A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer

Emma C. Scott, University of Maryland
Eugene J. Gardner, University of Maryland
Ashiq Masood, University of Maryland
Nelson T. Chuang, University of Maryland
Paula Vertino, Emory University
Scott E. Devine, University of Maryland

Journal Title: Genome Research
Volume: Volume 26, Number 6
Publisher: Cold Spring Harbor Laboratory Press | 2016-06-01, Pages 745-755
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1101/gr.201814.115
Permanent URL: https://pid.emory.edu/ark:/25593/rtn66

Final published version: http://dx.doi.org/10.1101/gr.201814.115

Copyright information:

© 2016 Scott et al.
This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/).

Accessed December 20, 2018 8:25 PM EST
A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer

Emma C. Scott,1,2,8 Eugene J. Gardner,1,2,8 Ashiq Masood,2,3,4,9 Nelson T. Chuang,1,2,5 Paula M. Vertino,6,7 and Scott E. Devine1,2,3,4

1Graduate Program in Molecular Medicine, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA; 2Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA; 3Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA; 4Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA; 5Division of Gastroenterology, Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA; 6Department of Radiation Oncology, Emory University School of Medicine, Atlanta, Georgia 30322, USA; 7Winship Cancer Institute, Emory University, Atlanta, Georgia 30322, USA

Although human LINE-1 (L1) elements are actively mobilized in many cancers, a role for somatic L1 retrotransposition in tumor initiation has not been conclusively demonstrated. Here, we identify a novel somatic L1 insertion in the APC tumor suppressor gene that provided us with a unique opportunity to determine whether such insertions can actually initiate colorectal cancer (CRC), and if so, how this might occur. Our data support a model whereby a hot L1 source element on Chromosome 17 of the patient’s genome evaded somatic repression in normal colon tissues and thereby initiated CRC by mutating the APC gene. This insertion worked together with a point mutation in the second APC allele to initiate tumorigenesis through the classic two-hit CRC pathway. We also show that L1 source profiles vary considerably depending on the ancestry of an individual, and that population-specific hot L1 elements represent a novel form of cancer risk.

[Supplemental material is available for this article.]

Human LINE-1 (L1) elements are autonomous retrotransposons that continue to produce new “offspring” L1 insertions in human genomes (Beck et al. 2010, 2011; Ewing and Kazazian 2010; Huang et al. 2010; Iskow et al. 2010; Stewart et al. 2011; The 1000 Genomes Project Consortium 2015; Sudmant et al. 2015). Until recently, L1 elements were thought to be mobilized primarily in the germline and then silenced in somatic cells throughout adulthood. However, several recent reports have shown that L1 elements are active in at least some adult somatic tissues, including the brain (Muotri et al. 2005; Coufal et al. 2009; Baillie et al. 2011; Evrony et al. 2012; Upton et al. 2015) and epithelial somatic tumors (Miki et al. 1992; Iskow et al. 2010; Lee et al. 2012; Solyom et al. 2012; Shukla et al. 2013; Helman et al. 2014; Tubio et al. 2014; Doucet-O’Hare et al. 2015; Ewing et al. 2015; Rodic et al. 2015). These observations have led to the suggestion that L1 might play a role in initiating human cancers by mutating specific oncogenes or tumor suppressor genes in somatic cells.

However, several broad surveys of somatic L1 insertions in human cancers have detected only a few strong L1 driver candidates (Lee et al. 2012; Shukla et al. 2013; Helman et al. 2014; Tubio et al. 2014), and it is unclear whether any of these insertions could have initiated tumorigenesis in the cancers in which they were discovered (see Discussion). Thus, somatic L1 drivers that affect the earliest stages of tumorigenesis have been elusive in these studies, and it is presently unclear whether L1 has the capacity to initiate tumorigenesis in somatic cells. This could, in principle, reflect a lack of knowledge of the oncogenes and tumor suppressor genes that act at the earliest stages of tumorigenesis in human cancers. However, it might instead indicate that L1 elements are not generally capable of initiating tumorigenesis in somatic cells because they are effectively repressed in most cells. An alternative possibility is that somatic L1 mobilization generally occurs only after an L1-permissive environment is established in an emerging tumor. Under this scenario, L1 could not actually initiate tumorigenesis but might instead become active during the more advanced stages of tumor progression and metastasis (e.g., Rodic et al. 2015).

In perhaps the strongest study implicating L1 in tumor initiation thus far, a somatic L1 insertion was identified in a case of colorectal cancer (CRC) that disrupted the APC tumor suppressor gene (Miki et al. 1992). The APC gene is the earliest gatekeeper that is mutated in the majority (~85%) of CRC cases and, in fact, both copies of APC must be mutated to initiate CRC through this classic two-hit pathway (Kinzler and Vogelstein 1996; Fearon 2011). In the case of familial adenomatous polyposis (FAP) and other CRC syndromes, one mutated copy of APC is inherited in the germline, and the other is acquired during the lifetime of the individual. In the sporadic form of the disease, both copies of APC must be mutated independently in a somatic cell in order to initiate tumorigenesis (Kinzler and Vogelstein 1996; Fearon 2011). In the remaining 15% of CRC cases, tumorigenesis is instead initiated by a hypermutation phenotype that is associated with microsatellite instability (MSI) and faulty DNA repair (Fearon 2011). APC also can be mutated in this less common

© 2016 Scott et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

26:745–755 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/16; www.genome.org
form of MSI-CRC, but it does not promote the earliest stages of tumor initiation in such cases.

Although one of the two APC alleles clearly was disrupted by L1 in the sporadic case of CRC described by Miki et al. (1992), the status of the second APC allele was not investigated in that study. Furthermore, the MSI phenotype of the tumor was not investigated, leaving open the possibility that tumorigenesis was initiated by a hypermutation phenotype associated with faulty DNA repair, rather than by the more common pathway involving biallelic mutations in APC (Fearon 2011). It also is unclear how the L1 insertion in APC could have been generated sufficiently early to initiate tumorigenesis in a normal colon cell. Thus, many questions remain unanswered as to the genesis of this somatic L1 insertion in APC and its impact on tumorigenesis (Miki et al. 1992).

Results

A novel somatic L1 insertion disrupts the APC tumor suppressor gene

Inspired by Miki et al. (1992), we screened CRC samples obtained from our institution to identify additional somatic L1 insertions in APC and other genes that have been implicated in CRC (Methods; Iskow et al. 2010). Among the first 10 tumors that we screened with L1-seq assays, we discovered a somatic L1 insertion in patient 20444 that disrupted the open reading frame (ORF) of the APC tumor suppressor gene (Fig. 1). PCR assays and subsequent Sanger sequencing confirmed that this novel somatic L1 insertion disrupted the sixteenth exon of the APC gene at codon 1396 (p.F1396L1 in GenBank gene ID NM_000038; hg19 coordinate Chr 5:112,175,479) (Table 1; Fig. 1; Supplemental Tables S1, S5; Supplemental Methods). This insertion was identified in the same exon as the Miki et al. (1992) insertion, just 388 bp upstream of that insertion (Fig. 1). We determined that the new L1 insertion is 1387 bp in length, ends with a ~27-bp poly(A) tail, and is flanked by a 14-bp target site duplication (TSD) with the sequence 5′-CAGTGGATGTTTTT-3′ (Fig. 1A; Supplemental Table S1). The insertion is truncated at the 5′ end and also contains a 5′ inversion and a small internal deletion at the inversion junction that presumably were caused by twin priming (Ostertag and Kazazian 2001). PCR validation assays confirmed that the APC L1 insertion was found only in the tumor and was absent from normal adjacent tissues, thus confirming it as a true somatic L1 insertion (Supplemental Methods).

Whole-genome sequencing (WGS) confirms the L1 insertion in APC and reveals additional somatic mutations

We next performed Illumina whole-genome sequencing (WGS) on the normal/tumor pair from patient 20444 to gain a better understanding of how this L1 insertion might have worked together with other somatic mutations to promote tumorigenesis (Supplemental Data S1, S2; Supplemental Tables S1, S6; Supplemental Methods). We began by scanning the WGS data to identify additional somatic mobile element insertions (MEIs) that might have been generated in the tumor using the Mobile Element Locator Tool (MELT) that we developed for the 1000 Genomes Project (Methods; Sudmant et al. 2015). MELT confirmed the somatic L1 insertion in the APC gene with the features outlined above and detected 26 additional somatic L1 insertions in the tumor that were absent from adjacent normal tissues (for a total of 27 somatic L1 insertions) (Table 1; Supplemental Table S1). All 27 were validated as somatic insertions in the tumor with at least one junction PCR and/or sequencing assay (Supplemental Table S5; Supplemental Methods). In addition, we amplified and sequenced the entire insertion and flanking regions for 16 of these somatic L1 insertions, including the APC insertion (Supplemental Tables S1, S5; Supplemental Methods). Our data revealed that all 16 of these fully sequenced insertions terminated in poly(A) tails and were flanked by TSDs, indicating that they were generated by target primed reverse transcription (TPRT) as expected for

Figure 1. Mutagenesis of APC by a somatic L1 insertion. (A) A schematic of the L1 insertion in APC. The top diagram shows the location of the APC gene (vertical red bar) in band q22.2 of Chromosome 5. The diagram below depicts the 1387-bp somatic L1 insertion (dark and light green) in exon 16 of APC with the associated hallmarks of retrotransposition, including a flanking 14-bp TSD (orange), poly(A) tail (T in reverse order; red) and evidence for twin priming (two green boxes separated by an 18-bp deletion at the inversion point). (B) The two diagrams show the inactivating mutations that were discovered in both alleles of APC in the tumor. The top allele is inactivated by the L1 insertion at codon 1396 (multicolored bar; p.F1396L1) and has the reference codon at position 1450 (gray circle). The bottom allele is inactivated by the p.R1450* stop codon (red circle) and does not have an L1 insertion (gray circle). (C) The diagram shows the sites that are affected by these mutations within the APC protein. Also depicted is the L1 insertion identified by Miki et al. (1992) (red square; p.P1526L1). All three of these mutations occur within or near the somatic mutation cluster region (black bar) (Miyoshi et al. 1992) and are similar to other inactivating APC mutations in CRC (see Discussion). This image is adapted from the output of the Protein Painter tool (http://explore.pediatriccancergenomeproject.org/proteinPainter). The APC protein is 2843 amino acids long and consists of the following domains: Suppressor APC (red; involved in nuclear export and other functions); Armadillo/beta-catenin-like repeats (orange; mediate protein–protein interactions); APC cysteine-rich regions (yellow; bind beta-catenin); SAMP (green; binds axin); APC basic (blue; interacts with microtubules); and MAPRE1-binding (aka EB1-binding; purple; binds the microtubule-associated protein MAPRE1).
Table 1. Twenty-seven somatic L1 insertions identified in the patient’s tumor

<table>
<thead>
<tr>
<th>ID (Chr:pos)</th>
<th>Strand</th>
<th>Source</th>
<th>Length</th>
<th>Location</th>
<th>Gene symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:79306674 + ND 1524 Intergenic RDGR2</td>
<td>1:922334470 + ND 1382 Intronic RGS21</td>
<td>2:21127375 + Chr 12 1822 Intergenic RDGR2</td>
<td>2:46142515 - Unk 417 Intronic PRKCE</td>
<td>2:108074967 + Chr 14 1786 Intergenic RDGR2</td>
<td></td>
</tr>
<tr>
<td>1:115370008 + Chr 14 1066 Intronic DPP10</td>
<td>2:226441747 - Chr 14 367 Interonic NYAP2</td>
<td>3:16436230 - Unk 342 Intergenic RDGR2</td>
<td>4:65697624 - ND ND Intergenic RDGR2</td>
<td>4:120879367 + Chr 17 578 Intergenic RDGR2</td>
<td></td>
</tr>
<tr>
<td>4:163870337 - Chr 14 1118 Intergenic RDGR2</td>
<td>5:15917732 + Chr 14 610 Intronic FBXL7</td>
<td>5:36700111 + Chr 14 1921 Intergenic RDGR2</td>
<td>5:112175479 + Chr 17 1387 Exonic APC</td>
<td>6:146839867 - ND 1196 Intergenic RDGR2</td>
<td></td>
</tr>
<tr>
<td>7:68130460 + Chr 14 1424 Intergenic RDGR2</td>
<td>7:82409579 - Chr 14 837 Intronic PCLO</td>
<td>8:50819246 - ND ND Intergenic RDGR2</td>
<td>8:76890366 + Chr 17 1394 Intergenic RDGR2</td>
<td>9:108750714 + Chr 17 402 Intergenic RDGR2</td>
<td></td>
</tr>
<tr>
<td>9:72023991 + Chr 12 3087 Intergenic RDGR2</td>
<td>9:76488906 + Unk 97 Intergenic RDGR2</td>
<td>11:115649909 + Chr 12 286 Intergenic RDGR2</td>
<td>16:56298643 + Chr 14 5616 Intronic CNOT1</td>
<td>18:39948814 - Unk 188 Intronic D1CAP1-AS4</td>
<td></td>
</tr>
</tbody>
</table>
| 18:65529080 + Chr 14 1511 Intergenic LOC643542 | X:104566886 - Chr 14 1057 Intronic IL1RAPL2 | The data are tabulated for the 27 somatic L1 insertions in the patient’s tumor. See Supplemental Table S1 for additional details. (ID) location of insertion (Chromosome:position); (Unk) the source cannot be identified because the mutation profile is not informative; (ND) cannot be determined because of missing data.

genuine L1 retrotransposition events (Table 1; Supplemental Table S1; Luan et al. 1993; Jurka 1997).

Vogelstein and colleagues have described the landscape of genes that are frequently mutated in CRC and also have mapped the temporal order in which these genes are mutated (Kinzler and Vogelstein 1996; Fearon 2011; The Cancer Genome Atlas Network 2012). Therefore, we next identified somatic mutations in the tumor, including single nucleotide variants (SNVs) and short insertions/deletions (indels), to determine how L1 might have worked together with other types of mutations to initiate and drive tumorigenesis. We detected a C to T somatic SNV that created a premature stop codon in the sixteenth exon of APC just 160 bp downstream from the somatic L1 insertion (p.R1450*). This SNP was previously documented in the COSMIC database (Forbes et al. 2014). In fact, codon 1450 is one of the most frequently mutated APC sites in patients with sporadic CRC (Fearon 2011). These data indicate that L1 was responsible for inactivating one APC allele, and the p.R1450* stop codon was responsible for inactivating the second APC allele.

A microsatellite stable CRC

TCGA and others have identified two subtypes of CRC: one subtype with microsatellite stability (MSS) and a second subtype with high levels of microsatellite instability (MSI) (Fearon 2011; The Cancer Genome Atlas Network 2012). These two CRC subtypes also progress along different pathways: The MSS subtype is initiated by mutations in APC, whereas the MSI subtype is initiated by a hypermutation phenotype that is associated with faulty DNA repair (Markowitz and Bertagnolli 2009; Fearon 2011). Thus, we next sought to assign our CRC case to one of these two subtypes. We used the MSIsensor tool (Niu et al. 2014), together with the Illumina WGS data described above, to measure microsatellites in our tumor and adjacent normal control tissues (Methods). The microsatellite mutation rate in our CRC sample (0.01%) was clearly within the MSS range (Niu et al. 2014) and was well below the rates observed in tumors that have MSI phenotypes (0.5%–41%). Thus, our CRC sample was unambiguously assigned to the MSS subtype. Likewise, no somatic mutations were detected in DNA repair genes that previously have been linked to hypermutation phenotypes in CRC (MLH1, MLH3, MSH2, MSH3, MSH6, PMS2, POLE) (The Cancer Genome Atlas Network 2012). Collectively, these data demonstrate that our CRC case progressed along the classic MSS route whereby APC serves as the earliest gatekeeper. Therefore, these data formally demonstrate that the L1 insertion, combined with the p.R1450* stop codon, were responsible for initiating tumorigenesis in this case.

Other somatic mutations supporting an L1-initiated pathway of tumorigenesis

As outlined above, Vogelstein and colleagues and TCGA previously have defined the landscape of genes that are frequently mutated in CRC (including genes mutated in MSS-CRC), and we identified somatic mutations in three additional MSS-CRC genes in our tumor (Supplemental Data S1; Supplemental Methods). First, we identified a somatic point mutation in the PIK3CA gene (CAT to CGT; p.H1047R) that causes an amino acid substitution of histidine to arginine at codon 1047 of the encoded protein. PIK3CA is mutated in 18% of MSS-CRCs, and this precise mutation has been found previously in five independent CRCs by TCGA (The Cancer Genome Atlas Network 2012). We also identified an 18-bp tandem-duplication in the KRAS gene that includes codons 59 and 61 and retains the ORF of the encoded protein (p.L56_E62dup). KRAS is mutated in 43% of MSS-CRCs, and this mutation likely contributed to tumorigenesis since activating point mutations affecting codons 59 and 61 have been reported previously for this gene (The Cancer Genome Atlas Network 2012; Forbes et al. 2014). This specific tandem duplication has not been identified previously by TCGA nor is it present in the COSMIC database. However, COSMIC has several entries of somatic indels spanning positions 59 and 61 that previously have been implicated in a variety of cancer types, suggesting that indels in this region can activate KRAS (Forbes et al. 2014). We also identified a 2-bp somatic deletion in the ACVR1B gene that causes a frameshift in the third exon of this gene (p.K177fs). ACVR1B is mutated in 4% of MSS-CRC cases (The Cancer Genome Atlas Network 2012).
The p.R1450* mutation in APC, along with the PIK3CA, KRAS, and ACVR1B mutations outlined above, were validated through PCR, cloning, and Sanger capillary sequencing (Supplemental Table S5; Supplemental Methods). RNA-seq analysis of the normal and tumor specimens also revealed that the mutant alleles of these genes were all expressed in the tumor but were not expressed in the adjacent normal control tissues, as would be expected of somatic mutations (including the L1 insertion allele of APC, the p.R1450* allele of APC, and the mutant alleles of PIK3CA, KRAS, and ACVR1B described above) (Supplemental Table S3; Supplemental Methods). As an independent validation of the somatic mutations, we also performed Illumina WGS on a second genomic DNA preparation from a separate region of the tumor and confirmed the presence of the L1 insertion in APC along with the other somatic mutations outlined above (p.R1450* in APC, and the somatic mutations in PIK3CA, KRAS, and ACVR1B) (Supplemental Data S1, S2; Supplemental Table S6). Our data suggest that these mutations were clonal driver mutations, because they were found in two separate tumor locations and thus, must have occurred early in tumorigenesis. Likewise, at least 25 of 27 (92.6%) of the somatic L1 insertions were unique for each FL-L1Hs element in this patient (Fig. 2A; Supplemental Table S2). Moreover, the complete mutational signatures of these elements, which ranged from 7 to 147 mutations, alone could be used to distinguish each individual element from all other FL-L1Hs elements in the patient’s genome (Fig. 2B,C; Supplemental Table S2). Therefore, the complete mutational signatures of these elements, which ranged from 7 to 147 mutations, were unique for each FL-L1Hs element in this patient (Fig. 2A; Supplemental Table S2).

By comparing these FL-L1Hs profiles with the mutation profiles of the L1 insertion along with four additional somatic L1 insertions in the tumor (Figs. 2D, 3A; Supplemental Table S2), we identified a candidate source element (GenBank ID L19088) (Dombroski et al. 1993) that shared an identical pattern of interior mutations with the L1 insertion in APC (Figs. 2D, 3A). This Chr 17 FL-L1Hs element had a profile of 25 mutations that was unique among FL-L1Hs elements in the patient’s genome and also was consistent with this source giving rise to the APC insertion along with four additional somatic L1 insertions in the tumor (Figs. 2D, 3A; Supplemental Table S2). We identified two additional FL-L1Hs source elements on Chromosomes 14 (Chr 14:59160899) and 12 (Chr 12:117814460) that accounted for most of the remaining somatic L1 insertions in the tumor (Table 1; Figs. 2E,F, 3B,C). The unique signatures within these

Figure 2. Interior mutations in 295 FL-L1Hs source elements in the patient’s genome. (A) The total number of mutations in each FL-L1Hs source element is depicted, grouped by pre-Ta and Ta subfamilies (Boisjoly et al. 2000). (B) The Chr 17 FL-L1Hs source element profile is compared to the three closest FL-L1Hs elements in the patient’s genome. Although the three most similar elements have 1S, 14, and 14 mutations in common with the Chr 17 source element, respectively (middle of Venn diagram), they have 18 total differences (Δ) in all three examples. Similar results were obtained with the remaining elements in the patient’s genome (Supplemental Table S2). (C) Mutation frequencies in FL-L1Hs source elements. Individual mutations are plotted by the total number of FL-L1Hs elements in which they are found. A total of 2788 mutations are confined to a single source element (leftmost bar), whereas only a few mutations are shared by the majority of the 295 FL-L1Hs elements in the patient’s genome (right bars). This large collection of singleton mutations, and the profiles that are generated by combining these mutations, has allowed us to identify source elements that generated specific somatic offspring insertions in the tumor (Fig. 3) and also allow us to evaluate the expression of these elements (Fig. 4). (D–F) Mutation profiles for the Chr 17 (D), Chr 14 (E), and Chr 12 (F) source elements. All three of these source elements are heterozygous in our patient’s genome. Differences from the reference L1.3 element (GenBank ID L19088) (Dombroski et al. 1993) are marked in green, red, blue, and yellow and represent mutations in common with the Chr 17 source element, respectively (Supplemental Data S1). (Δ) The total number of mutations in each FL-L1Hs source element is depicted, grouped by pre-Ta and Ta subfamilies (Supplemental Table S2). Moreover, the complete mutational signatures of these elements, which ranged from 7 to 147 mutations, were unique for each FL-L1Hs element in this patient (Fig. 2A; Supplemental Table S2).
source elements allowed us to unambiguously assign 18 of 27 offspring insertions to one of these three source elements (Fig. 3; Table 1; Supplemental Table S1). One additional offspring was mapped to the Chr 12 source element using a 3′ transduction that was associated with the offspring insertion (Fig. 3C; Moran et al. 1999). The eight remaining insertions could not be mapped unambiguously to source elements, because we had limited sequence information for these elements or they did not span regions containing unique mutations.

The interior sequences of the Chr 17, Chr 14, and Chr 12 FL-L1Hs source elements revealed two intact ORFs and also indicated that all three are L1Ta-1d elements, which include the most active “hot” L1s in humans (Supplemental Table S2; Boissinot et al. 2000; Brouha et al. 2003; Beck et al. 2010). All three also are nonreference insertions that are absent from the hg19 human genome reference sequence but are present in the 1000 Genomes Phase III MEI data set (Sudmant et al. 2015). Upon comparing our sequenced elements to additional data sets, we learned that Beck et al. (2010) had previously identified and sequenced the Chr 17 source element. The Chr 17 source element from Beck et al. (2010) was cloned from a Yoruban individual (NA19129, Coriell) and is identical in sequence to the Chr 17 source element that we cloned from our patient (Supplemental Table S2). Beck et al. (2010) previously tested the Chr 17 source element in a cell culture-based retrotransposition assay (Moran et al. 1996) and reported that it is indeed a hot L1, with activity levels of 137% compared to the hot L1.3 control (GenBank ID L19088) (Beck et al. 2010). Therefore, since the Chr 17 source element sequences were identical in both studies, these data indicate that the somatic L1 insertion in our patient’s APC gene was generated by an exceptionally hot L1 source element.

**Expression of L1 source elements in normal versus tumor tissues**

We next sought to better understand how the somatic L1 insertion in APC was generated in tissues that should have repressed the Chr 17 source element. The Chr 17 source element must have been active in normal cells at the earliest stages of tumorigenesis in order to generate the L1 insertion in APC sufficiently early to initiate tumorigenesis. This suggested the possibility that the Chr 17 source element might have been expressed in the normal somatic colon tissues of this patient. To determine whether this was the case, we performed strand-specific RNA-seq analysis and learned that the Chr 17 source element indeed was expressed in both the normal and tumor tissues of this patient (Fig. 4A; Supplemental Table S3; Methods). To perform this analysis, we leveraged the interior mutation profiles of the FL-L1Hs elements (Fig. 2; Supplemental Table S2) to identify source elements that were expressed in the patient’s...
DNA methylation status of the Chr 17 element in the affected tissues (Hata and Sakaki 1997). We therefore sought to determine the promoter region of this element was hypomethylated in both the normal and tumor tissues of this patient (Fig. 4C; Supplemental Methods). Therefore, it appears that the Chr 17 source element may have evaded somatic repression because its promoter was not sufficiently methylated at many CpG sites, including the critical CpG site at position 60 and three other critical sites that were identified in the Hata study (Fig. 4C; Hata and Sakaki 1997; Discussion).

The Chr 17 and Chr 14 source elements are restricted to African and African-derived populations

We next evaluated the population genetics of the Chr 17, Chr 14, and Chr 12 source elements in 26 diverse human populations using the MEI data that we generated for the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015; Sudmant et al. 2015). Interestingly, we found that the Chr 17 and Chr 14 source elements are restricted to African and African-derived populations (Fig. 5; Supplemental Table S4). These data are consistent with the fact that our patient is African American (Methods) and that Beck et al. (2010) sequenced the same Chr 17 source element from a Yoruban individual. In contrast, the Chr 12 source element is found in all continental groups and populations (Fig. 5; Supplemental Table S4). Overall, these data indicate that the source FL-L1Hs content of an individual’s genome is likely to vary considerably depending on the ancestry of the individual, and that differences in L1 content are likely to influence cancer risk. Our patient’s genome had a population-specific hot L1 source element that is absent from most genomes and apparently increased her cancer risk considerably.

Discussion

Next-generation sequencing has revolutionized somatic L1 discovery over the past 6 yr, leading to the identification of thousands of somatic L1 insertions in several types of human epithelial cancers (Iskow et al. 2010; Lee et al. 2012; Solyom et al. 2012; Shukla et al. 2013; Helman et al. 2014; Tubio et al. 2014; Doucet-O’Hare et al. 2015; Ewing et al. 2015; Rodic et al. 2015). The fact that L1 mobilization occurs frequently in human tumors raises the possibility that somatic L1 insertions could act as driver mutations during tumor initiation, progression, and metastasis. However, very few somatic L1 insertions have been recovered in known oncogenes and tumor suppressors, where a clear role in tumorigenesis could be established in the tumors in which they were discovered.

In addition to the Miki et al. (1992) insertion described above, one of the strongest driver candidates identified to date is a somatic L1 insertion in the ST18 gene that led to up-regulation of this patient. Bisulfite sequence analysis of the Chr 17 source element indicated that most of the CpG sites in the promoter region of this element were hypomethylated in both the normal and tumor tissues of this patient (Fig. 4C; Supplemental Methods). Therefore, it appears that the Chr 17 source element may have evaded somatic repression because its promoter was not sufficiently methylated at many CpG sites, including the critical CpG site at position 60 and three other critical sites that were identified in the Hata study (Fig. 4C; Hata and Sakaki 1997; Discussion).
The identification of a second somatic L1 insertion in the Legacy system (Fearon 2011; The Cancer Genome Atlas Network 2012).

The precise role of ST18 in the development of hepatic carcinoma is unknown, and thus it is unclear whether this L1 insertion might have influenced tumor initiation, or instead, later stages of tumorigenesis. Another strong example of a potential L1 driver mutation was identified in the sixth exon of the Pten tumor suppressor gene in a case of uterine corpus endometrial carcinoma (Helman et al. 2014). Because this insertion disrupted a coding exon, the tumor likely had lower levels of PTEN activity. However, once again, the precise role of this somatic insertion in this case of uterine cancer is unclear.

A number of other genes have been hit by somatic L1 insertions multiple times independently in tumors, suggesting that these insertions might also constitute driver mutations (e.g., Helman et al. 2014; Ewing et al. 2015; Rodic et al. 2015). However, most of these insertions map to sites within genes that are difficult to interpret (such as introns), or they map to genes with tenuous connections to the cancers in which they were discovered. Thus, even when L1 driver candidates are identified in a given tumor, it can be difficult to assign these mutations to clear roles in tumorigenesis, particularly in tissues for which the landscape of driver mutations has not been well established. In most tumor types, the temporal order whereby specific gene mutations influence tumorigenesis also has not been well established, making it even more difficult to determine whether L1 can initiate tumorigenesis.

Here, we present a clear example of a somatic L1 driver mutation initiating tumorigenesis through the classic route of CRC progression that has been mapped out by Vogelstein and colleagues (Fig. 6; Kinzler and Vogelstein 1996; Markowitz and Bertagnolli 2009; Fearon 2011; The Cancer Genome Atlas Network 2012). The identification of a second somatic L1 insertion in the APC gene >20 yr after the original Miki et al. (1992) insertion provided us with an opportunity to determine whether such insertions can actually initiate tumorigenesis. We show that both gatekeeper APC alleles are mutated in our patient’s tumor, and this occurred in an MSS genetic background where such mutations would be expected to initiate CRC. A key factor in this study was the availability of relatively inexpensive whole-genome sequencing, which was not available at the time of the Miki et al. (1992) study. This allowed us to determine how the L1 insertion in APC was generated and then how it worked together with other mutations in APC, PIK3CA, Kras, and ACVR1B to drive tumorigenesis (Fig. 6).

Our data support a model whereby an exceptionally hot L1 source element on Chromosome 17 of the patient’s genome evaded somatic repression and produced an off-spring insertion that disrupted the APC gene in a normal colon cell. This hot L1 source element has a mutation in one of the four CpGs that are necessary for repression of the L1 promoter by DNA methylation (Fig. 4C; Hata and Sakaki 1997). Indeed, our data indicate that this source element likely evaded somatic repression because its promoter was not sufficiently methylated (Fig. 4C). Moreover, all four of the CpGs that previously were shown to...
be essential for repression of the L1 promoter were either mutated or hypomethylated in the normal colon and tumor tissues of our patient (Fig. 4C). One intriguing possibility is that the G61A mutation that eliminated one of these critical CpGs also served as an “epimutation” that somehow caused the hypomethylation of the entire promoter region of the Chr 17 element. Similar mechanisms have been observed previously in several types of human cancers. For example, a recent report demonstrated that a 2-bp mutation altered the DNA methylation pattern of the entire RB1 gene promoter in a pedigree of human retinoblastomas (e.g., Quinonez-Silva et al. 2016). If this was the case, then it appears that only a single mutation in L1 might have enabled the transposon to explore an entirely different niche of mutagenesis in somatic cells, thereby expanding upon its ability to mutagenize the germline.

Several previous studies have demonstrated that somatic L1 insertions occur relatively frequently in human CRC (Lee et al. 2012; Solyom et al. 2012; Tubio et al. 2014; Ewing et al. 2015). Somatic L1 insertions also are abundant in adenomatous polyps of the colon, suggesting that L1 mobilization can occur relatively early in the process of CRC (Tubio et al. 2014; Ewing et al. 2015). We now show that somatic L1 insertions can initiate CRC in normal colon cells, and thus, can act at the earliest stages of tumorigenesis that precede adenomatous polyp formation. Once adenomas are formed, it appears that additional somatic L1s can be generated (Tubio et al. 2014; Ewing et al. 2015). Although PIK3CA and KRAS mutations are thought to play key roles in driving the subsequent transition from adenoma to adenocarcinoma (Fearon 2011), it seems unlikely that L1 insertions could play a role in this transition, because such insertions probably have a limited capacity to activate oncogenes. In fact, in our patient’s tumor, PIK3CA and KRAS were mutated by an activating somatic point mutation and a short duplication, respectively, underscoring the idea that L1 works together with other types of somatic mutations to drive human cancers. However, L1 could be envisioned to impact tumor suppressors such as PTEN, e.g., which has been implicated in later stages of CRC (Markowitz and Bertagnolli 2009; Fearon 2011).

Cancer risk from an African-specific hot L1

The hot FL-L1Hs Chr 17 source element, and a second source element on Chromosome 14 that generated most of the remaining somatic L1 insertions in our patient’s tumor, are both restricted to African and African-derived individuals (Fig. 5). Such source elements appear to cause a novel form of ancestry-specific cancer risk. Remarkably, a wide range of epithelial cancers have been diagnosed in our patient’s immediate family, suggesting the possibility that these two FL-L1Hs source elements might be responsible for increased levels of cancer risk in her family. In fact, the patient’s father and seven of 12 siblings had a history of cancer (including vocal cord, breast, liver, prostate, lung, and esophageal cancers). These two FL-L1Hs source elements could potentially be active in a range of somatic tissues, where L1 is normally repressed and hence, may help to drive tumorigenesis in these other tissues. Such elements also may be active in somatic cells of the brain where they could impact neurological diseases such as Alcardi-Goutieres syndrome (Upton et al. 2015) and schizophrenia (Bundo et al. 2014) in an ancestry-specific manner. Finally, these elements would presumably remain active in the germline as well, where they could impact other human traits and diseases in an ancestry-specific manner.

As outlined above, several previous studies have suggested that L1 driver mutations may be rare in human tumors. However, a possible explanation for this observation is that the L1 insertions in these previous studies might have been generated by source elements that were derepressed only after tumorigenesis was well under way. As a consequence, such tumors might be enriched for passenger L1 insertions that played no clear role in tumorigenesis. Our study suggests that more emphasis needs to be placed on L1 source elements that evade somatic repression in normal tissues as a means to identify tumors that harbor L1 driver mutations. Follow-up studies with cancer-prone families like our patient’s family, who carry source elements that can evade somatic repression, may help to establish a firmer link between somatic L1 mutagenesis and human cancers.

Impact of APC mutations

The somatic L1 insertion and the p.R1450∗ stop codon that we identified in APC both map to the mutation cluster region (MCR) of the gene, where the majority of disease-causing mutations have been discovered in sporadic CRCs (Fig. 1C). Most of the mutations that have been identified in sporadic CRCs introduce frameshifts or premature stop codons into the MCR and thus truncate the encoded APC protein within or near this region (Miyoshi et al. 1992). Both of the mutations discovered in our study cause these very same types of disruptions: The F1L396L1 insertion causes a frameshift and premature termination of the encoded protein within the inserted L1, whereas the p.R1450∗ mutation introduces a stop codon that terminates the protein within the MCR. Moreover, the p.R1450∗ mutation occurred at one of the most commonly mutated APC codons in sporadic CRC (Polakis 1995; Fearon 2011). Thus, both of these terminations are consistent with previously identified disease-causing APC alleles in sporadic CRCs. Likewise, our RNA-seq data indicate that both of these mutant alleles were expressed in the tumor, suggesting that these variants escape nonsense-mediated mRNA decay (NMD) and likely produce truncated proteins (Supplemental Table S3).

Interestingly, the original Miki et al. (1992) L1 insertion lies slightly downstream from the MCR but is still very close to this region (codon 1526) (Fig. 1C). The Miki et al. (1992) insertion is in the same orientation as the APC gene, whereas our L1 is in the opposite orientation of APC. However, both L1 insertions cause frameshifts and terminate the protein within the inserted L1 sequences, suggesting that they effectively truncate APC in similar ways. Importantly, we now formally demonstrate that our somatic L1 insertion in APC initiated tumorigenesis. We cannot rule out that the Miki et al. (1992) insertion at codon 1526 also might have initiated CRC, but without knowing the status of the second allele and whether the tumor had stable microsatellites, it remains possible that the tumor was instead initiated by faulty DNA repair (Markowitz and Bertagnolli 2009; Fearon 2011). However, APC mutations are known to contribute to later stages of tumor progression in mutator-initiated CRCs, suggesting that both L1 insertions likely acted as driver mutations, perhaps at different stages of tumorigenesis. The fact that two independent L1 insertions have been identified in APC suggests that gateway tumor suppressors are sensitive targets for L1 mutagenesis, particularly when an inherited hot L1 evades somatic repression.

Methods

Genetic analysis of CRC patient tissue samples

Ten CRC samples with adjacent normal colon tissues were obtained from the Greenebaum Cancer Center Tumor Bank at the
University of Maryland Medical Center under IRB protocol HP-00060447. The following samples were obtained: adenocarcinomas 19079, 19084, 19202, 20085, 20105, 20255; mucinous adenocarcinomas 19120 and 20444; tubular adenoma 20267; and carcinoid 20558. Patient 20444 (the major focus of this study) was a 62-yr-old female African American with a high-grade mucinous adenocarcinoma of the colon stage T4 N0. Genomic DNA was isolated from frozen tissue samples using the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer’s instructions. Total RNA for patient 20444 was isolated from both the normal and tumor tissues with the RNeasy Kit (Qiagen). Normal/tumor pairs were screened for somatic L1 insertions in CRC using the L1-seq method as described previously (Supplemental Methods; Iskow et al. 2010). WGS, RNA-seq, PCR validation of variants, and bisulfite sequencing were performed as described in the Supplemental Methods.

**Microsatellite assessment using MSIsensor**

Using MSIsensor (Niu et al. 2014), we determined the total number of somatically mutated microsatellites in our normal/tumor pairs plus three control TCGA normal/tumor colorectal cancer pairs in which the microsatellite status was known (two MSS samples and one MSI sample) (The Cancer Genome Atlas Network 2012; data not shown). Cases with <3.5% of microsatellites somatically mutated are considered to have stable microsatellites (MSS), whereas cases with >3.5% of microsatellites somatically mutated are considered to have microsatellite instability (MSI).

**MELT analysis**

L1 discovery was performed on normal/tumor pairs using WGS Illumina data and the MELT algorithm with default parameters (http://melt.igs.umaryland.edu) (Supplemental Table S6; Supplemental Data S2; Sudmant et al. 2015). Sites were manually examined using the IGV (Thorvaldsdottir et al. 2013) and also were validated by PCR and/or Sanger sequencing (Supplemental Table SS; Supplemental Methods). Analysis of reference FL-L1Hs elements was conducted by genotyping all reference elements using the MELT-Deletion algorithm (http://melt.igs.umaryland.edu). The MELT-Deletion algorithm genotypes 4645 reference L1Hs elements using L1 sites provided by RepeatMasker (Smit et al. 1996–2010; Jurka et al. 2005). Of these, 19 FL-L1Hs source elements (L1Hs >5900 bp) were absent from the patient’s genome and were excluded from further analysis (Supplemental Table S2; Supplemental Data S3).

**Analysis of FL-L1 Hs source elements in the patient’s genome**

The Chr 14 and Chr 17 FL-L1Hs source elements were amplified with long-range PCR using LA Taq DNA Polymerase (Takara), cloned into plasmids with the TOPO XL PCR Cloning Kit (Invitrogen), and sequenced in triplicate with Sanger sequencing using 20 custom L1 primers and universal flanking primers (Supplemental Table S5). Thirty-one nonreference FL-L1Hs elements that were detected in the patient’s genome with MELT were sequenced with Pacific Biosciences sequencing. Each element was amplified from either the normal sample DNA from patient 20444 or from a 1000 Genomes Project sample that also had the element (see Supplemental Table S5 for a listing of DNAs and primers). Amplification was performed using LA Taq DNA Polymerase (Takara) with the following reaction conditions (Supplemental Table S5): 90 sec at 94°C, followed by 32 cycles for 30 sec at 94°C, for 30 sec at 57°C, and for 8 min 30 sec at 68°C. A final elongation step was performed for 10 min at 68°C. All amplicons were subsequently pooled in approximately equimolar amounts and then purified with AMPure Selective Beads (Beckman Coulter Genomics). The amplicons were prepared for PacBio sequencing using the DNA Template Prep Kit 1.0 (Pacific Biosciences) by following the manufacturer’s protocol. Small fragments and extra adapters were removed from the sample using a BluePippin with a 0.75% agarose cassette (Sage Science). One SMRT cell was sequenced per amplicon pool, using P6C4 chemistry and a 240-min movie on the PacBio RS II.

Following PacBio sequencing, reads were aligned to their genomic locations using BLASR 1.3.1 (Chaisson and Tesler 2012). Reads were clustered based on genomic locations and assembled to form a consensus for each amplicon using version 2.3.0.140936 of ConsenseTools, which is available as part of the SMRTAnalysis software package from Pacific Biosciences (https://github.com/PacificBiosciences/SMRT-Analysis/wiki/ConsenseTools-v2.3.0-Documentation). SNVs and FL-L1Hs subfamilies were then identified using the LINEU tool found in the MELT analysis package (http://melt.igs.umaryland.edu). We used the Sanger sequencing and PacBio sequencing data for FL-L1Hs elements outlined above, together with the FL-L1Hs sequences available in hg19, to develop a profile of 295 potential source elements in the patient’s genome (Fig. 2A,B; Supplemental Table S2). Equivalent results were obtained using FL-L1Hs sequences available in the GRCh38/hg38 build of the human genome reference sequence together with our nonreference FL-L1Hs elements.

**Strand-specific RNA-seq analysis of FL-L1Hs source elements**

Total RNA samples were prepared for strand-specific RNA-seq analysis by first treating the samples with DNase I (Invitrogen). Ribosomal RNA was reduced prior to library construction using the Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat; Illumina). Illumina RNA-seq libraries were prepared from this material with the TruSeq RNA Sample Prep Kit (Illumina) according to the manufacturer’s protocol. Between first and second strand cDNA synthesis, the primers and nucleotides were removed from the samples with NuCleave spin columns (Ambion). The second strand was synthesized with a dNTP mix containing dUTP. Adapters containing indexes 6 nt in length were ligated to the double-stranded cDNA. After adapter ligation, the second strand cDNA was digested with 2 units of Uracil-N-Glycosylase (Applied Biosystems). The DNA was purified between enzymatic reactions, and size selection of the library was performed with AMPure XT beads (Beckman Coulter Genomics). Libraries were sequenced on the HiSeq4000 with paired end runs (one sample per lane) (Supplemental Table S6).

To evaluate FL-L1Hs RNA expression, raw RNA-seq FASTQ files that were generated by strand-specific RNA-seq (Supplemental Table S6) were aligned to the reference FL-L1Hs L1.3 element (GenBank ID L19088) (Dombroski et al. 1993) using Bowtie 2 version 2.2.4 (Langmead and Salzberg 2012). Reads with perfect matches to the FL-L1Hs element profiles outlined in Figure 2 were identified and quantified (Supplemental Table S3). To determine relative expression of all FL-L1Hs copies, the read coverages over element-specific SNV signatures were pooled to obtain a raw expression value for each element, and the mean coverage was reported (Fig. 4; Supplemental Table S3).

**Data access**

Sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under BioProject accession number PRJNA2992328. The BioSample numbers are listed in Supplemental Table S6.

---

**L1 initiates human colorectal cancer**

---

**Genome Research**

www.genome.org

753
Acknowledgments

We thank Brandi Cantarell for help with informatics analysis; Naomi Sengamalay, Sandra Ott, Kelly Klega, Xuechu Zhao, Lisa Sadowszicz, and Luke Tallon for assistance with Illumina and PacBio sequencing; Doris Powell for advice on methylation experiments; Shari Corin for critical evaluation of the manuscript; and the three anonymous reviewers for helpful comments. This work was funded by the following NIH grants: National Cancer Institute (NCI) grant T32 CA154274 (E.C.S.), National Institute of Diabetes and Digestive and Kidney Diseases grant T32 DK067872 (N.T.C.), NCI grant R01 CA077337 (P.M.V.), NCI grant R01 CA166661 (S.E.D.), and National Human Genome Research Institute grant R01 HG002898 (S.E.D.).

Author contributions: E.C.S. performed L1-seq assays, PCR validations, sequencing of somatic L1 insertions, analysis of the two APC alleles, cloning and sequencing of the Chr 17 and Chr 14 FL-L1Hs elements with Sanger capillary sequencing, and bisulfite methylation experiments; E.J.G. aligned WGS raw sequencing data to the reference genome, generated BAM files, and performed MELT analysis, RNA-seq analysis, FL-L1Hs analysis, PCR validations, and bisulfite analysis; A.M. performed non-MEI somatic variant analysis and patient data evaluation; E.J.G. and N.T.C. performed PacBio FL-L1Hs experiments and analysis, E.C.S., P.M.V., and S.E.D. designed methylation experiments; E.C.S., E.J.G., and S.E.D. designed experiments, performed data analysis, prepared data display items, and wrote the manuscript.

References


Received November 12, 2015; accepted in revised form April 19, 2016.