid, X. Li, Somatic whole genome dynamics of precancer in Barrett’s esophagus reveals features associated with disease progression, *Nat. Commun.* 13 (2022). <https://doi.org/10.1038/s41467-022-29767-7>.

[51] I. Martincorena, J.C. Fowler, A. Wabik, A.R.J. Lawson, F. Abascal, M.W.J. Hall, A. Cagan, K. Murai, K. Mahbubani, M.R. Stratton, R.C. Fitzgerald, P.A. Handford, P.J. Campbell, K. Saeb-Parsy, P.H. Jones, Somatic mutant clones colonize the human esophagus with age, *Science*. 362 (2018) 911–917. <https://doi.org/10.1126/science.aau3879>.

[52] A.R.J. Lawson, F. Abascal, T.H.H. Coorens, Y. Hooks, L. O’Neill, C. Latimer, K. Raine, M.A. Sanders, A.Y. Warren, K.T.A. Mahbubani, B. Bareham, T.M. Butler, L.M.R. Harvey, A. Cagan, A. Menzies, L. Moore, A.J. Colquhoun, W. Turner, B. Thomas, V. Gnanapragasam, N. Williams, D.M. Rassl, H. Vöhringer, S. Zumalave, J. Nangalia, J.M.C. Tubío, M. Gerstung, K. Saeb-Parsy, M.R. Stratton,

785 P.J. Campbell, T.J. Mitchell, I. Martincorena, Extensive heterogeneity in
786 somatic mutation and selection in the human bladder, *Science*. 370 (2020)
787 75–82. <https://doi.org/10.1126/science.aba8347>.

788 [53] C.H. Nam, J. Youk, J.Y. Kim, J. Lim, J.W. Park, S.A. Oh, H.J. Lee, J.W. Park, H.
789 Won, Y. Lee, S.Y. Jeong, D.S. Lee, J.W. Oh, J. Han, J. Lee, H.W. Kwon, M.J. Kim,
790 Y.S. Ju, Widespread somatic L1 retrotransposition in normal colorectal
791 epithelium, *Nature* 617 (2023) 540–547. [https://doi.org/10.1038/s41586-](https://doi.org/10.1038/s41586-023-06046-z)
792 023-06046-z.

793 [54] M. Shademan, K. Zare, M. Zahedi, H. Mosannen Mozaffari, H. Bagheri
794 Hosseini, K. Ghaffarzadegan, L. Goshayeshi, H. Dehghani, Promoter
795 methylation, transcription, and retrotransposition of LINE-1 in colorectal
796 adenomas and adenocarcinomas, *Cancer Cell Int.* 20 (2020) 1–16.
797 <https://doi.org/10.1186/S12935-020-01511-5>.

798

799

800 **FIGURE LEGENDS**

801 **Figure 1. RetroTest design and benchmarking.** **A.** RetroTest design. **B.**
 802 Scatterplot and correlation between L1 activation and orphan transductions (TD2)
 803 in the International Consortium of PanCancer (PCAWG) data. The number of L1 3'
 804 transductions measured L1 activation. **C.** Performance of RetroTest for different
 805 sequencing coverages using an artificially generated GRC37/hg19 genome with a
 806 total of 2,480 randomly distributed L1 transductions at 50% VAF. **D.** Performance
 807 of RetroTest with respect to the VAF of L1 integrations, using the artificially
 808 generated GRC37/hg19 genome with a total of 2,480 randomly distributed L1
 809 transductions at different VAFs. **E.** Performance of TraFiC with respect of the VAF
 810 of L1 integrations, using the artificially generated GRC37/hg19 genome with a
 811 total of 2,480 randomly distributed L1 transductions at different VAFs. **F.** Venn
 812 diagram of the number of L1 insertions detected by RetroTest and TraFiC in the
 813 artificial genome with a VAF of 50% for L1 insertions.

814

815 **Figure 2. L1 activation measured by RetroTest in the HNSCC cohort (n=96).** **A.**
 816 Quantification of L1 activation in HNSCC tumors as the number of orphan
 817 transduction detected by RetroTest. **B.** Boxplot of L1 activation with respect to
 818 early (T1-T2) and advanced (T3-T4) TNM stages. Differential activation p-value
 819 was derived by the Wilcoxon test. To correct coverage-related bias, L1 activation
 820 was calculated as the number of TD2 divided by its median coverage. **C.** Boxplot of
 821 L1 activation with respect to smoking status. Differential activation p-value was
 822 derived by the Wilcoxon test. To correct coverage-related bias, L1 activation was
 823 calculated as the number of TD2 divided by its median coverage. **D.** Kaplan-Meier
 824 curves for overall survival with respect to L1 activation rate at T1 stage. Patients

were grouped into high (above the median) and low (below the median) L1 activation rate. Log-rank test was used to calculate the p-value. To correct coverage-related bias, L1 activation was calculated as the number of TD2 divided by its median coverage. **E.** Oncoplot showing the genes affected by L1 insertions and their original source elements. A total of 27 patients presented genes affected by L1 insertions from 46 different source elements.

831

Figure 3. Characterization of the mutational profile of HNSCC patients by WGS (n=19). **A.** Oncoplot showing the genes harboring somatic mutations in HNSCC patients. The type of mutation in each gene and the L1 activation (number of L1 insertions in each patient) is shown. **B.** Boxplot of L1 activation with respect to TP53 mutation or wild-type status. To correct coverage-related bias, L1 activation was calculated as the number of TD2 divided by its median coverage. Differential activation p-value was derived by the Wilcoxon test. **C.** Barplot of the Pathway enrichment analysis based on the 918 genes somatically mutated genes. Enrichment p-values were calculated with the Fisher exact test. **D.** Barplot of the transcription factor binding enrichment analysis based on the 918 genes somatically mutated genes. Enrichment p-values were calculated with the Fisher exact test.

844

Figure 4. Evaluation of L1 activation in normal adjacent tissue. **A.** Schematic representation of the evaluation of the field cancerization process. **B.** Number of L1 active elements in normal adjacent tissue of HNSCC patients (n=9), compared to L1 activation in paired tumor tissue and PBMCs as germline control from the same patients.

850 **Table 1. Baseline characteristics of the HNSCC patients and**
851 **clinicopathological results in the series.**

N = 96		
Mean age (range)		
67.4 (38 – 90)		
Sex		
	Female	20
	Male	71
	NA	5
TNM stage		
	T1	16
	T2	19
	T3	18
	T4	34
	NA	9
Exitus		
	No	50
	Yes	38
	NA	8
Alcohol		
	Drinker	61
	Non-drinker	24
	NA	11
Smoking habits		
	Smoker	64

	Non-smoker	22
	NA	10
852		
853		
854		
855		
856		

857 **Table 2. Number of HNSCC patients showing L1 activation along TNM stages.**

	TNM stage			
	T1	T2	T3	T4
L1 active	10	12	17	25
L1 inactive	6	7	1	9

858

859 **ADDITIONAL FILES**

860 **Additional file 1 (pdf): Supplementary Methods:** RetroTest library and target
861 sequencing; RetroTest method in detail; RetroTest benchmarking; Whole Genome
862 Sequencing and determination of mutation profile; Gene pathway databases uses
863 for enrichment analysis

864 **Additional file 2 (xls): Table S1:** RetroTest benchmarking. **Table S2:** HNSCC
865 somatic variants. **Table S3:** Gene pathways affected by HNSCC somatic variants.
866 **Table S4:** Transcription factors binding to genes affected by HNSCC somatic
867 variants. **Table S5:** HNSCC NAT somatic variants

868 **Additional file 3 (pdf). Figure S1.** Kaplan-Meier curves for overall survival in
869 HNSCC cohort with respect to (A) L1 activation status (active vs inactive) and (B)
870 L1 activation rate (with respect to the median, being high above vs low as below
871 the median). Log-rank test was used to calculate the p-value.

872

873

874

875

876

877

878

879

880

881

882

883

884 **FIGURES**

885 **Figure 1.**

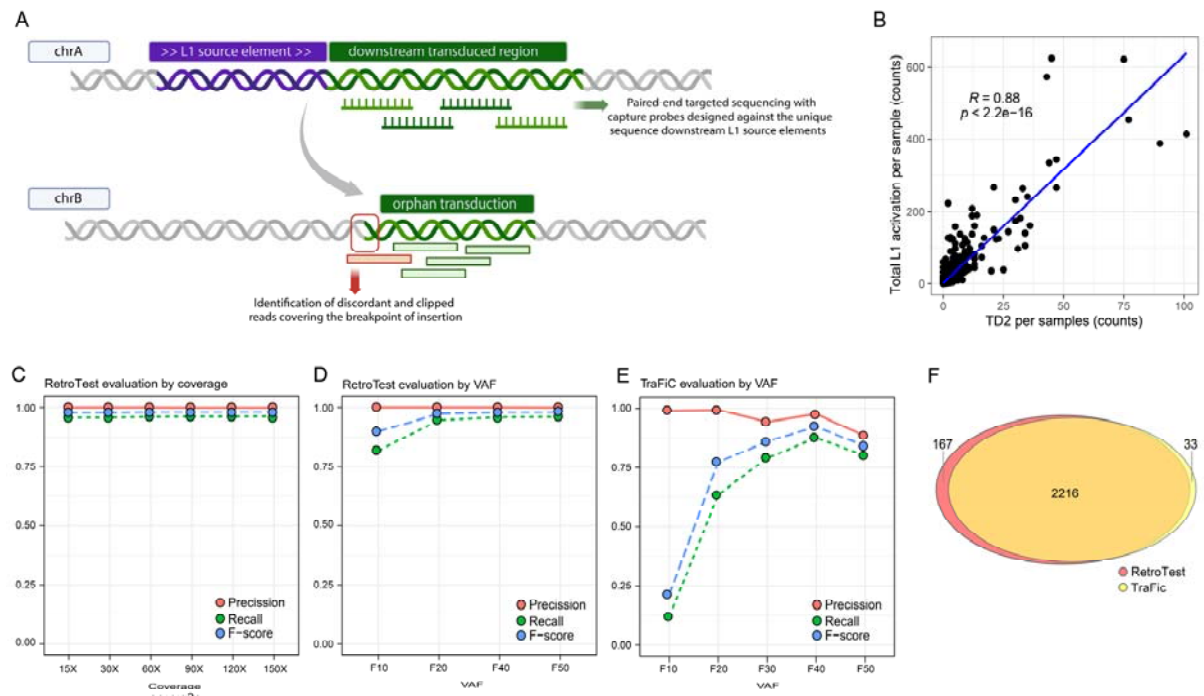
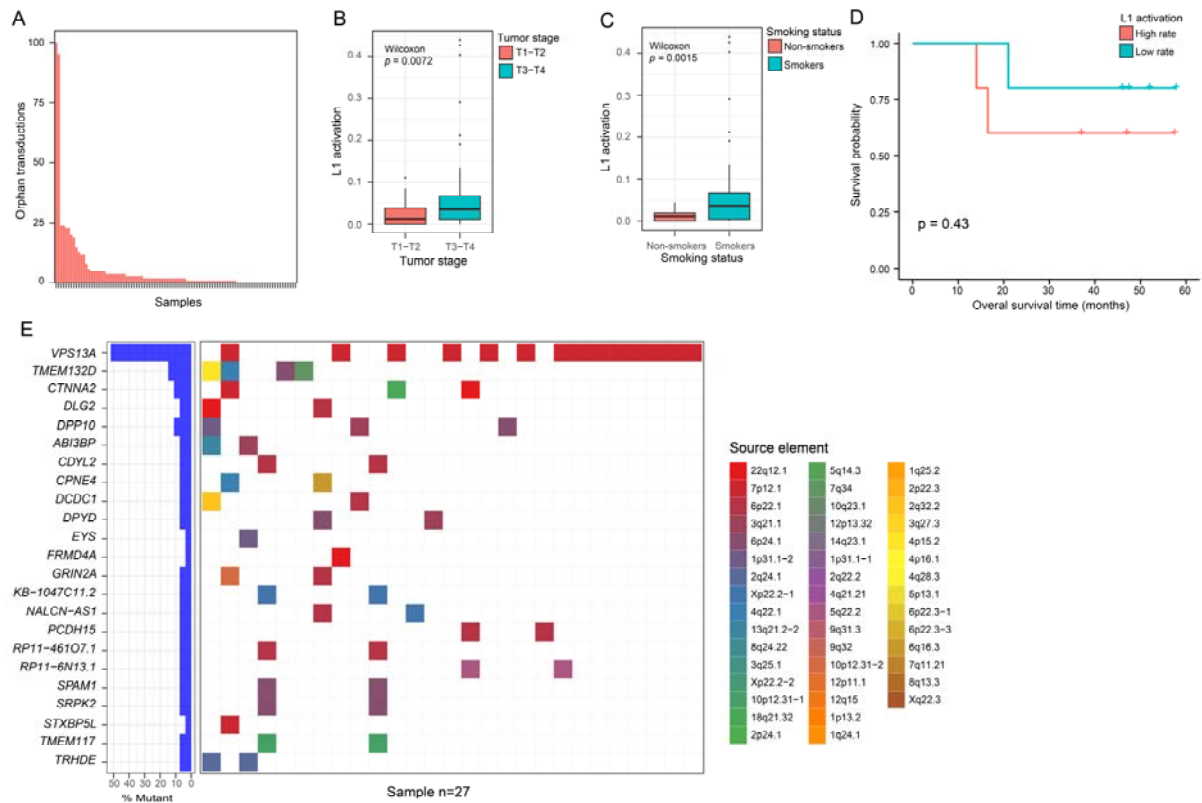
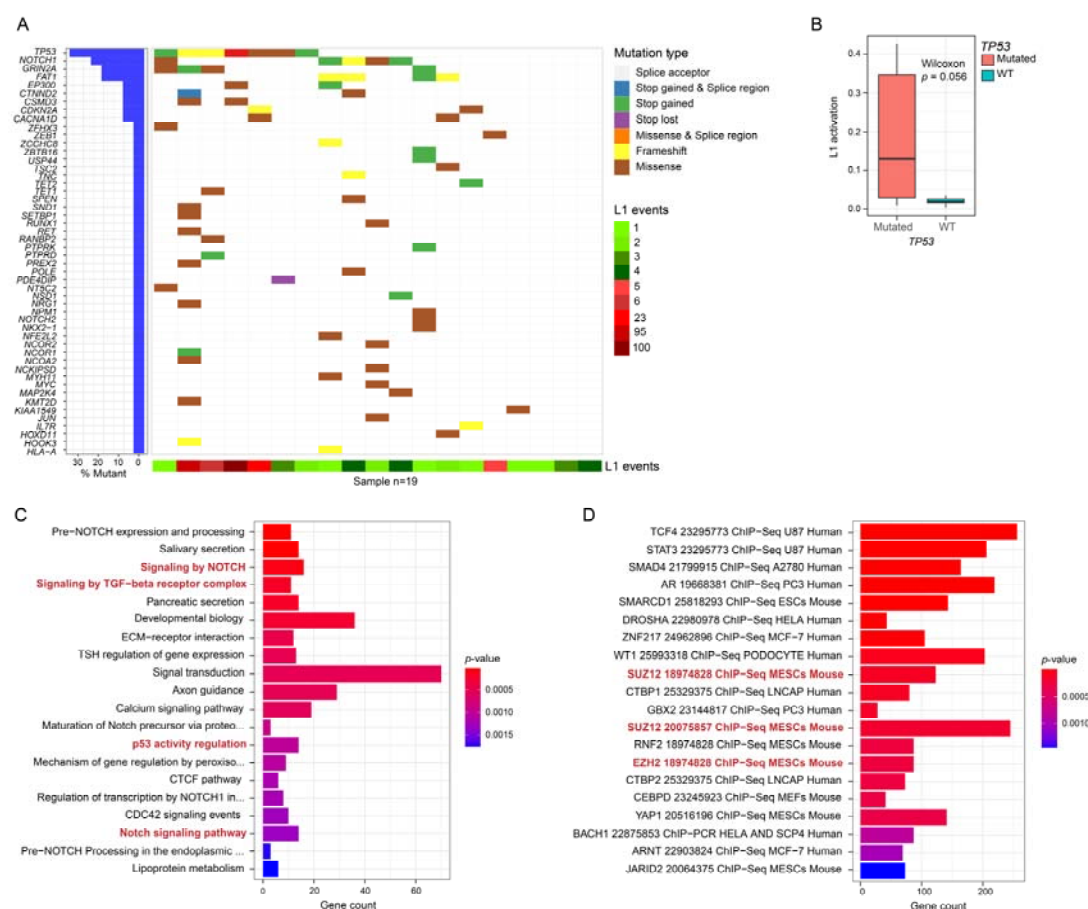


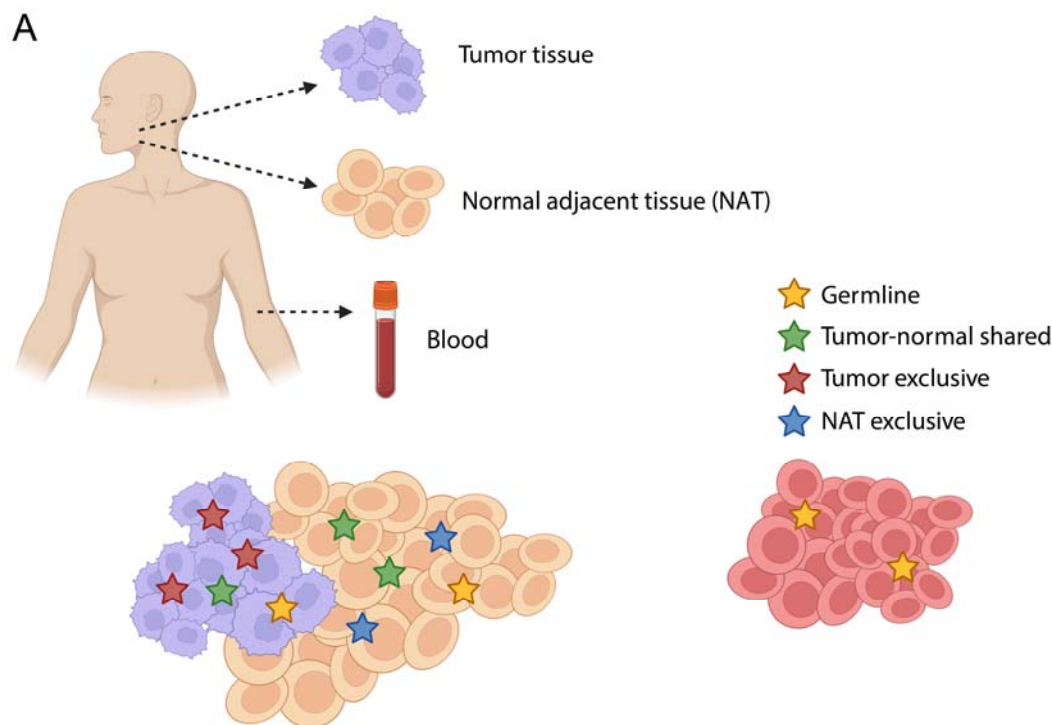
Figure 2.



914 **Figure 3.**



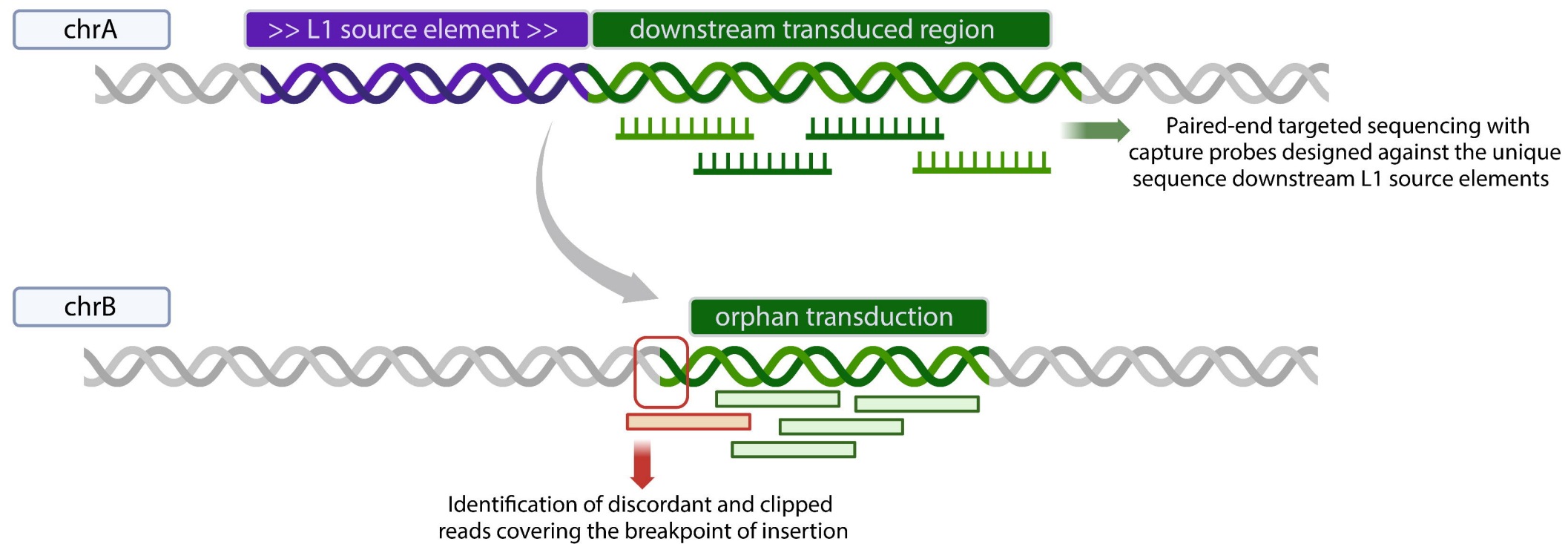
926 **Figure 4.**



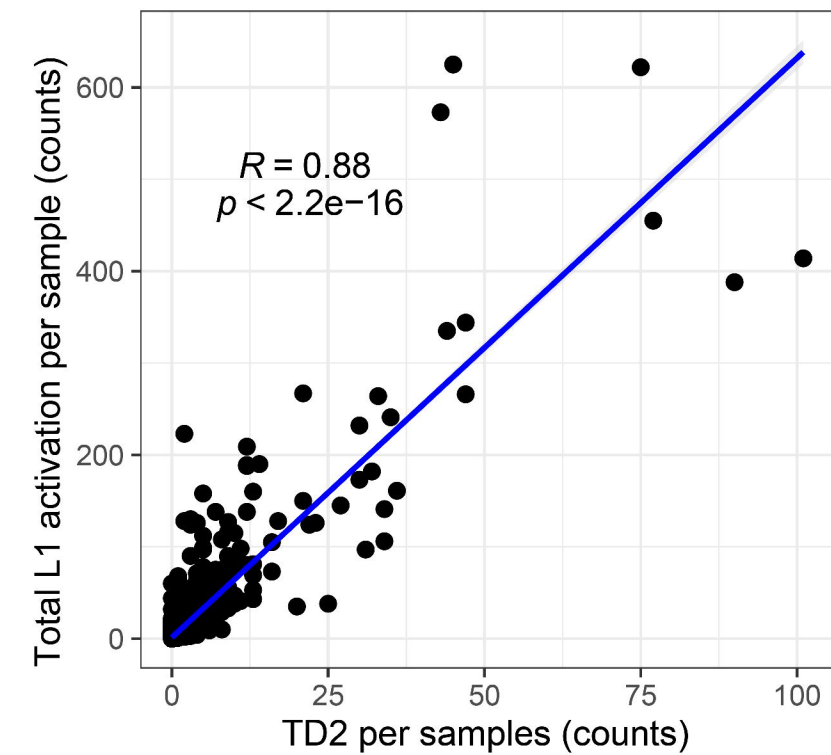
B

Patient	NAT exclusive	Tumor exclusive	Tumor-normal shared	Germline	Total
HNSCC_01	0	20	0	1	21
HNSCC_02	1	95	1	0	97
HNSCC_03	1	24	0	0	25
HNSCC_04	0	100	0	0	100
HNSCC_05	0	4	0	0	4
HNSCC_06	2	1	0	0	3
HNSCC_07	0	23	0	1	24
HNSCC_08	0	3	0	2	5
HNSCC_09	0	2	0	1	3

A

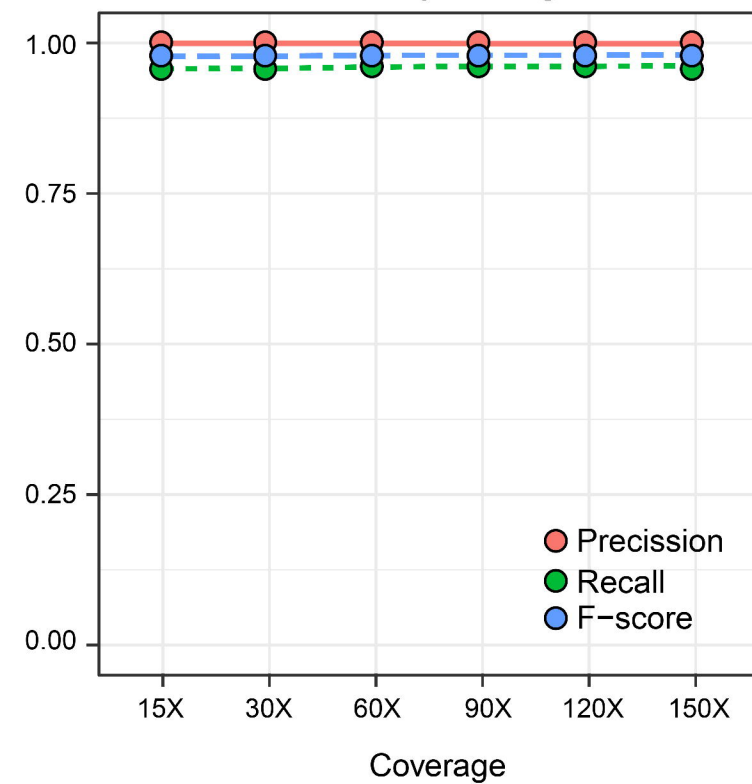


B



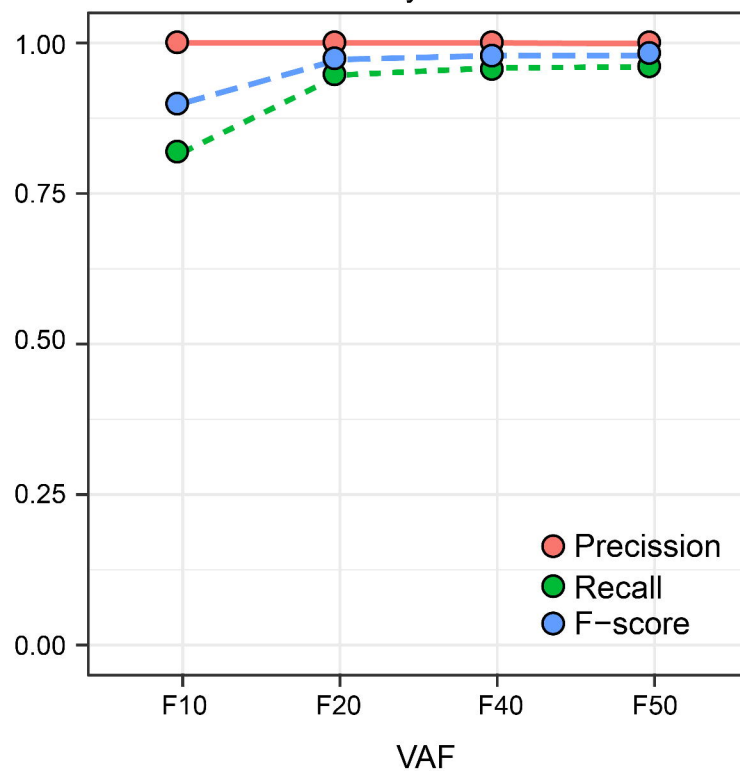
C

RetroTest evaluation by coverage



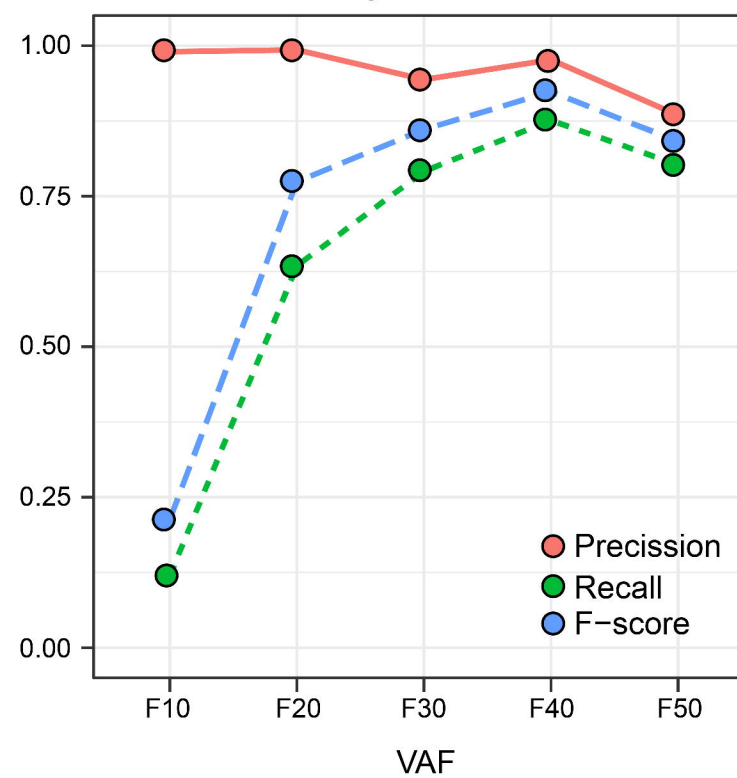
D

RetroTest evaluation by VAF

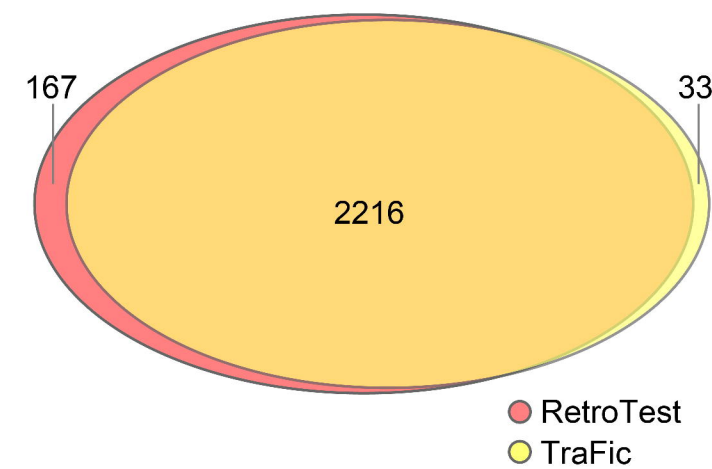


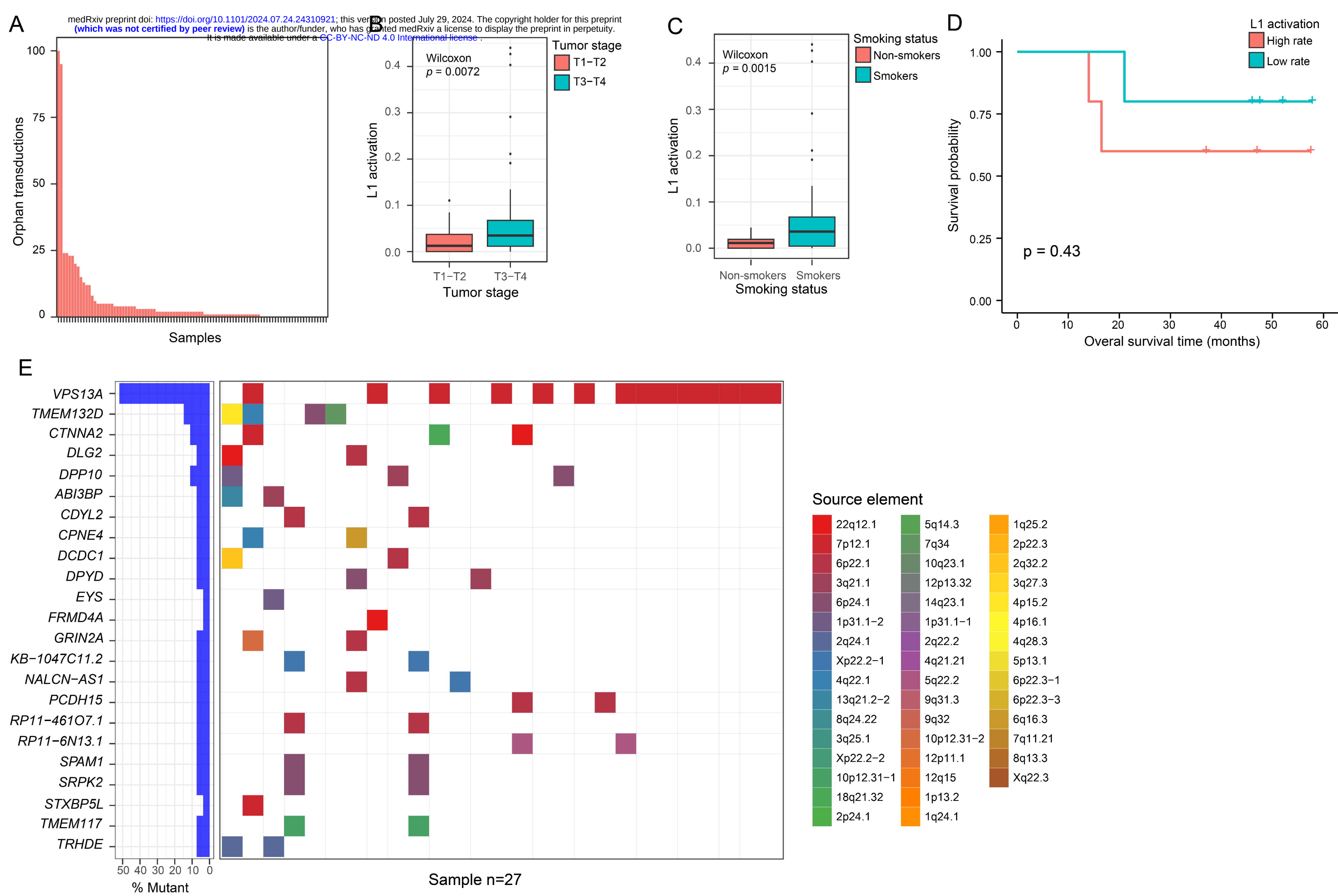
E

TraFiC evaluation by VAF

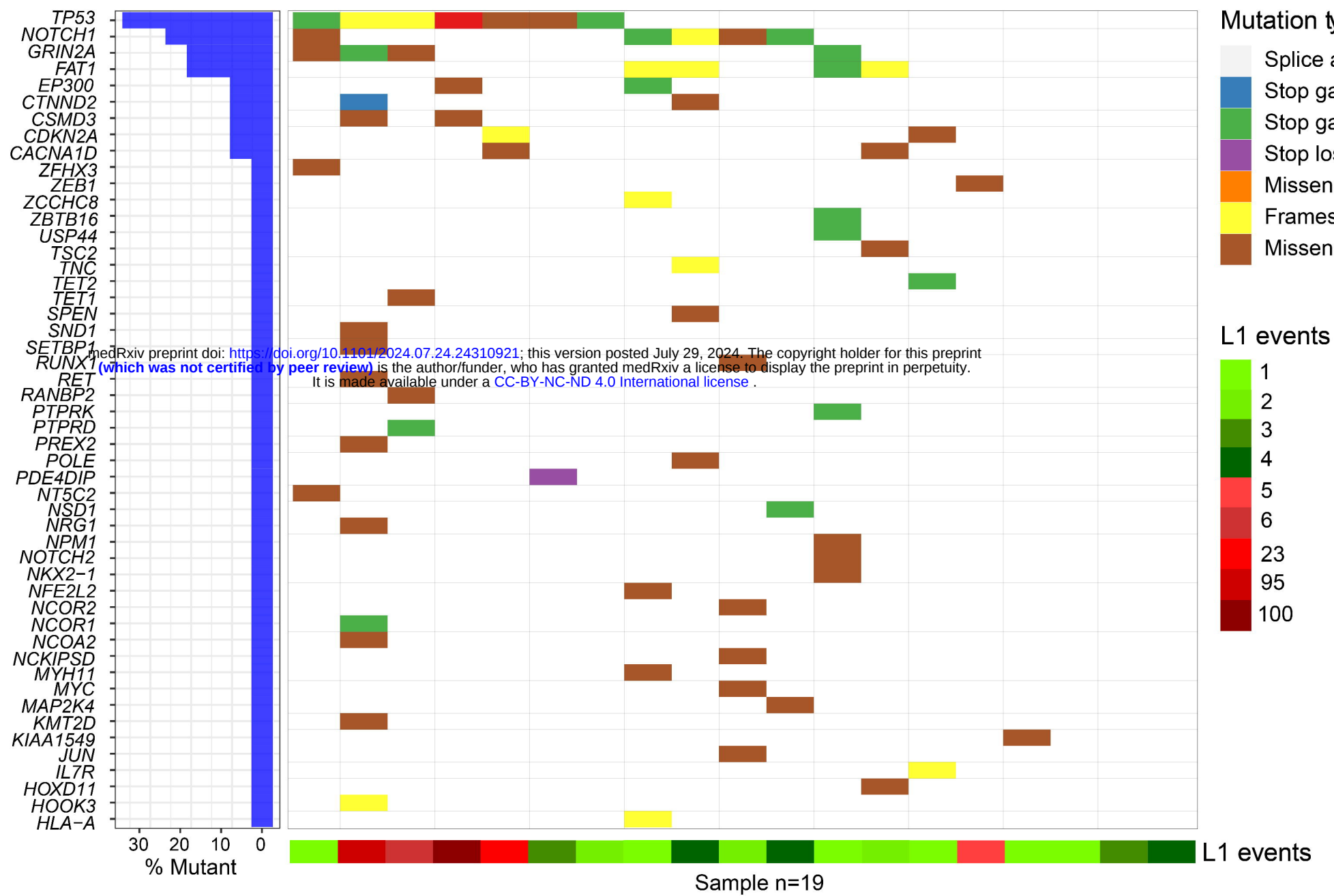


F

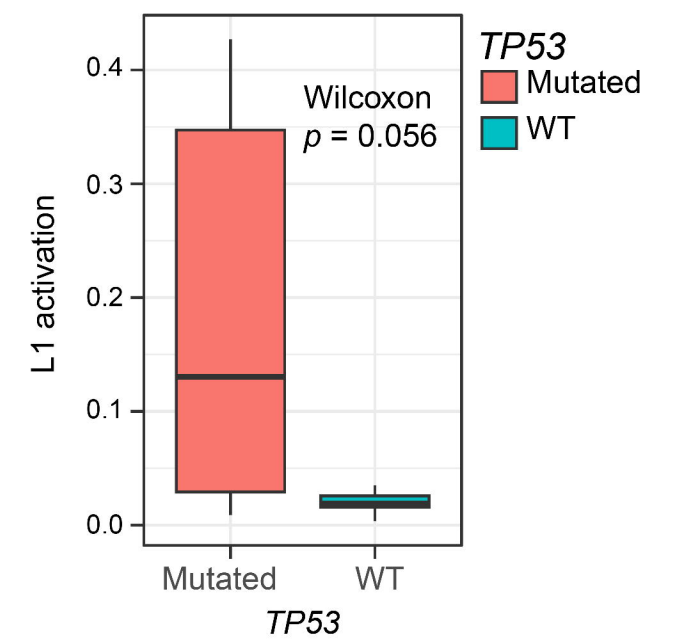




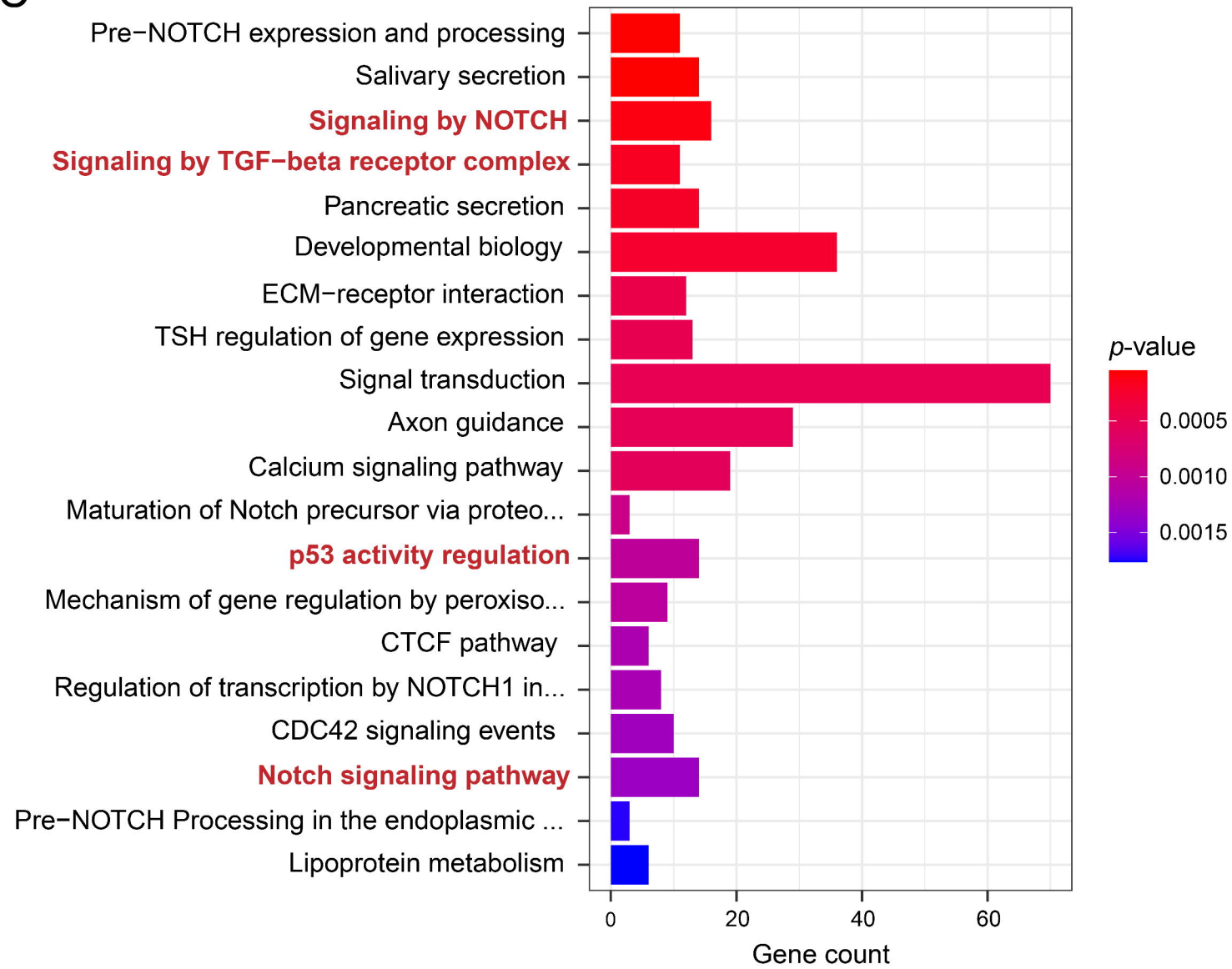
A



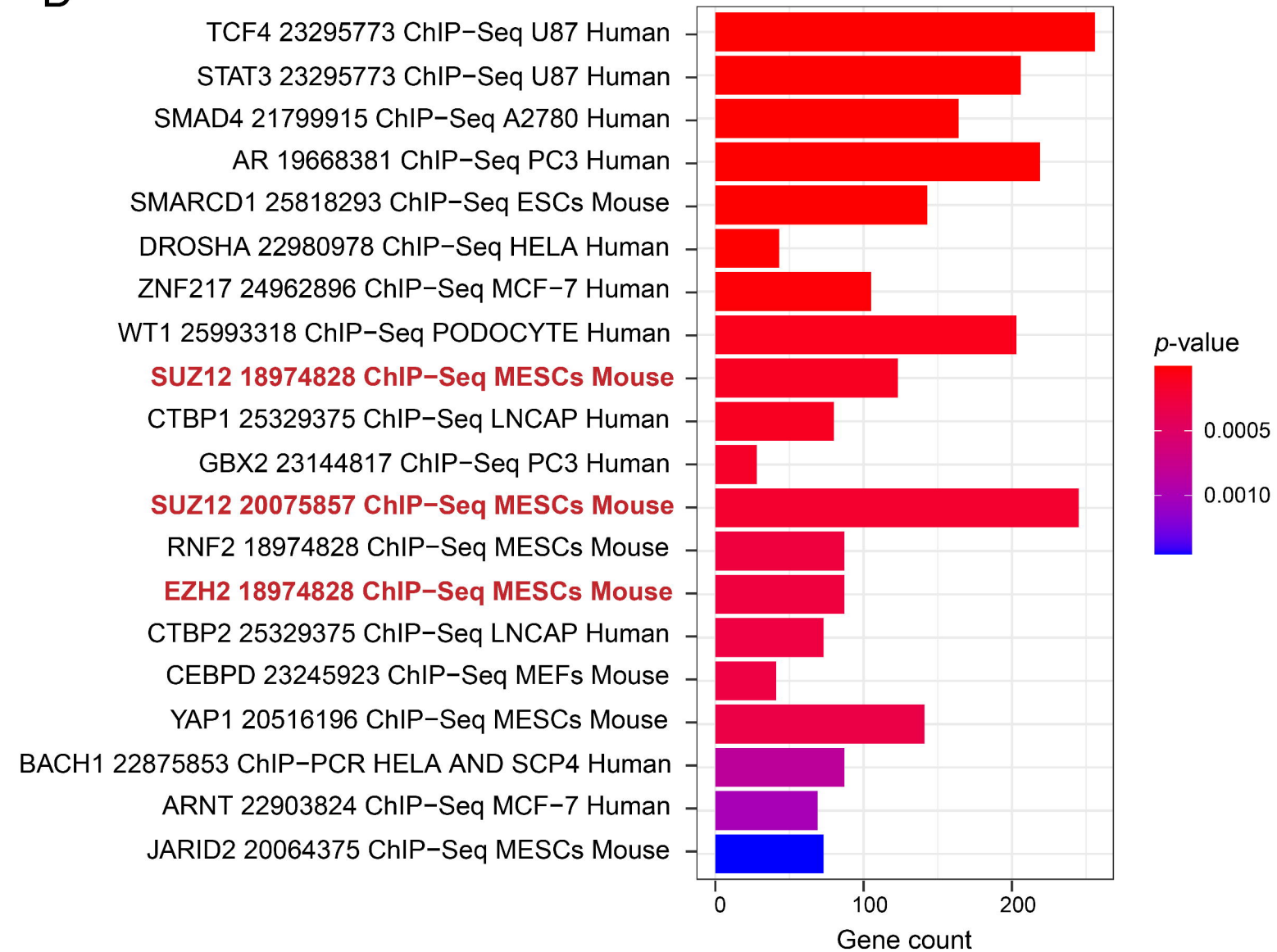
B



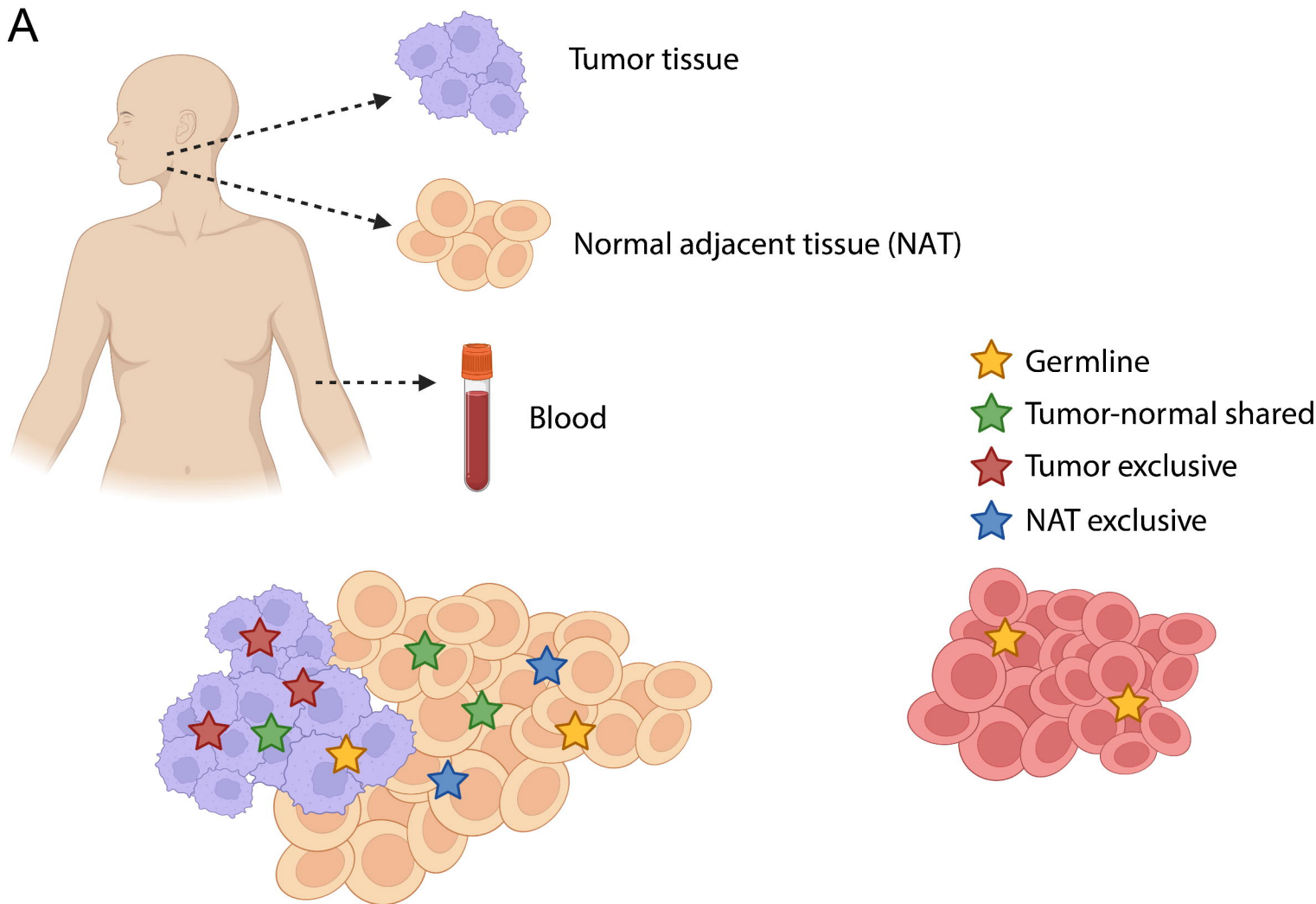
C



D



A



B

Patient	NAT exclusive	Tumor exclusive	Tumor-normal shared	Germline	Total
HNSCC_01	0	20	0	1	21
HNSCC_02	1	95	1	0	97
HNSCC_03	1	24	0	0	25
HNSCC_04	0	100	0	0	100
HNSCC_05	0	4	0	0	4
HNSCC_06	2	1	0	0	3
HNSCC_07	0	23	0	1	24
HNSCC_08	0	3	0	2	5
HNSCC_09	0	2	0	1	3