Table S1. Summary of sequence data generated in this study.

For each pooled library in this study, the total number of mapped and unmapped reads is given. Additionally, sensitivity estimates are given for L1 detection from each library based on the reasoning that reference L1 insertions present in one pooled sample must be present in all pooled samples, and so any reference L1s missing from one but present in another must be false negatives.

Table S2. Validated somatic L1 insertion sites.

Definitions for columns present in each table (**Table S2a-e**) are as follows. Chrom: chromosome; Position: best guess of L1 insertion point, Start: leftmost coordinate for aligned reads; End: rightmost coordinate for aligned reads; Strand: polarity in genome, plus or minus; Max Count: maximum number of supporting reads among all libraries in which the site was identified; Max Unique: maximum number of unique alignment positions found among all libraries in which the site was identified; Max Width: maximum distance between leftmost and rightmost peak coordinates among all libraries in which the site was identified; MapScore: mean mappability score across peak based on 50 bp mappability track available through the UCSC Genome Browser; MapQual: mean mapping quality for alignments (Bowtie 2); Gene Annotation: overlapping annotations from UCSC Known Genes; Exon Annotation: Whether the site overlaps an exon annotation based on UCSC Known Genes. Validated Tissues: the combination of tissues in which a given insertion was validated (see Table 1 for definitions), Site ID: an identifier used to refer to a given insertion; Patient ID: indicates in which patient a given insertion was validated; Tumor Type: the tissue of origin of the tumour in which the insertion was identified (TGCT: testicular germ cell tumor). Definitions for columns present in **Tables S2b-e** are as described for **Table S2a**, with the following exceptions and additions: Validated: indicates whether the site was successfully validated by PCR and capillary sequencing; Sample ID: specific sample IDs in which a given insertion was validated; and Patient ID: indicates in which patient a given insertion was validated. Several columns are cohort-specific: For (b): colorectal cancer N/P/T/M refer to whether a given insertion was detected in the normal, polyp, tumor (cancer), or metastasis pool, and the next several columns describe the read support for each pool in terms of read count, number of unique alignments, and peak width. This is done equivalently in (b) describing the overlap between TIP-seq results from pancreatic cancer and L1-seq results: the column labeled TIP-seq indicates whether a given insertion was detected from the orthogonal TIP-seq insertion discovery method (1 = Yes, 0 = No). The column ‘Added by TIP-seq’ indicates whether a given insertion was detected with both TIP-seq and L1-seq (1 = Yes, 0 = No), but PCR validation was not attempted based on the L1-seq data. A “1” in the “TIP-seq” column and a “0” in the same row in the “Added By TIP-seq” column indicates that an insertion was detected by both TIP-seq and L1-seq, and also validated by PCR. When both columns contain the value “1” this indicates that an insertion was found by both TIP-seq and L1-seq but PCR validation had not been previously attempted for that site – thus such sites are considered “Added by TIP-seq”. Validated TIP-seq sites were compared to L1-seq sites plus a 500 bp window up- and downstream of the predicted L1-seq sites using bedtools intersect (Quinlan et al. 2010). The following 5 patient samples were genotyped by both L1-seq and TIP-seq: A43, A55, A57, A82, and A146. (d) A single somatic insertion found in a testicular germ cell tumor. (e) Somatic insertions present in gastric cancers. Sheets S2f-o: Step-by-step detailed validation process and clinical details. Sheets Sf-i: step-by-step validation process of L1s in colorectal cancer cases. Sheets S2k-m: step-by-step validation process of L1s in pancreatic cancer cases. Sheet S2o: step-by-step validation process of L1s in TGCT cases. Sheets S2q-s: step-by-step validation process of L1s in gastric cancer cases. **Sheets S2j, S2n S2p, and S2t show clinical patient data of colorectal, pancreatic, testicular germ cell, and stomach tumor cases, respectively.** The fraction for validated insertions overall is 59%, but this is a slight over-representation of intronic insertions as validation efforts were occasionally focused on insertions most likely to affect genes. Considering all putative somatic insertions with at least two unique reads and a mappability of greater than 0.5, we find 41.0% are intronic, which given that gene set used (UCSC Known Genes) covers 42.7% of the reference genome, is not significantly different from random expectation. We find comparable fractions of intronic events using various cutoffs for calling putative insertions.

Table S3. Putative polymorphic L1 insertions.

Sites present in all libraries/tissues were selected with greater than 50 total reads, and at least 2 unique read alignments, that did not corresponding to a known L1Hs or L1PA element in the hg19/GRCh37 assembly. All column names are defined in the description of **Suppl. Table 2**, with the following addition: Known NonReference indicates whether a given insertion has been detected in one of a number of previous studies. The references are as follows: AE2010 (Ewing et al. 2010), AE2011 (Ewing et al. 2011), AK2014 (Kuhn et al. 2014), CB2010 (Beck et al. 2010), CS2011 (Stewart et al. 2011), DBRIP (Wang et al. 2006), EH2014 (Helman et al. 2014), EL2012 (Lee et al. 2012), JT2014 (Tubio et al. 2014), RI2010 (Iskow et al. 2010), RS2013 (Shukla et al. 2013), SS2012 (Solyom et al. 2012). The sub-tables are as follows: (**a**) putative polymorphic insertions in colorectal cancer cases; (**b**) putative polymorphic insertions in pancreatic cancer patients; (**c**) putative polymorphic insertions in TGCT patients; (**d**) putative polymorphic insertions in gastric cancer patients.

Table S4. Reference L1 insertions

Reference insertions are defined as those with a repeatmasker-derived L1Hs annotation present in hg19/GRCh37 in the proper position (<500 bp from the peak) and orientation for a given cluster of aligned reads. All column names are defined in the description of **Suppl. Table 2**. The sub-tables are as follows: (**a**) Reference L1 insertions in colorectal cancer cases; (**b**) reference L1 insertions in pancreatic cancer patients; (**c**) reference L1 insertions in TGCT patients; (**d**) reference L1 insertions in gastric cancer patients.

Table S5. Proteomics analysis of polyp and normal colon

Protein abundance of the polyp with the highest number of somatic L1 insertions (sample '10') was compared with that of its paired normal colon (sample '8') using mass spectrometry analysis. 457 proteins were at least 2-fold upregulated and 989 proteins were at least 2-fold downregulated in the polyp compared to adjacent normal colon.

Table S6. Counts and fractions corresponding to panels d-f in **Fig. S1**.

For each point on panels d, e, and f in **Fig. S1**, this table shows the percentage of validated insertions associated with that point, and the actual number of validated insertions (i.e. the numerator of the percentage). The interpretation of the columns is identical to the interpretation of the horizontal axis of **Fig. S1d-f**.