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

Tatiana Cajuso1,2, Päivi Sulo1,2, Tomas Tanskanen1,2, Riku Katainen1,2, Aurora Taira1,2, Ulrika A. Hänninen1,2, Johanna Kondelin1,2, Linda Forsström1,2, Niko Välimäki1,2, Mervi Aavikko1,2, Eevi Kaasinen1,2, Ari Ristimäki1,3, Selja Koskensalo4, Anna Lepistö4, Laura Renkonen-Sinisalo4, Toni Seppälä4, Teijo Kuopio5,6, Jan Böhm6, Jukka-Pekka Mecklin7,8, Outi Kilpivaara1,2, Esa Pitkänen1,2, Kimmo Palín1,2 & Lauri A. Aaltonen1,2

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.
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, however only a few elements remain retrotransposition competent and account for most retrotranspositions. 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. In cancer, ~24% of somatic retrotransposition events involve 3' transduction, a process characterized by mobilization of 3' flanking sequence which can serve as a unique sequence revealing the insertion origin.

LINE-1s are frequently repressed by promoter methylation and genome-wide hypomethylation is reported to lead to their activation during tumorigenesis, thus leading to high retrotransposition activity and genome instability. High retrotransposon activity has been reported in several human cancers, especially in tumors arising from the gastrointestinal tract, such as colorectal cancer (CRC). 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. Furthermore, ongoing retrotransposon activity has been reported in CRC. Insertion count is associated with patient age and LINE-1 hypomethylation is associated with poor survival in CRC. LINE-1 insertions in APC have been reported in two CRCs, indicating that these insertions may be early tumorigenic events. 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. Seventy-five percent of MSI-positive sporadic CRCs are attributed to the CpG island methylator phenotype (CIMP) 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 and DELLY 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. 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. The mean number of insertions per tumor was 25 (median, 17; interquartile range, 10–31) with high variability among tumors (Fig. 1a). 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, 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 replicating regions.

![Distribution of somatic insertions across 202 colorectal tumors and over replication time.](https://example.com/distribution.png)

**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.
Retrotranspositions 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 study21 (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)29 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 PTTRD 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 sites30 and recurrent hotspots for HPV integration31 (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-triphosphate binding for ARAP2; and RNA polymerase II transcription activity and DNA binding transcription factor activity for ZNF251. None of these genes have been classified as cancer genes29, fragile sites30 or hotspots of HPV integration31. 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 insertions24,25. 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 reported24,25 (Fig. 4). Taken together these findings, as well as the previous extensive knowledge of the tumor-initiating role of APC in most CRCs33,34, 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 study30 (Supplementary Data 3). Since common fragile sites are prone to copy number alterations (CNAs)35, we evaluated whether retrotransposition and CNAs—in this study detected as AI36—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

### 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 | Gene ID       | Gene name | Number of insertions (n = 202) | Cancer census role |
|---------------|-----------|-------------------------------|--------------------|---------------|-----------|-------------------------------|--------------------|
| ENSG000000168702 | LRP1B     | 19                            | TSG                | ENSG000000178568 | ERBB4     | 7                            | Oncogene, TSG      |
| ENSG000000171094 | ALX       | 5                            | Oncogene, fusion   | ENSG000000185811 | IKZF1     | 3                            | TSG                |
| ENSG000000196090 | PTPTR1A   | 3                            | TSG                | ENSG000000466899 | PREX2     | 3                            | Oncogene           |
| ENSG000000183454 | GRIN2A2   | 3                            | TSG                | ENSG000000144218 | AFF3      | 2                            | Oncogene, fusion   |
| ENSG00000015768 | NRG1      | 2                            | TSG, fusion        | ENSG000000079102 | RUNX1T1   | 2                            | Oncogene           |
| ENSG000000156702 | FUT1      | 2                            | Oncogene, fusion   | ENSG000000134982 | APC       | 2                            | TSG                |
| ENSG000000189283 | FHT1      | 2                            | TSG, fusion        | ENSG000000085276 | MECOM     | 2                            | Oncogene, fusion   |
| ENSG000000196159 | FAT4      | 2                            | TSG                |

Gene names are shown in italics. Cancer census role, role in cancer as defined by the Cancer Gene Census 29. TSG tumor suppressor gene.

replicating regions (replication time < 0.2, 0.73 insertions per Mbp) (Fig. 1b).
**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.31,36. Figure plotted with genoPlotR.64. 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 (c.1049.1T) and p.P1526L1 (c.310.1T), turquoise lollipops; p.F1396L1 and p.P1526L124,25 black lollipops. Figure modified from cBio cancer genomics portal59,60.

**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\(^2\); 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\(^3\) and 28 were reported to be active in cancer (Supplementary Data 4)\(^2\). Recurrent transductions were detected from 24 LINE-1s, and in concordance with our previous study\(^1\) 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\(^2\) across 31 different tumor subtypes (Supplementary Data 4)\(^2\).

**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, the genomic fraction of AI, 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 (Y600E) (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 (±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 covariates, insertion count was independently associated with poor disease-specific survival (Cox proportional hazards model, \( p = 0.0029 \)) (Fig. 6, Table 3).

**Discussion**

Although retrotransposition activity is a hallmark of tumors of the gastrointestinal tract\(^10,11,17–20\), 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.
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 retrotransposition activity was positively associated with the genomic fraction of AI, and paradoxically with CIMP even though LINE-1s are frequently repressed by promoter methylation. Moreover, LINE-1 methylation was associated with MSI and the CIMP in a previous study in CRC\(^2\). 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\(^3,\)\(^4\). 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 retrotransposition 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 retrotransposition insertions not only in CRC, but also in other cancer types.

By characterizing the landscape of retrotransposition 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\(^3,\)\(^4\) 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 431/04/04/06, Dnro 8048/06.01.03.01/2014, 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 33/07/2000, Dnro THL/107/5.05.00/2011, 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 40x median coverage. Data was processed similar to GATK best practices\(^3,\)\(^4\).

**Transposon detection.** The identification of somatic retrotransposition insertions was conducted utilizing the Transposon Finder in Cancer (TraFiC)\(^5\). TraFiC default parameters were applied except for: \(a = 1\) (RepeatMasker accuracy), \(s = 3\) (minimum of three reads in tumor cluster), and \(g = 3\) (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\(^5\)\(^,\)\(^6\). Somatic filtering was performed against germline calls from 234 normal samples (202 corresponding CRC normals, 20 myometrium samples\(^4\), and 12 blood samples\(^7\)) with a 200 bp window as described in TraFiC\(^10\). Furthermore, calls in decay sequences from 1000 Genomes Project Phase 2 (hs37d5)\(^8\) were filtered away.
Detection of LINE-1 transpositions. We identified 3′ and orphan transpositions utilizing DELLY structural variant (SV) calls (v 0.9)53,54. Filtering criteria utilized in this study were SV calls supported by at least one supporting discordant read, 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 1 LHS, full-length) from The European database of L1 HS retrotransposon insertions in humans (euL1db)57, 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 removed. Transcriptional events were defined as female coming in 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 (MRG-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 instructions48 (http://www.mrc-holland.com Accessed December 2015). In short, the assay targets the 5′ end of the eight tumor suppressor genes; CACNA1C, CDKN2A, CRABP1, IGF2, MLH1, NEUROG1, RUNX3, SOCS1. The methylation level for each probe was calculated using the Coffalyser software (MRC-Holland, Amsterdam, The Netherlands). If ≥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 RNA sequencing samples. MS-MLPA was performed according to manufacturer’s instructions48. Subsequent filtering criteria were minimum coverage of 4, minimal allelic fraction of 10 and minimum quality score 2052. For KRAS, mutations in codons 12, 13, 117, and 146 in any transcript were classified as mutation positive and for BRAF, only hotspot in V600E in any transcript were genes with recurrent insertions were transcribed. All non-synonymous changes in any transcript (ENST00000457016) from 234 MSS tumors36 were utilized for Fig. 4. Figure 4 was created with http://www.cbioportal.org/tools.jsp53,59 and modified with Inkscape (http://www.inkscape.org)56.

GO analysis. We applied GO analyser for RNA-seq and other length biased data (Goseq)57 with R version 3.5.1 to identify enrichment of biological processes in genes with recurrent events. 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 5′ end of being fragile (Rst from forest 3 predictors)52. Genomic coordinates were lifted to GRCh37/Hg19 with https://genome.ucsc.edu/cgi-bin/hgLiftOver56 and regions with no converted coordinates were excluded (chr1:046597226–48877831, chr3:45970128–48447930). The fraction of AI was calculated as the number of focal AI events per fragile site (both breakpoints of each fragile site were compiled) divided by the total number of AI events in 1699 tumors56. 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 MiTeic (version 1.1.4) with default parameters (GRCh37_78)56,41. Subsequent filtering criteria were minimum coverage of 4, minimal allelic fraction of 10 and minimum quality score 2052. For KRAS, mutations in codons 12, 13, 117, and 146 in any transcript were classified as mutation positive and for BRAF, only hotspot in V600E in any transcript were genes with recurrent insertions were transcribed. All non-synonymous changes in any transcript (ENST00000457016) from 234 MSS tumors56 were utilized for Fig. 4. Figure 4 was created with http://www.cbioportal.org/tools.jsp53,59 and modified with Inkscape (http://www.inkscape.org)56.

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 variables61,62. 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 3.3.261.

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. Influent observations were assessed with dfbeta and martingale residuals (Supplementary Figs. 7 and 8). All tests were performed using R version 3.3.261.

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 EGAS0000103010. 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.
Acknowledgements
The authors thank Alison Ollikainen, Iina Vuoristo, Inga-Lili 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 (NIAASC) 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