








ARTICLE

<https://doi.org/10.1038/s41467-019-11770-0>

OPEN

# Retrotransposon insertions can initiate colorectal cancer and are associated with poor survival

Tatiana Cajuso<sup>1,2</sup>, Päivi Sulo <sup>1,2</sup>, Tomas Tanskanen<sup>1,2</sup>, Riku Katainen<sup>1,2</sup>, Aurora Taira<sup>1,2</sup>, Ulrika A. Hänninen <sup>1,2</sup>, Johanna Kondelin<sup>1,2</sup>, Linda Forsström<sup>1,2</sup>, Niko Välimäki<sup>1,2</sup>, Mervi Aavikko <sup>1,2</sup>, Eevi Kaasinen<sup>1,2</sup>, Ari Ristimäki <sup>1,3</sup>, Selja Koskensalo <sup>4</sup>, Anna Lepistö<sup>4</sup>, Laura Renkonen-Sinisalo<sup>4</sup>, Toni Seppälä <sup>4</sup>, Teijo Kuopio<sup>5,6</sup>, Jan Böhm<sup>6</sup>, Jukka-Pekka Mecklin<sup>7,8</sup>, Outi Kilpivaara<sup>1,2</sup>, Esa Pitkänen<sup>1,2</sup>, Kimmo Palin <sup>1,2</sup> & Lauri A. Aaltonen<sup>1,2</sup>

Genomic instability pathways in colorectal cancer (CRC) have been extensively studied, but the role of retrotransposition in colorectal carcinogenesis remains poorly understood. Although retrotransposons are usually repressed, they become active in several human cancers, in particular those of the gastrointestinal tract. Here we characterize retrotransposon insertions in 202 colorectal tumor whole genomes and investigate their associations with molecular and clinical characteristics. We find highly variable retrotransposon activity among tumors and identify recurrent insertions in 15 known cancer genes. In approximately 1% of the cases we identify insertions in *APC*, likely to be tumor-initiating events. Insertions are positively associated with the CpG island methylator phenotype and the genomic fraction of allelic imbalance. Clinically, high number of insertions is independently associated with poor disease-specific survival.

<sup>1</sup> Applied Tumor Genomics Research Program, Faculty of Medicine University of Helsinki, Biomedicum Helsinki, PO Box 63 (Haartmaninkatu 8), FI-00014 Helsinki, Finland. <sup>2</sup> Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Biomedicum Helsinki, PO Box 63 (Haartmaninkatu 8), FI-00014 Helsinki, Finland. <sup>3</sup> Department of Pathology, HUSLAB, University of Helsinki and Helsinki University Hospital, (Haartmaninkatu 3), FI-00290 Helsinki, Finland. <sup>4</sup> Department of Gastrointestinal Surgery, Helsinki University Hospital, University of Helsinki, (Haartmaninkatu 4), FI-00290 Helsinki, Finland. <sup>5</sup> Biological and Environmental Science, University of Jyväskylä, PO Box 35 (Seminaarinkatu 15), FI-40014 Jyväskylä, Finland. <sup>6</sup> Department of Pathology, Central Finland Health Care District, (Keskussairaalantie 19), FI-40620 Jyväskylä, Finland. <sup>7</sup> Department of Surgery, Jyväskylä Central Hospital, (Keskussairaalantie 19), FI-40620 Jyväskylä, Finland. <sup>8</sup> Department of Health Sciences, Faculty of Sport and Health Sciences, University of Jyväskylä, PO Box 35 (Seminaarinkatu 15), FI-40014 Jyväskylä, Finland. Correspondence and requests for materials should be addressed to L.A.A. (email: [lauri.aaltonen@helsinki.fi](mailto:lauri.aaltonen@helsinki.fi))

Retrotransposons are transposable genetic sequences that copy themselves into an RNA intermediate and insert elsewhere in the genome. Almost half of the human genome consists of transposon derived sequences<sup>1</sup>, however only a few elements remain retrotransposition competent and account for most retrotranspositions<sup>2,3</sup>. Two types of retrotransposons have been identified in the human genome; autonomous and non-autonomous. Autonomous elements, such as Long Interspersed Nuclear Element-1s (LINE-1s) and Endogenous retroviruses (ERVs), provide the required machinery for retrotransposition. On the contrary, non-autonomous elements, such as Alus and SINE-VNTR-Alu (SVAs), require the LINE-1 machinery to retrotranspose<sup>4–7</sup>. In cancer, ~24% of somatic retrotranspositions involve 3' transduction, a process characterized by mobilization of 3' flanking sequence which can serve as a unique sequence revealing the insertion origin<sup>8–11</sup>.

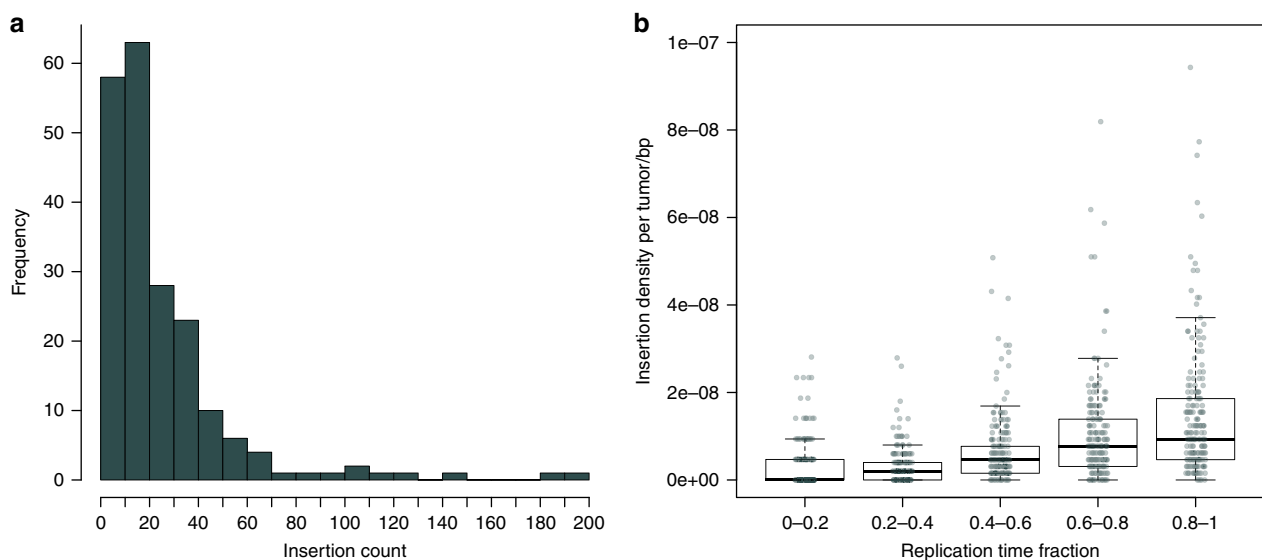
LINE-1s are frequently repressed by promoter methylation<sup>12</sup> and genome-wide hypomethylation is reported to lead to their activation during tumorigenesis<sup>13,14</sup>, thus leading to high retrotransposon activity and genome instability<sup>14–16</sup>. High retrotransposon activity has been reported in several human cancers, especially in tumors arising from the gastrointestinal tract, such as colorectal cancer (CRC)<sup>10,11,17–20</sup>. Somatic insertion density in tumors is higher in closed chromatin and late replicating regions. Among insertions in genes, insertion density is higher in genes with low expression<sup>10,21</sup>. Furthermore, ongoing retrotransposon activity has been reported in CRC<sup>22</sup>. Insertion count is associated with patient age<sup>18</sup> and LINE-1 hypomethylation is associated with poor survival in CRC<sup>23</sup>. LINE-1 insertions in *APC* have been reported in two CRCs, indicating that these insertions may be early tumorigenic events<sup>24,25</sup>. CRC can develop through two distinct pathways; chromosomal instability (CIN) or microsatellite instability (MSI). Most sporadic CRCs follow the CIN pathway, characterized by a large number of chromosomal alterations. Fifteen percent of CRC cases follow the MSI pathway, characterized by a high number of base substitutions and short insertions and deletions<sup>26</sup>. Seventy-five percent of MSI-positive sporadic CRCs are attributed to the CpG island methylator phenotype (CIMP)<sup>27</sup> which is characterized by gene promoter hypermethylation. Although genomic instability pathways have

been studied extensively in CRC, the tumorigenic role of retrotransposition is not fully understood. Retrotransposon insertions have been difficult to detect with previous methodological approaches and very few genome-wide studies have been reported. Here, we characterize somatic retrotransposon insertions in 201 CRCs and one colorectal adenoma utilizing whole genome sequencing (WGS), and investigate the associations between somatic retrotransposon activity and clinical characteristics.

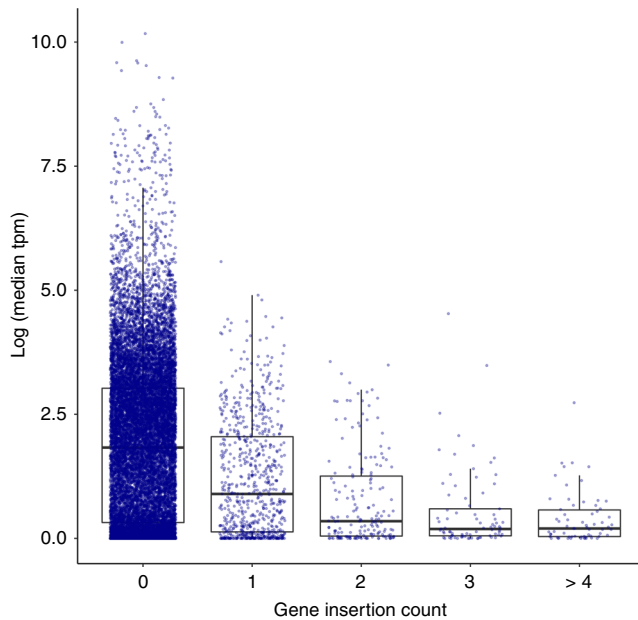
## Results

### Genome-wide detection of somatic retrotransposition in CRC.

To characterize the landscape of somatic retrotransposon insertions in CRC we applied TraFiC<sup>10</sup> and DELLY<sup>28</sup> to WGS data from 202 colorectal tumors and matched normal samples. TraFiC was used to detect insertions without 3' flanking sequence and DELLY was used to detect LINE-1 transductions that were not identifiable by TraFiC. From the 202 tumors, 12 were MSI and 190 were microsatellite stable (MSS) including three ultra-mutated tumors, harboring somatic *POLE* mutations. After strict somatic filtering, we identified a total of 5072 insertions (Supplementary Data 1). We detected 4726 insertions with TraFiC, and 346 transduction calls with DELLY. Based on visual inspection of the paired-end read data on 100 random insertion calls, 76 calls were evaluated as true somatic insertions, giving a false positive rate of 24% (95% confidence interval [CI], 16–34%) (Supplementary Data 1). Additionally, 14 out of 15 3' transductions from two samples were validated by long-distance inverse-PCR (LDI-PCR) and Nanopore sequencing in a separate study<sup>22</sup>. The mean number of insertions per tumor was 25 (median, 17; interquartile range, 10–31) with high variability among tumors (Fig. 1a). Mean number of insertions in MSS, MSI and the *POLE* ultra-mutated tumors was 25, 34, and 24 respectively. The majority of insertions (99%, 5024/5072) were LINE-1 retrotranspositions, however we also detected 20 SVA, 13 Alu and 15 ERV insertions (Supplementary Data 1). In concordance with previous studies<sup>10,21</sup>, insertion density was higher in closed chromatin (1.78 insertions per Mbp) than in open chromatin (0.96 insertions per Mbp) and in late replicating regions (replication time > 0.8, 3.06 insertions per Mbp) than in early



**Fig. 1** Distribution of somatic insertions across 202 colorectal tumors and over replication time. **a** Frequency of somatic insertion counts in 202 colorectal tumors. **b** Insertion density over replication time. The genome was stratified by replication time in five categories where 0 referred to the earliest replication timing. Each point represents insertion density in the corresponding category for each of the 202 tumors. Boxplot shows median, interquartile range (IQR), and whiskers extend to the most extreme data points which are no more than 1.5 times the IQR



**Fig. 2** Retrotransposon insertions in protein-coding genes. Gene expression (median TPM values from 34 tumors) over gene insertion count groups. Boxplot shows median, interquartile range (IQR), and whiskers extend to the most extreme data points which are no more than 1.5 times the IQR

**Table 1** Genes from the Cancer Gene Census with two or more insertions

Gene ID	Gene name	Number of insertions (n = 202)	Cancer census role
ENSG00000168702	<i>LRP1B</i>	19	TSG
ENSG00000178568	<i>ERBB4</i>	7	Oncogene, TSG
ENSG00000171094	<i>ALK</i>	5	Oncogene, fusion
ENSG00000196090	<i>PTPR</i>	3	TSG
ENSG00000046889	<i>PREX2</i>	3	Oncogene
ENSG00000185811	<i>IKZF1</i>	3	TSG, fusion
ENSG00000183454	<i>GRIN2A</i>	2	TSG
ENSG00000144218	<i>AFF3</i>	2	Oncogene, fusion
ENSG00000157168	<i>NRG1</i>	2	TSG, fusion
ENSG00000079102	<i>RUNX1T1</i>	2	Oncogene, TSG, fusion
ENSG00000151702	<i>FLI1</i>	2	Oncogene, fusion
ENSG00000134982	<i>APC</i>	2	TSG
ENSG00000189283	<i>FHIT</i>	2	TSG, fusion
ENSG00000085276	<i>MECOM</i>	2	Oncogene, fusion
ENSG00000196159	<i>FAT4</i>	2	TSG

Gene names are shown in italics. Cancer census role, role in cancer as defined by the Cancer Gene Census 30  
TSG tumor suppressor gene

replicating regions (replication time < 0.2, 0.73 insertions per Mbp) (Fig. 1b).

**Retrotransposons are predicted to initiate ~1% of CRCs.** To characterize retrotransposon insertions in genes, all protein-coding transcripts and the insertion polyA/T in conjunction with gene orientation were used to assess insertion orientation (sense/antisense). Of the 5072 insertions, 1680 (33%) were detected within protein-coding genes, with 98% in introns (Supplementary Data 1, Supplementary Fig. 1). We identified 353 insertions in antisense orientation and 349 in sense orientation (Supplementary Data 1). Insertion count was higher in genes with lower

expression (median transcript per million reads [TPM] from 34 tumors) in concordance with a previous study<sup>21</sup> (Fig. 2).

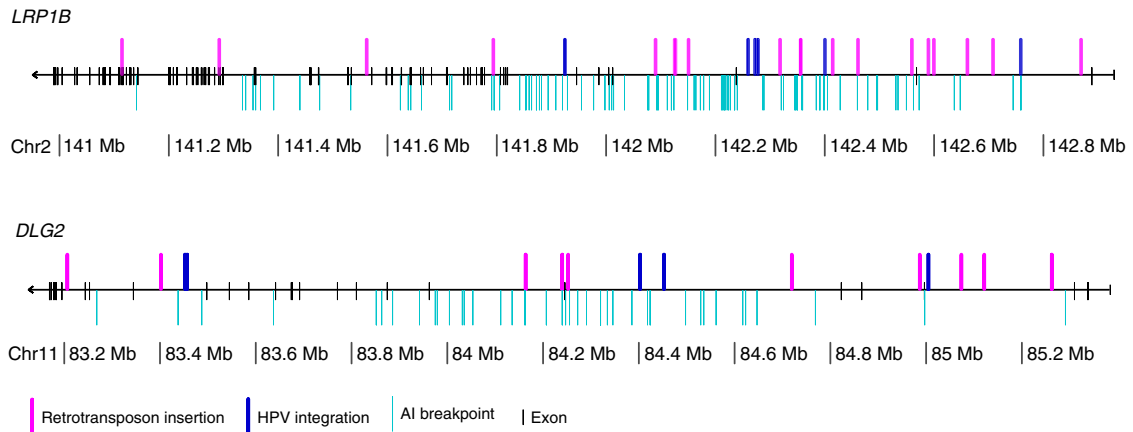
Recurrent insertions (at least two insertions) were identified in 333 protein-coding genes (Supplementary Data 2) and no significant enrichment of biological processes was observed after correcting for gene length. Fifteen genes in the Cancer Gene Census (CGC)<sup>29</sup> displayed recurrent insertions and no clear bias towards tumor suppressor genes or oncogenes was apparent (Table 1).

The most frequently affected protein-coding genes were *LRP1B* with 19 insertions, *DLG2* with 10 insertions and *PTPRD* and *LSAMP* both with 9 insertions. All the insertions were located in the introns and no insertion clusters were observed (Supplementary Data 1). Higher number of insertions in antisense orientation was observed in *LRP1B* where insertion orientation was available for more insertions (Supplementary Data 1). Both *LRP1B* and *DLG2* have been reported to be fragile sites<sup>30</sup> and recurrent hotspots for HPV integration<sup>31</sup> (Fig. 3). However, no clusters of insertions and HPV integrations nor allelic imbalances (AI) were apparent (Fig. 3).

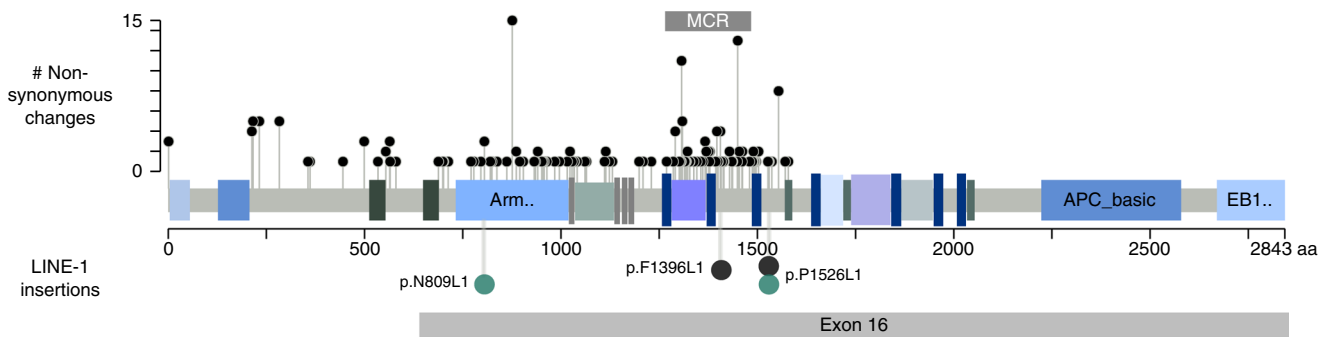
Genes with highest density of recurrent insertions were *RCN1* with three insertions, and *COL25A1*, *ARAP2*, and *ZNF251* with two insertions. Gene Ontology (GO) annotations associated to these genes include protein binding for *RCN1*; heparin binding and amyloid-beta binding for *COL25A1*; GTPase activator activity, small GTPase binding and phosphatidylinositol-3,4,5-trisphosphate binding for *ARAP2*; and RNA polymerase II transcription activity and DNA binding transcription factor activity for *ZNF251*<sup>32</sup>. None of these genes have been classified as cancer genes<sup>29</sup>, fragile sites<sup>30</sup> or hotspots of HPV integration<sup>31</sup>. We also investigated whether insertions had an overall effect on the expression of the closest genes but no significant effect was detected (Supplementary Fig. 2, Methods section Association test between insertions and RNA expression).

Seventy-two insertions were identified in exons of protein-coding genes (Supplementary Data 1, Supplementary Fig. 1). We identified one insertion in the last exon/3' UTR of *PIK3CA* (Supplementary Data 1) and two insertions in exon 16 of *APC* (Fig. 4, Supplementary Data 1). Loss of heterozygosity and copy number loss encompassing *APC* were found in both tumors, and no other sequence variations were identified. Moreover, both insertions were in close proximity (2,151 bp) to two previously reported insertions<sup>24,25</sup>. The location of the insertions was consistent with the distribution of non-synonymous point mutations detected in *APC* and were predicted to disrupt the protein reading frame of *APC* as previously reported<sup>24,25</sup> (Fig. 4). Taken together these findings, as well as the previous extensive knowledge of the tumor-initiating role of *APC* in most CRCs<sup>33,34</sup>, suggest that retrotransposon insertions may have contributed to the early steps of tumorigenesis in 2 of the 202 colorectal tumor patients.

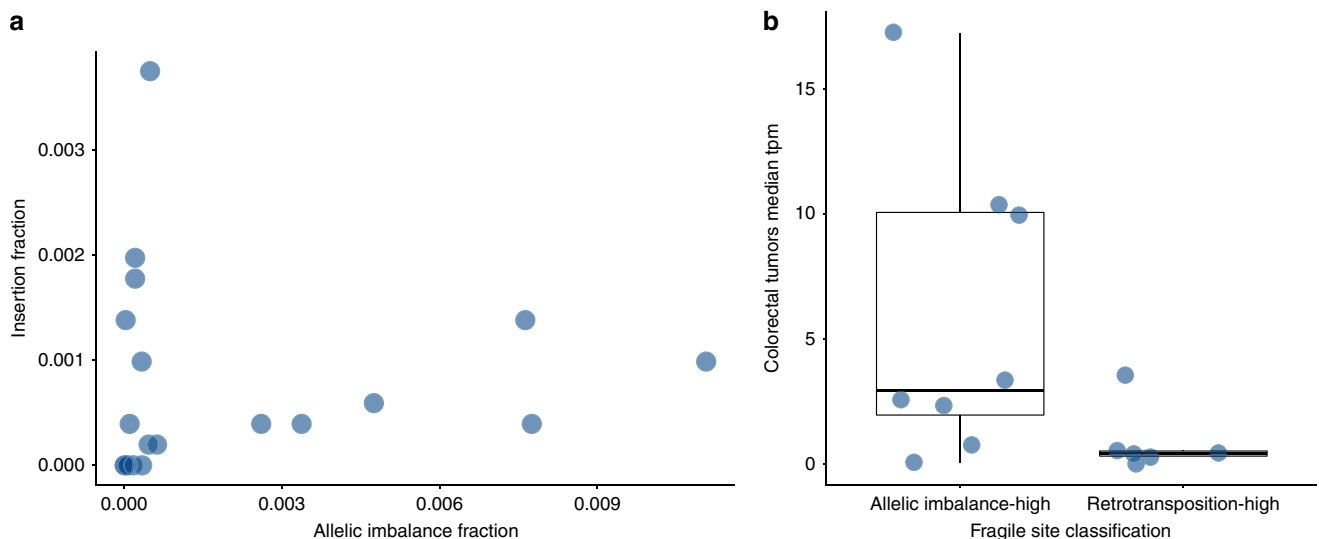
**Recurrent insertions in lowly expressed fragile sites.** We observed recurrent insertions in 12 out of 21 genes with high probability of being fragile as estimated in another study<sup>30</sup> (Supplementary Data 3). Since common fragile sites are prone to copy number alterations (CNAs)<sup>35</sup>, we evaluated whether retrotransposition and CNAs—in this study detected as AI<sup>36</sup>—were correlated (Supplementary Data 3). Fragile sites with high frequency of insertions seemed to display lower frequency of AI (Fig. 5a). Next, we investigated whether high frequency of insertions (insertion fraction/AI fraction > 1) and high frequency of AIs (0 < insertion fraction/AI fraction < 1) could result from differences in gene expression within fragile sites. Indeed, insertion frequency seemed to be higher in genes with lower



**Fig. 3** Retrotransposon insertion distribution in *LRP1B* and *DLG2*. Mapping of retrotransposon insertions identified in 202 colorectal tumors, HPV integration hotspots reported in 135 cervical cancers and allelic imbalance breakpoints identified in 1,699 CRCs<sup>31,36</sup>. Figure plotted with genoPlotR<sup>64</sup>. Source data are provided as a Source Data file



**Fig. 4** Distribution of non-synonymous changes and LINE-1 insertions on the linear protein of APC. Non-synonymous changes in 187 MSS CRCs, small lollipops. LINE-1 insertions, larger lollipops. p.N809L1 (c1049.1T) and p.P1526L1 (c310.1T), turquoise lollipops; p.F1396L1 and p.P1526L1<sup>24,25</sup> black lollipops. Figure modified from cBio cancer genomics portal<sup>59,60</sup>



**Fig. 5** Insertion and AI frequency in 21 fragile sites. **a** Insertion fraction over the fraction of allelic imbalance in 21 fragile sites. **b** Gene expression (median TPM values from 34 tumors) in fragile sites with high insertion fraction and fragile sites with high allelic imbalance fraction (Supplementary Data 3). Boxplot shows median, interquartile range (IQR), and whiskers extend to the most extreme data points which are no more than 1.5 times the IQR

expression (Exact Two-Sample Fisher-Pitman Permutation Test for log-transformed gene expression,  $p = 0.04$ ) (Fig. 5b). These results are concordant with our data and those of another study<sup>21</sup>; insertion density is overall negatively correlated with gene expression.

**Few active reference LINE-1s account for most transductions.**

We utilized the 3' unique sequence from the transduced regions to identify the reference source elements of LINE-1 transductions. We detected a total of 346 transductions arising from 56 of 315 human specific full-length reference LINE-1s. Fourteen out of the 56 reference elements were previously reported to be active in humans<sup>2,3</sup> and 28 were reported to be active in cancer (Supplementary Data 4)<sup>21</sup>. Recurrent transductions were detected from 24 LINE-1s, and in concordance with our previous study<sup>11</sup> the most active was the LINE-1 located in 22q12.1, which alone accounted for 160 transductions (46%). Seven and six percent of the transductions arose from the LINE-1s located in 9q32 and Xp22.2, respectively. Moreover, the insertion frequencies are in concordance with the frequencies reported by another study across 31 different tumor subtypes (Supplementary Data 4)<sup>21</sup>.

**Insertion count associates with CIMP and AI.** We investigated the associations between insertion count and molecular and clinical characteristics. We utilized 196 colorectal tumors with

complete information on molecular and clinical variables that were included in the model (Table 2, Supplementary Data 5). We applied a multiple linear regression model for log-transformed insertion counts, and hypothesized that the number of somatic insertions may be associated with tumor location, TP53 mutation, MSI, genomic fraction of AI<sup>36</sup> and CIMP. The model was adjusted for mean sequencing coverage, tumor stage, sex and age at diagnosis (Table 2). Goodness-of-fit was tested by Pearson's chi-square test ( $p = 0.99$ ). We found that insertion count was positively associated with CIMP (Multiple linear regression model,  $p = 0.00032$ ) and the genomic fraction of AI (Multiple linear regression model,  $p = 0.0036$ ) (Table 2). Moreover, both associations remained significant after including BRAF mutation (V600E) (Multiple linear regression model, CIMP,  $p = 0.004$ ; and genomic fraction of AI,  $p = 0.004$ ) and when only including MSS samples (Multiple linear regression model, CIMP,  $p = 0.001$ ; and the genomic fraction of AI,  $p = 0.006$ ). We also investigated whether insertion breakpoints were located at sites of chromosomal AI ( $\pm 5000$  bp from each AI breakpoint,  $n = 40,718$ ) however, only one colocalizing event was identified in one sample (c827, id4279).

**Insertion count associates with poor CRC survival.** We applied the Cox proportional hazards model in 192 patients with complete information on molecular and clinical variables that were used in the model (Table 3, Supplementary Data 5). Patients were followed for 1,370 person-years (Supplementary Data 5). We hypothesized that insertion count may be associated with disease-specific survival (Fig. 6). The model was adjusted for tumor stage, sex, MSI, the genomic fraction of AI, BRAF mutation and CIMP status (Table 3). As expected, advanced tumor stage (Dukes C and D) was strongly associated with CRC-specific survival. However, even after adjusting for the above-mentioned covariables, insertion count was independently associated with poor disease-specific survival (Cox proportional hazards model,  $p = 0.0029$ ) (Fig. 6, Table 3).

**Discussion**

Although retrotransposon activity is a hallmark of tumors of the gastrointestinal tract<sup>10,11,17-20</sup>, the role of retrotransposon insertions in CRC remains unclear with very few studies reported. Here, we characterized the somatic landscape of retrotransposon insertions in the largest dataset of colorectal tumor whole-genomes reported to date, and identified significant associations with clinical characteristics.

**Table 2 Multiple linear regression model for log insertion counts**

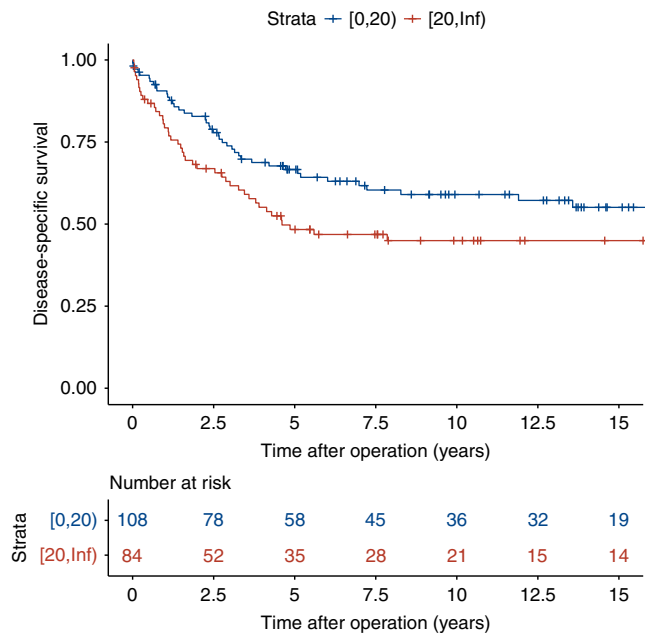
	Coefficient	Std. err.	z	p	Signif.
Intercept	0.408	0.647	0.630	5.29e-01	
CIMP-H	0.607	0.169	3.60	3.22e-04	***
Allelic Imbalance (/10% of reference)	0.0826	0.0284	2.91	3.64e-03	**
TP53 mutation	-0.0684	0.134	-0.509	6.10e-01	
MSI	0.150	0.285	0.527	5.98e-01	
Mean coverage (/10 reads)	0.309	0.0862	3.59	3.33e-04	***
Age at diagnosis (/10 years)	0.0331	0.0603	0.549	5.83e-01	
Male	0.0570	0.123	0.464	6.42e-01	
Dukes B	-0.0578	0.168	-0.344	7.31e-01	
Dukes C	-0.0483	0.186	-0.259	7.96e-01	
Dukes D	0.00490	0.211	0.0232	9.81e-01	
Proximal location	0.206	0.144	1.43	1.52e-01	

MSI microsatellite instability, CIMP-H CpG methylator phenotype high  
Significance codes: \*\*\*  $\leq 0.001$  < \*\*  $\leq 0.01$  < \*  $\leq 0.05$  < .  $\leq 0.1$

**Table 3 Cox proportional hazards model for disease-specific survival**

	Coefficient	Std. err.	z	p	HR [95% CI]	Signif.
Insertion count (/10)	0.108	0.0362	2.98	2.93e-03	1.11 [1.04, 1.20]	**
MSI	-0.258	0.642	-0.402	6.88e-01	0.773 [0.219, 2.72]	
CIMP-H	0.174	0.341	0.510	6.10e-01	1.19 [0.610, 2.32]	
BRAF mutation	0.790	0.447	1.77	7.68e-02	2.20 [0.918, 5.29]	.
Age [55, 75] years	-0.147	0.408	-0.360	7.19e-01	0.863 [0.388, 1.92]	
Age $\geq 75$ years	0.188	0.427	0.439	6.60e-01	1.21 [0.523, 2.78]	
Male	0.311	0.232	1.34	1.80e-01	1.37 [0.866, 2.15]	
Dukes B	0.452	0.449	1.01	3.13e-01	1.57 [0.652, 3.79]	
Dukes C	1.77	0.431	4.12	3.82e-05	5.89 [2.53, 13.7]	***
Dukes D	2.78	0.454	6.12	9.07e-10	16.2 [6.64, 39.4]	***
Allelic Imbalance (/10% of reference)	-0.0583	0.0539	-1.08	2.80e-01	0.943 [0.849, 1.05]	

The model was stratified by tumor location  
HR Hazard ratio, CI confidence interval, MSI microsatellite instability, CIMP-H CpG island methylator phenotype high  
Significance codes: \*\*\*  $\leq 0.001$  < \*\*  $\leq 0.01$  < \*  $\leq 0.05$  < .  $\leq 0.1$



**Fig. 6** Kaplan-Meier curves by insertion count. Tumors with less than 20 somatic insertions (blue line) and tumors with 20 or more insertions (red line)

We observed high retrotransposon activity with wide variability among tumors. We confirmed higher insertion density in late replicating regions, closed chromatin. Among insertions in genes, we also observed higher insertion count in genes with lower expression. The list of the most active reference LINE-1s became also validated in this extended set of CRCs<sup>10,11,21</sup>.

A number of additional observations were made. We identified recurrent insertions in 333 protein-coding genes, 15 of which are included in the CGC<sup>29</sup>. The most recurrent hit was *LRP1B* with 19 intronic insertions. The high frequency of insertions in this gene could be a result of various characteristics such as chromatin state, replication timing as well as gene length and/or expression. However, other causes such as sequence composition or somatic selection cannot be excluded. We also observed a high frequency of insertions at fragile sites with lower gene expression and lower AI fraction. These findings suggest that while AI is a recurrent feature of some fragile sites, sites with lower gene expression are more prone to retrotransposon insertions in CRC. Yet, the molecular basis of frequent retrotransposition in fragile sites, in particular *LRP1B*, and whether these insertions are important for the tumorigenic process remain as open questions.

Among the exonic insertions, we identified one in *PIK3CA* and two in *APC*. *PIK3CA* is a known oncogene involved in colorectal tumor progression and mutations in *APC* lead to colorectal tumor initiation<sup>33,34</sup>. The insertion locations in *APC* were similar to two previously reported insertions<sup>24,25</sup> and consistent with the distribution of pathogenic somatic changes in this gene. Similar to most pathogenic *APC* mutations, the insertions were predicted to disrupt the open reading frame of the gene. These observations, together with the previous extensive work showing the key role of *APC* loss early in colorectal neoplasia<sup>33,34</sup>, suggest that retrotransposon insertions in *APC* is one mechanism of CRC initiation as previously proposed<sup>24,25</sup>, although the inactivation of *APC* by retrotransposition should be functionally assessed in future studies. In addition, we identified recurrent intronic retrotranspositions in other genes frequently mutated in CRC, such as *FAT4*, and whether these insertions are important for the tumorigenic process remains to be investigated.

The availability of patient data allowed us to investigate possible associations between somatic insertion count and various molecular and clinical characteristics. We applied a multiple linear regression model and found that retrotransposon activity was positively associated with the genomic fraction of AI, and paradoxically with CIMP even though LINE-1s are frequently repressed by promoter methylation<sup>12</sup>. Moreover, LINE-1 methylation was associated with MSI and the CIMP in a previous study in CRC<sup>37</sup>. Of note, both CIMP and the genomic fraction of AI are characteristic of the two distinct genetic instability pathways in CRC. No associations with age at diagnosis, *TP53* mutations or tumor stage were detected in contrast to other studies<sup>18,38</sup>. These studies had significantly smaller sample sizes and fewer covariables were taken into account, which may explain the discrepancy. Importantly, survival analysis revealed a significant association between insertion count and poor disease-specific survival independently of other prognostic factors. Our findings indicate that tumors with high retrotransposon activity present characteristics of both MSS and MSI tumors, and are associated with poor CRC-specific survival. Although, further studies need to confirm the prognostic value of retrotransposon insertions not only in CRC, but also in other cancer types.

By characterizing the landscape of retrotransposon insertions in a large dataset of CRCs, we found that retrotranspositions appear to have the ability to serve as tumor-initiating events in CRC. The association of retrotransposition events with clinical characteristics—in particular poor prognosis—suggest that retrotransposition may play a more important role in CRC than previously thought. Further work should elucidate the timing and mechanisms leading to high somatic retrotransposition activity in some individuals, while others are spared. Understanding these could provide tools for management of CRC, including prevention.

## Methods

**Study subjects.** The samples and the clinical data utilized in this study were obtained from a population based series of 1042 CRCs<sup>39,40</sup> and from a subsequently collected series of additional Finnish CRCs. The tumors were fresh frozen and the corresponding normal tissues were obtained from either blood or from the normal colon tissue. Originally 202 CRCs entered the analyses. However, one tumor was later classified as an advanced adenomatous lesion (c232.1T). All samples were collected after informed consent. In the great majority of cases the consent was signed, in few cases collected before 1999 a verbal consent was derived (signed informed consent was not required in the Finnish legislation prior to that). For these early samples subsequent authorization for research use was derived from the National Supervisory Authority for Welfare and Health (Dnro 421/04/044/06, Dnro 8048/06.01.03.01/2014, Dnro 358/32/300/05, Dnro 1476/06.01.03.01/2012). The study has been reviewed by the Ethics Committee of the Hospital district of Helsinki and Uusima (Dnro 133/E8/03, 408/13/03/03/2009). Permission to use patient information was obtained from the National Institute for Health and Welfare (Dnro 53/07/2000, Dnro THL/1071/5.05.00/2011, Dnro THL/151/5.05.00/2017).

**Whole genome sequencing.** WGS was performed on Illumina HiSeq 2000 with 100 bp paired-end reads. Each normal and tumor DNA was sequenced to at least 40× median coverage. Data was processed similar to GATK best practices<sup>36,41</sup>.

**Transposon detection.** The identification of somatic retrotransposon insertions was conducted utilizing the Transposon Finder in Cancer (TraFiC)<sup>10</sup>. TraFiC default parameters were applied except for;  $a = 1$  (RepeatMasker accuracy),  $s = 3$  (minimum of three reads in tumor cluster), and  $gm = 3$  (minimum of three reads in normal cluster). In addition, paired-end reads with both ends having equal mapping quality and above 0 were included. In these cases, the first end of the pair was selected as the anchor read (end mapping to non-repetitive sequence). RepeatMasker (version open-4.0.5) and NCBI/RMBLAST 2.2.27+ were used for retrotransposon alignment as part of TraFiC. The RepeatMasker Database release utilized in this study was 20140121<sup>42,43</sup>. Somatic filtering was performed against germline calls from 234 normal samples (202 corresponding CRC normals, 20 myometrium samples<sup>44</sup>, and 12 blood samples<sup>45</sup>) with a 200 bp window as described in TraFiC<sup>10</sup>. Furthermore, calls in decoy sequences from 1000 Genomes Project Phase 2 (hs37d5)<sup>46</sup> were filtered away.

**Detection of LINE-1 transductions.** We identified 3' and orphan transductions utilizing DELLY structural variant (SV) calls (v 0.0.9)<sup>28,41</sup>. Filtering criteria utilized in this study were: SV calls supported by at least three supporting discordant reads and mapping quality > 37. SV calls were merged if they were the same DELLY type and were within 200 bp. Merged SVs (SV length > 1000 bp) in tumors were filtered against merged SVs in the normal samples. Subsequently, we extracted the SVs with one end of the pair within 1000 bp from the 3' end of a reference human-specific LINE-1 (Reference LIHS, full-length) from The European database of LIHS retrotransposon insertions in humans (eu.L1db)<sup>47</sup>, database version v1.0, date 05-10-14. The other end of the pair was used in the somatic filtering, where a 200 bp window and transduction calls from the pool of normal samples above mentioned were applied. One transduction detected in a female coming from an LINE-1 in Yp11.2 was filtered away. Furthermore, transduction calls within 200 bp, from the same retrotransposon family, and in the same sample were regarded as the same insertion and merged together. The same rationale was applied for calls detected by both DELLY and TraFiC.

**Methylation.** A Methylation-Specific Multiplex Ligation-dependent Probe Amplification assay (MS-MLPA) (Nygren AO, 2005) with the SALSA MLPA ME042 CIMP probemix (MRC-Holland, Amsterdam, The Netherlands) was used to determine the CIMP in an extended set of 255 tumor samples and the corresponding normal colon tissue of 175 samples as a separate study. Data from normal samples were used to determine the threshold for hypermethylation in the tumor samples. MS-MLPA was performed according to manufacturer's instructions<sup>48</sup> (<http://www.mrc-holland.com> Accessed December 2015). In short, the assay targets the promoter region of eight tumor suppressor genes: *CACNA1G*, *CDKN2A*, *CRABP1*, *IGF2*, *MLH1*, *NEUROG1*, *RUNX3*, *SOC1*. The methylation level for each probe was called using the Coffalyser software (MRC-Holland, Amsterdam, The Netherlands). If  $\geq 25\%$  of the probes for one gene were methylated, the gene was scored as methylated. If 5–8 genes, were scored as methylated, the tumor was classified as CIMP-high (CIMP-H), and if 0–4 genes were scored as methylated it was classified as CIMP-low (CIMP-L) tumor (Source data are provided as a Source Data file).

**RNA sequencing.** Total RNA from consecutive cryosections was extracted using RNeasy Mini Kit (Qiagen) from 34 tumors that displayed more than 50% of cancer cell percentage (HE staining of cryosections) and RNA integrity > 6 (Agilent RNA 6000, Agilent 2100 Bioanalyzer). Paired-end RNA sequencing was performed on the Illumina HiSeq 2000<sup>49</sup>. RNA-seq data was processed using Kallisto (version 0.43.0) software<sup>50</sup>. Kallisto quantification was executed in paired-end mode and aligned against the Ensembl Human reference transcriptome (GRCh37\_79). Quantification results from Kallisto were normalized and aggregated to gene-level utilizing sleuth (version 0.28.1) R package with default filtering settings<sup>51</sup>.

**Visual inspection of paired-end read data.** We selected 100 random insertions to ascertain the rate of true somatic calls based on visual inspection of the paired-end read data. Visualization was performed with BasePlayer<sup>52</sup>. Somatic calls were visually validated as true if the insertion call was supported by discordant reads (three + three for TraFiC calls) and at least two split reads supporting the insertion breakpoint and/or the polyA/T. Furthermore, the corresponding normal tissue was also visualized to confirm the somatic origin of the insertion calls.

**Insertion annotation.** Annotation of the insertion calls was applied by using the inner genomic coordinates of the reciprocal clusters provided by TraFiC (P\_R\_POS & N\_L\_POS) (Supplementary Data 1). Insertion breakpoints hitting an intron or an exon of any protein-coding transcript (GRCh37\_87) were annotated as intron/exon hit. Insertion orientation was determined by the presence of a polyA or a polyT (within a 200 bp window from mid point between positive breakpoint and negative breakpoint) in conjunction with gene orientation. PolyA was called when at least two forward strand reads started with three or more consecutive "A" bases. We used sequences of other reads to detect "A" repeats in the reference (i.e., the polyA call was discarded if "A" repeat was found in the middle of other overlapping read). PolyT was called using the reverse strand reads with three or more consecutive "T" bases at the end of the read sequence not present in the reference as described above (Supplementary Data 1). Insertion strand with respect to reference was defined as reverse when a polyA was called, and defined as forward when a polyT was called. Sense insertions were defined when the insertion strand with respect to reference was reverse in genes in plus orientation or forward in genes in minus orientation. Antisense insertions were defined when insertion strand with respect to reference was reverse in genes in minus orientation or forward in genes in plus orientation (Supplementary Data 1). Replication time fractions were extracted from Chen et al.<sup>53</sup>. Insertion density was defined as number of insertions divided by the total number of base pairs of each replication time fraction. Open chromatin was defined as DNase regions that were overlapping in at least two out of the four cell lines (RKO, LoVo, CaCo2, and Gp5D) (GSE83968)<sup>54</sup> in the 1000 Genomes Project pilot style callable regions<sup>55</sup>. Closed chromatin regions were defined as the above-mentioned callable regions minus the open chromatin regions.

**GO analysis.** We applied GO analyser for RNA-seq and other length biased data (Goseq)<sup>56</sup> with R version 3.5.1 to identify enrichment of biological processes in genes with recurrent insertions. Genes with recurrent insertions were defined as genes with two or more insertions. We utilized Wallenius approximation and p-values were corrected using Benjamini and Hochberg method. Enrichment was considered for FDR corrected p-values above 0.05.

**Fragile sites.** The 21 fragile sites were defined as genes with more than 0.85 probability of being fragile (Random forest 3 predictors)<sup>30</sup>. Genomic coordinates were lifted to GRCh37/Hg19 with <https://genome.ucsc.edu/cgi-bin/hgLiftOver57> and regions with no converted coordinates were excluded (chr10:46597226–48877831, chr10:45970128–48447930). The fraction of AI was calculated as the number of focal AI events per fragile site (both breakpoints of each AI call within the fragile site coordinates) divided by the total number of AI events in 1699 tumors<sup>36</sup>. Insertion fraction was calculated as the number of insertions per fragile site divided by the total number of insertions detected in 202 patients. Fragile site categories were defined based on the ratio of insertion fraction/AI fraction. AI high; 0 < ratio < 1, and Retrotransposon-high; ratio > 1.

**Mutation analysis in CRC genes.** Somatic changes in *BRAF*, *KRAS*, *TP53* and *APC* were called using MuTect (version 1.1.4) with default parameters (GRCh37\_78)<sup>36,41</sup>. Subsequent filtering criteria were minimum coverage of 4, minimal allelic fraction of 10 and minimum quality score 20<sup>52</sup>. For *KRAS*, mutations in codons 12, 13, 61, 117, and 146 in any transcript were classified as mutation positive and for *BRAF*, only hotspots in V600E in any transcript were considered as mutation positive. All non-synonymous changes in any transcript of *TP53* were classified as mutation positive. In addition, non-synonymous changes in *APC* (ENST00000457016) from 234 MSS tumors<sup>36</sup> were utilized for Fig. 4. Figure 4 was created with <http://www.cbioportal.org/tools.jsp58,59> and modified with Inkscape (<http://www.inkscape.org60>).

**Association test between insertions and RNA expression.** For the 827 insertions identified in any of the 34 tumors, we investigated the effect on the expression of the 642 distinct closest genes. For each sample and each gene the TPM values were extracted and ranked in ascending order. Consequently, the rank number corresponding to the sample with the insertion was recorded for each gene. We computed the sum-of-squared error statistic (Chi-square test) for the frequency table to test whether the rank values of the samples with insertion were uniformly distributed (no insertion effect on gene expression). Furthermore, 100,000 permutations with randomized rank numbers were applied but no significant effect was observed (Supplementary Fig. 2). Tests were performed using R versions 3.4.3 or 3.3.0.

**Multiple linear regression analysis.** To model retrotransposon insertion counts we applied a multiple linear regression model for log-transformed insertion counts. Spearman correlation matrix (R package PerformanceAnalytics) and variance inflation factors (vif function in R package car) were computed to evaluate possible collinearity among explanatory variables<sup>61,62</sup>. Model fit was assessed by plotting residuals against fitted values, theoretical normal quantiles and leverage (Supplementary Figs. 3–5). All tests were performed using R version 3.3.2<sup>63</sup>.

**Cox proportional hazards regression analysis.** We applied the Cox proportional hazards regression to study the association between disease-specific survival with retrotransposon insertion counts. The time variable was defined as days since diagnosis or operation. Patients that were alive in the last status assessment were censored at that date (survival status was assessed periodically using the Population Register Centre of Finland with the most recent assessment in 2016). Death from other causes than CRC were also defined as censored events. Proportional hazards assumptions were assessed by Grambsch-Therneau test for proportional hazards and evaluation for a non-zero slope of the scaled Schoenfeld residuals versus time (Supplementary Fig. 6). Based on inspection of the scaled Schoenfeld residuals, the model was stratified by tumor location. Influential observations were assessed with dfbeta and martingale residuals (Supplementary Figs. 7 and 8). All tests were performed using R version 3.3.2<sup>63</sup>.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The raw sequencing data produced in this study is not available due to the presence of germline data- and thus identifiable information-which we do not have the specific consent to distribute. The whole-genome somatic point mutations have been deposited in the EGA database under the accession code [EGAS00001003010](https://ega-archive.org/studies/EGAS00001003010). Gene expression values have been deposited in the Zenodo database under the Digital Object Identifier [<https://doi.org/10.5281/zenodo.3241399>]. The methylation source data and the data underlying Fig. 3 are provided as Source Data files. All the other data supporting the findings of this study are available within the article and its Supplementary Information

files. A reporting summary of this article is available as a Supplementary Information files.

### Code availability

The code is available in <https://github.com/cajuso/RetroCRCmanu>

Received: 23 July 2018 Accepted: 31 July 2019

Published online: 06 September 2019

### References

- Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Brouha, B. et al. Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. USA* **100**, 5280–5285 (2003).
- Beck, C. R. et al. LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010).
- Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35**, 41–48 (2003).
- Hancks, D. C., Goodier, J. L., Mandal, P. K., Cheung, L. E. & Kazazian, H. H. Jr. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.* **20**, 3386–3400 (2011).
- Raiz, J. et al. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* **40**, 1666–1683 (2012).
- Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H. Jr. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum. Genet.* **73**, 1444–1451 (2003).
- Holmes, S. E., Dombroski, B. A., Krebs, C. M., Boehm, C. D. & Kazazian, H. H. Jr. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat. Genet.* **7**, 143–148 (1994).
- Goodier, J. L., Ostertag, E. M. & Kazazian, H. H. Jr. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000).
- Tubio, J. M. C. et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
- Pitkänen, E. et al. Frequent L1 retrotranspositions originating from TTC28 in colorectal cancer. *Oncotarget* **5**, 853–859 (2014).
- Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).
- Alves, G., Tatro, A. & Fanning, T. Differential methylation of human LINE-1 retrotransposons in malignant cells. *Gene* **176**, 39–44 (1996).
- Shukla, R. et al. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111 (2013).
- Daskalos, A. et al. Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. *Int. J. Cancer* **124**, 81–87 (2009).
- Burns, K. H. Transposable elements in cancer. *Nat. Rev. Cancer* **17**, 415–424 (2017).
- Lee, E. et al. Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).
- Solyom, S. et al. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res.* **22**, 2328–2338 (2012).
- Helman, E. et al. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.* **24**, 1053–1063 (2014).
- Ewing, A. D. et al. Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome Res.* **25**, 1536–1545 (2015).
- Rodriguez-Martin, B. et al. Pan-cancer analysis of whole genomes reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours. Preprint at <https://www.biorxiv.org/content/10.1101/179705v3>. (2017).
- Pradhan, B. et al. Detection of subclonal L1 transductions in colorectal cancer by long-distance inverse-PCR and Nanopore sequencing. *Sci. Rep.* **7**, 14521 (2017).
- Ye, D., Jiang, D., Li, Y., Jin, M. & Chen, K. The role of LINE-1 methylation in predicting survival among colorectal cancer patients: a meta-analysis. *Int. J. Clin. Oncol.* **22**, 749–757 (2017).
- Miki, Y. et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res.* **52**, 643–645 (1992).
- Scott, E. C. et al. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* **26**, 745–755 (2016).
- Boland, C. R., Richard Boland, C. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**, 2073–2087.e3 (2010).
- Toyota, M. et al. CpG island methylator phenotype in colorectal cancer. *Proc. Natl Acad. Sci. USA* **96**, 8681–8686 (1999).
- Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
- Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
- Rajaram, M. et al. Two distinct categories of focal deletions in cancer genomes. *PLoS ONE* **8**, e66264 (2013).
- Hu, Z. et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat. Genet.* **47**, 158–163 (2015).
- Carbon, S. et al. AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288–289 (2009).
- Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
- Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.* **6**, 479–507 (2011).
- Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
- Palin, K. et al. Contribution of allelic imbalance to colorectal cancer. *Nat. Commun.* **9**, 3664 (2018).
- Ogino, S. et al. LINE-1 hypomethylation is inversely associated with microsatellite instability and CpG island methylator phenotype in colorectal cancer. *Int. J. Cancer* **122**, 2767–2773 (2008).
- Jung, H., Choi, J. K. & Lee, E. A. Immune signatures correlate with L1 retrotransposition in gastrointestinal cancers. *Genome Res.* **28**, 1136–1146 (2018).
- Aaltonen, L. A. et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N. Engl. J. Med.* **338**, 1481–1487 (1998).
- Salovaara, R. et al. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *J. Clin. Oncol.* **18**, 2193–2200 (2000).
- Katainen, R. et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).
- Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
- RepeatMasker. Available at: <https://www.repeatmasker.org>.
- Mehine, M. et al. Characterization of uterine leiomyomas by whole-genome sequencing. *N. Engl. J. Med.* **369**, 43–53 (2013).
- Välimäki, N. et al. Whole-genome sequencing of growth hormone (GH)-secreting pituitary adenomas. *J. Clin. Endocrinol. Metab.* **100**, 3918–3927 (2015).
- 1000 Genomes Project Consortium. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- Mir, A. A., Philippe, C. & Cristofari, G. euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Res.* **43**, D43–D47 (2015).
- Nygren, A. O. H. Methylation-Specific MLPA (MS-MLPA): simultaneous detection of CpG methylation and copy number changes of up to 40 sequences. *Nucleic Acids Res.* **33**, e128–e128 (2005).
- Ongen, H. et al. Putative cis-regulatory drivers in colorectal cancer. *Nature* **512**, 87–90 (2014).
- Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
- Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687–690 (2017).
- Katainen, R. et al. Discovery of potential causative mutations in human coding and noncoding genome with the interactive software BasePlayer. *Nat. Protoc.* **13**, 2580–2600 (2018).
- Chen, C.-L. et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.* **20**, 447–457 (2010).
- Liu, N. Q. et al. The non-coding variant rs1800734 enhances DCLK3 expression through long-range interaction and promotes colorectal cancer progression. *Nat. Commun.* **8**, 14418 (2017).
- Katsnelson, A. 1000 Genomes Project reveals human variation. *Nature* (2010). <https://doi.org/10.1038/news.2010.567>.
- Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
- Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
- Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
- Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, 11 (2013).



60. Inkscape Project. Inkscape. Available at: <https://inkscape.org>. (2011)
61. Fox, J. & Weisberg, S. *An R Companion to Applied Regression*. (SAGE, London, UK, 2011).
62. Peterson, B. G. & Carl, P. PerformanceAnalytics: econometric tools for performance and risk analysis. (2014).
63. R Core Team. *R: A language and environment for statistical computing*. (Vienna, Austria, 2016).
64. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).

## Acknowledgements

The authors thank Alison Ollikainen, Iina Vuoristo, Inga-Lill Åberg, Sini Marttinen, Marjo Rajalaakso, Sirpa Soisalo, Jiri Hamberg, and Heikki Metsola for technical assistance and Alison Ollikainen also for proofreading the manuscript. This work was supported by grants from the Academy of Finland (Finnish Center of Excellence Program 2012–2017, 250345 and 2018–2025, 312041), The Finnish Cancer Society, The European Research Council (268648), The Sigrid Juselius Foundation; Jane and Aatos Erkko Foundation, the Nordic Information for Action eScience Center (NIASC) and Nordic Center of Excellence financed by NordForsk (Project number 62721). We also thank SYSCOL (an EU FP7 Collaborative Project, 258236) for sequencing the RNA samples. The following foundations are acknowledged for personal funding: Ida Montinin Säätiö foundation, Cancer Society of Finland, Juhani Ahon Foundation for Medical Research and The Maud Kuistila Memorial Foundation. The authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

## Author contributions

T.C., P.S., and T.T. analyzed insertion data. T.C., P.S. and E.P. contributed and performed insertion and transduction calling. T.C., T.T., U.A.H. and J.K. prepared WGS samples. O.K. supervised WGS sample preparation. T.C. and O.K. contributed and organized RNA sample preparation. T.C. and T.T. performed statistical analysis. R.K., E.P., K.P. and N.V. were involved in primary WGS data analysis. T.C. and R.K. performed and analyzed insertion polyA/T calls and insertion orientation bias, and R.K. developed BasePlayer. A.T. and L.F. performed CIMP analysis. A.T. and K.P. designed and performed the analysis of insertion effect on gene expression. N.V. performed primary RNA-seq analysis. A.R. and J.B. reviewed tumors. S.K., A.L., L.R.S., T.K., T.S. and J.P.M. provided patient samples. E.K. and M.A. contributed to the study design. T.C., O.K., E.P.,

K.P. and L.A.A. designed the study. O.K., E.P., K.P. and L.A.A. supervised the study. All authors contributed to writing the manuscript.

## Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-019-11770-0>.

**Competing interests:** The Authors declare the following competing interests. L.A.A. has received a lecture fee from Roche Oy and Bayer. M.A. has received a non-financial support from Roche Diagnostics Oy in a form of paid travel expenses for the Roche Sequencing Solutions Sample preparations EMEA User Meeting in Heidelberg, Germany. The remaining authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Peer review information:** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019