Supplemental Material

Supplemental Figures

Supplemental Tables

TranspoSeq

TranspoSeq parses tumor and normal alignment files to identify clusters of unique reads whose pair-mates align to a consensus retrotransposon sequence. To increase sensitivity and prevent filtering out almost half of the genome, we do not discard insertions that fall into all reference retrotransposons, only those that land in reference elements within the same subfamily (i.e., L1PA, L1PB, etc.). Reads with a mapping quality score of 0 are discarded in our analysis. Additionally, pair-mates where both reads align to a retrotransposon consensus sequence are discarded. Supplemental Figure 1 shows the steps of TranspoSeq as described in Online Methods. The current implementation of TranspoSeq uses the Broad Institute's load sharing farm to run parallel processing on each chromosome arm. The alignment parameters listed in Online Methods were used in this study; however, they are input parameters that are easily modifiable for future runs of the pipeline. The code for TranspoSeq will be available at www.broadinstitute.org/cancer/cga/transposeq.

Supplemental Figure 1. Schematic overview of TranpoSeq algorithm. TranspoSeq is a computational framework that takes in paired-end sequencing data and produces a list of annotated putative somatic retrotransposon insertion sites. First, input BAMs are parsed for discordant read-pairs; these pairs are then aligned to a consensus retrotransposon sequence. Pairs with one read aligning to the retrotransposon database and the other aligning to the reference genome with little ambiguity are clustered in the forward and reverse directions. Clusters are overlapped and annotated to support a putative non-reference retrotransposon at the given genomic position. Finally, the read-pairs within each cluster are assembled *de-novo* and the resulting contig is aligned to both the reference and retrotransposon database to annotate the

element that was inserted. Events with strong evidence that pass filtering criteria are retained and classified as somatic or germline.

TranspoSeq Performance

We used a simulated alignment file with retrotransposons elements computationally inserted in order to assess TranspoSeq's sensitivity and specificity. Simulated alignment data were created by inserting 226 full length L1HS and 732 *Alu*Y consensus sequences into a 22Mb region of chromosome 20 (chr20: 2500000-24500000) of the human reference hg19. This region has comparable GC (40.8%), simple repeat (1.63%), large repeat (49.03%), segmental duplication (3.12%) and microsatellite (0.06%) content as the rest of the genome and was chosen arbitrarily to represent typical genomic sequence. The SAMTOOL's package wgsim (https://github.com/lh3/wgsim) was used to create a simulated BAM file with read length 100bp, fragment length 500bp, 20x coverage and default values for all other parameters. TranspoSeq was able to correctly identify 225/226 L1 and 730/732 *Alu* elements with no false positive calls. A second simulated dataset was created by inserting 1000 5'-truncated L1HS elements of varying lengths: 100 elements each of lengths ranging from 40bp to 6000bp. TranspoSeq's ability to detect both germline and somatic insertions of 100bp L1HS elements drops to 60% sensitivity, and continues to decrease for elements <80bp (Supplemental Fig. 2).

To computationally assess the performance of TranspoSeq, we compared our findings to those of Lee et al. (2012) on the colorectal sample TCGA-AA-3518. The method used by Lee et al. (2012) to analyze whole-genome data for novel retrotransposition was developed concurrently with, but independent of, TranspoSeq (as TranspoSeq was first presented to the public in 2011). Both methods use a similar approach, but differ in the stringency of filtering criteria, such that TranspoSeq's is more conservative than Lee et al. (2012) requiring more read-pair and split-read evidence to support any call. Of the 92 high-confidence somatic retrotransposon insertions we identify in TCGA-AA-3518, 63 insertions are common to both studies (TranspoSeq identifies

60% of the putative events reported by Lee et al.; conversely, Tea detects 68% of the somatic insertions discovered by TranspoSeq). We find most (~90%) of the PCR validated calls reported in Lee et al. (2012). Their validation protocol verified the existence of only the 3' L1 insertion junction, so it remains unclear whether all of these events are true bona fide retrotransposons insertions.

Finally, we swapped tumor and normal BAM file labels and re-ran TranspoSeq on a random subset of five HNSC samples. We find no retrotransposon insertions that pass our filtering criteria and are unique to the normal sample, implying that the somatic insertions we detect are likely tumor-associated and not due to normal variation.

Supplemental Figure 2. Sensitivity of TranspoSeq to identify germline (blue) and somatic (red) L1HS insertions of different lengths. 100 elements of each length were inserted into a simulated BAM file and the fraction of elements identified by TranspoSeq was recorded.

Non-reference germline retrotransposon insertions

Non-reference germline retrotransposon insertions were identified as putative insertion events present in both tumor and matched normal sample. The number of non-reference germline retrotransposon insertions per individual was on average 880 +/- 275. All putative retrotransposon insertion events were assessed for presence of target site duplications (TSDs) and endonuclease consensus sites. TSDs were determined by the distance between forward and reverse clipped reads whenever available. The distribution of TSD lengths in somatic and germline insertions differed significantly, KS-test of p-value < 2.2e-16. Endonuclease consensus

sites were determined from assembled contig sequences for both strand directions.

Supplemental Figure 3. Non-reference germline retrotransposon insertions across individuals. *(A)* Distribution of TSD or microdeletion length at breakpoints of germline insertions. *(B)* Sequence logo of insertion motifs for germline retrotransposon insertions. *(C)* Length of non-reference germline L1 elements. *(D)* Number of individuals in which each known, or previously annotated (left panel), and novel (right panel) germline retrotransposon insertion is found.

Distribution of somatic retrotransposon insertions

To assess the genome-wide distribution of retrotransposon insertions, we determined for each chromosomal arm, whether there was an enrichment of insertion events given the length of the arm. These fold-enrichments were determined by the ratio of the number of insertion events in a chromosome arm divided by total number of insertion events to the length of chromosome arm divided by length of human genome. To assess the difference between somatic and germline events, a Fisher's exact test was performed using the fisher.test R function.

Germline retrotransposons are distributed evenly across the genome (data not shown). The distribution of retrotransposon insertions across chromosomal arms significantly differs between germline and somatic events (Wilcoxon p=3.706e-08).

Supplemental Figure 4. *(A)* Barplots displaying the number of somatic retrotransposons insertions per individual analyzed, grouped by tumor type. These data are whole-genome sequences from 200 individuals collected and sequenced through The Cancer Genome Atlas, across 11 tumor types: lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian carcinoma (OV), rectal adenocarcinoma (READ), colon adenocarcinoma (COAD), kidney clear cell carcinoma (KIRC), uterine corpus endometrioid carcinoma (UCEC), head and neck squamous cell carcinoma (HNSC), breast carcinoma (BRCA), acute myeloid leukemia (LAML), and glioblastoma multiforme (GBM). *(B)* Table of the mean number of somatic retrotransposons insertions per individual across tumor types. *(C)* Genomic distribution of somatic retrotransposon insertions. Positions of somatic retrotransposon insertions (red) overlayed on human chromosomes.

Somatic retrotransposon insertions across tumors

We find 133 somatic events that do not have the expected TPRT TSD lengths, comprising 16% of all somatic insertions identified. We sought to determine a sequence motif enrichment separately in the two groups of somatic events: one group with the expected TSD length and one group lacking this. We find that the set of candidate insertions lacking TSDs (defined as no TSD or with TSD of length <=2bp) does not display the canonical L1 endonuclease target sequence, or any enriched sequence motif. These target sites do contain a slight GC bias (Supplemental Figure 5) with a one-sided KS p-value of 8.188E-05 when compared to the GC content of the set of insertions with expected TSDs. The observation of a possible additional class of somatic events was also noted in Lee et al. (2012) and Solyom et al. (2012), where they describe a similar peak around a TSD length of 0-2bp, consistent with L1 endonuclease-independent somatic insertion.

We find 184 inversion events in the set of candidate somatic insertions and 27 full-length L1 insertions, comprising approximately 3% of all somatic events. Finally, there does not appear to be any correlation between number of somatic insertions and age of patient at tumor diagnosis in our cohort (Spearman correlation of 0.09).

Supplemental Figure 5. GC content (%GC) of somatic target sites for events with canonical TSD lengths and the set of events lacking TSDs.

We compared somatic retrotransposition sites with the 73 annotated common fragile sites across the genome from Fungtammasan et al. (2012). Of the 810 somatic events, 130 (16%) fall in a known fragile site. Of the 286 genes with a somatic retrotransposon insertion, 60 (21%) are

common fragile site (CFS) genes. Similarly, 15% of germline retrotransposon insertions fall in common fragile sites and 18% of genes with germline insertions are CFS genes. However, both germline and somatic insertion genes contain more CFS than expected from all RefSeq genes (Fisher's exact p<2.16E-16, Supplemental Fig. 6A).

We sought to determine whether longer genes have a higher propensity for retrotransposon insertions. We compared the lengths of genes with germline and somatic retrotransposon insertions to the distribution of all genes and found that indeed, somatic insertions tend to target (or be tolerated) in longer genes (Supplemental Fig. 6B).

There is evidence that both sense and antisense L1 insertions can attenuate gene expression(Han et al. 2004). We find that about half of the retrotransposons somatically inserted in cancer are present in the sense orientation with respect to the disrupted gene, consistent with previous findings for disease-causing L1 insertions(Chen et al. 2005), but significantly different from germline retrotransposon insertions (Fisher's exact p=1.4E-10, Supplemental Fig. 6C).

Supplemental Figure 6. *(A)* Proportion of germline and somatic retrotransposon insertions that land in common fragile sites as determined by Fungtammasan et al. (2012) (Fungtammasan et al. 2012). *(B)* Length of genes that harbor germline and somatic retrotransposon insertions as compared to lengths of all RefSeq genes. *(C)* Proportion of germline (left) and somatic (right) retrotransposon insertions that land in the same orientation (sense) as the gene in which they are inserted.

Biological Processes associated with somatic retrotransposition

We used the Gene Ontology to assess whether any biological processes are enriched in the genes found to harbor somatic retrotransposon insertions in our study. Cell adhesion was highly enriched in this set as well as the neuronal synapse cellular component (data not shown) possibly due to the bias toward large genes. Many of these genes are frequently mutated in cancer, but have been suggested as passenger events due to their size and propensity for mutation (Lawrence et al. 2013).

Supplemental Figure 7. Gene Ontology (GO) Analysis. GO Biological Processes enriched in genes with somatic retrotransposon insertions.

Association with other genomic events

Many of the correlations performed between retrotransposon clusters and other genomic features were inconclusive due to the small sample size within each tumor type. In addition to *TP53* and *CDKN2A* associates, however, the HNSC samples also showed significant differential expression of *PRSS12* and *ALPK1* between samples with high and low rates of retrotransposition.

Microsatellite Instability

Eight samples, including three LUSC, two HNSC, one UCEC, one LUAD, and one COAD, exhibit an extremely high amount of somatic retrotransposon insertion events (>30

events). Although the lung tumors couldn't be assessed, the other four samples all have high levels of microsatellite instability (MSI). MSI status, however, does not predict somatic retrotransposon insertion load, as many MSI-high tumors do not have any somatic insertions.

TranspoSeq-Exome

TranpoSeq-Exome gathers split reads identified by BWA that are at least 10bp and align to the database of consensus retrotransposon sequences. Post-processing filtering leaves only putative insertions that have more than 4 reads supporting them. Since poly-adenylation tracts will not align significantly to the database, we allow events with evidence from only the forward or reverse direction. Manual inspection enabled us to find several events with evidence from both directions, including the *PTEN* event. In cases where support is only captured in one direction, we cannot distinguish between a possible rearrangement with a retrotransposon and a retrotransposon insertion.

Supplemental Figure 8. Schematic of TranspoSeq-Exome workflow. TranspoSeq-Exome consists of three steps. Get Reads parses tumor and normal BAM files for split reads identified by BWA that are at least 10bp in length. These clipped portions of the read are aligned to the database of consensus retrotransposon sequences using blastn. Reads where the clipped portion aligns with a BLAST e-value less than 2E-07 are gathered for the next step. Process Reads takes these reads and clusters them by read strand in the forward and reverse direction, then overlaps these clusters. Here, we keep all clusters even if there is no overlapping cluster identified in the opposing direction. Assemble Reads gathers the identified split reads and assembles them *de*

novo using Inchworm to get longer potential contigs and then aligns these contigs back to the database of consensus retrotransposons.

Experimental Validation

We further validated putative retrotransposons insertions identified in the 200 samples discussed in this manuscript. We designed PCR primers to span each 5' and 3' insertion junction using Primer3 {Rozen:2000wg} for each target; a primer set consisted of one unique primer and the other hybridized to the putative retrotransposon at its predicted 3' or 5' position. We designed primer sets for 51 targets: 4 germline events, 42 somatic events from whole-genome data (including an SVA insertion and several full-length L1HS inserts), and 5 somatic events from exome data.

All four predicted germline transpositions were validated. Of the 47 predicted somatic retrotranspositions, PCR-based validation showed:

Two-sided somatic validation (5' and 3' junctions support insertion): 32

One-sided somatic validation (5' or 3' junction supports insertion): 7

Possibly germline transposition (#reads in normal >= #reads in tumor/100): 2

Failure of amplification: 6 (amplification of 6 putative retrotranspositions from lung adenocarcinoma sample LUAD-38-4630 did not yield any amplicons in either tumor or normal sample; this failure may represent false positive calls or a technical failure for the new DNA aliquot obtained for this sample).

In summary, we find 39/47 (83%) of predicted somatic insertions have experimental evidence for a transposition event by amplification of either 5' or 3' junctions in the tumor, but no junctional amplification from the matched normal sample. Moreover, 32 of 47 (68%)

predicted somatic insertions have evidence for amplification of both 5' and 3' junctions in the tumor sample and no evidence in the matched normal. Finally, 2/47 putative somatic retrotranspositions have some evidence of the insertion in the matched normal. These 'possibly germline' events are defined as an event in which the number of reads supporting the insertion in the normal is greater than $1/100^{th}$ of the number of supporting reads in the tumor. We provide sequence analysis of the junctions of all 45 candidate insertions in the validation set that produced amplicons (Supplemental Table 1).

Supplemental Figure 9. Site-specific PCR confirms presence of retrotransposon insertion in *PTEN* exon. *(A)* Diagram of PCR primer design for experimental validation of predicted retrotransposon insertions, top panel; capillary gel electrophoresis for amplicons from 5' junction, from 3' junction, and from primers spanning the entire insert for tumor (T) and matched normal (N) samples of an individual with endometrial carcinoma. (B) Illumina sequencing reveals a 5'-truncated L1HS insertion, with TSDs flanking the insertion, a canonical TTAAA target site sequence, and a ~37bp polyA tail.

Supplemental Tables

Supplemental Table 1. PCR primers and sequencing information.

Supplemental Table 2. Germline non-reference retrotransposon insertions.

Supplemental Table 3. Somatic retrotransposon insertions across all cancer types.

Supplemental Table 4. 3' transduction events.

Supplemental Table 5. Somatic retrotransposon insertions from capture data.

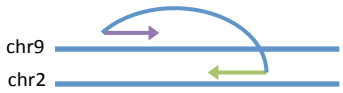
Supplemental Table 6. Consensus retrotransposon sequences.

Supplemental Figure 1

Step 1: Align Reads

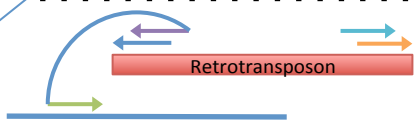
1. Obtain BAM

2. Parse out discordant read-pairs

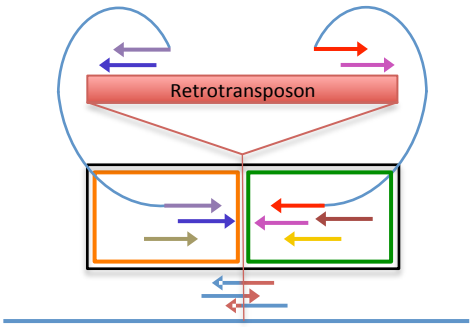


3. Align reads to consensus retrotransposon database

4. Retrieve mates that align uniquely to the genome



Step 2: Process Reads



1. Cluster reads
Identify forward and reverse clusters

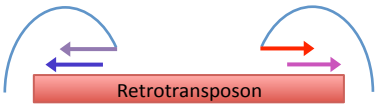
2. Overlap forward and reverse clusters
Identify candidate insertion regions

3. Add split read information

4. Filter and annotate candidate regions

Step 3: Assemble Reads

1. Retrieve mates of unique reads supporting insertion

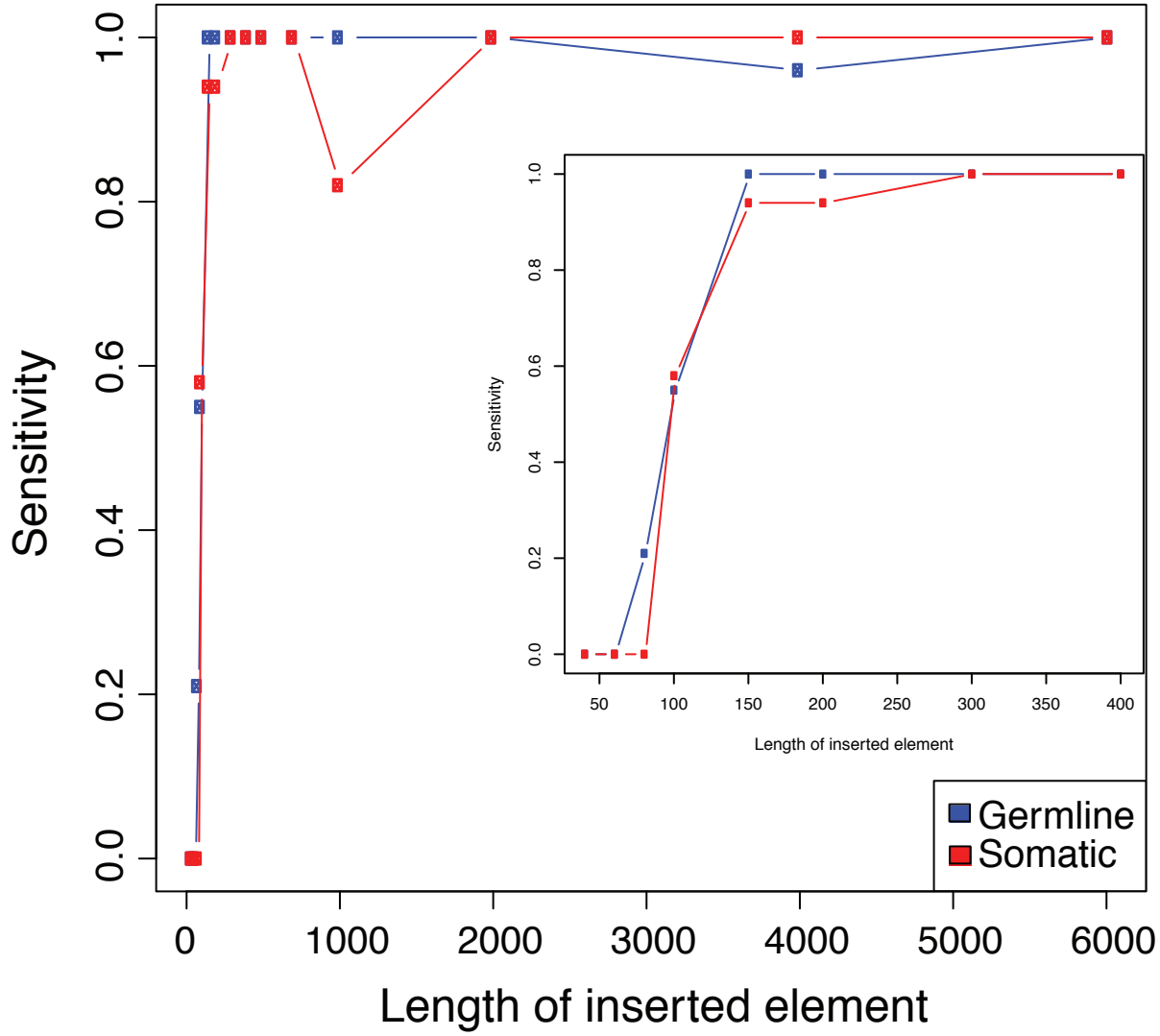


2. Assemble *de-novo*

3. Align assembled contigs to retrotransposon database

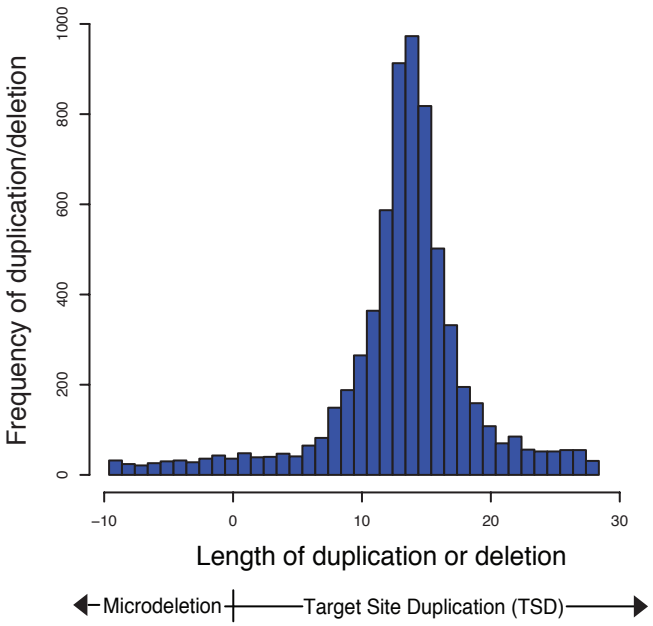
4. Add assembly information, filter further, and output somatic and germline retrotransposon insertion

Supplemental Figure 2

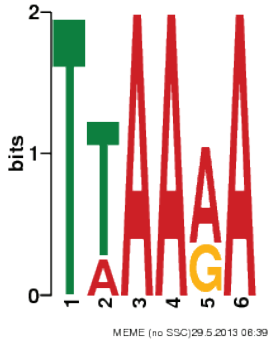


Supplemental Figure 3

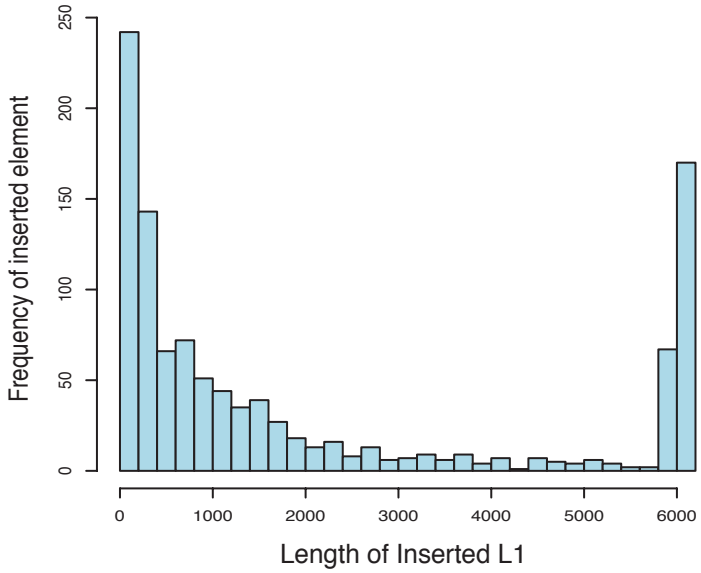
A



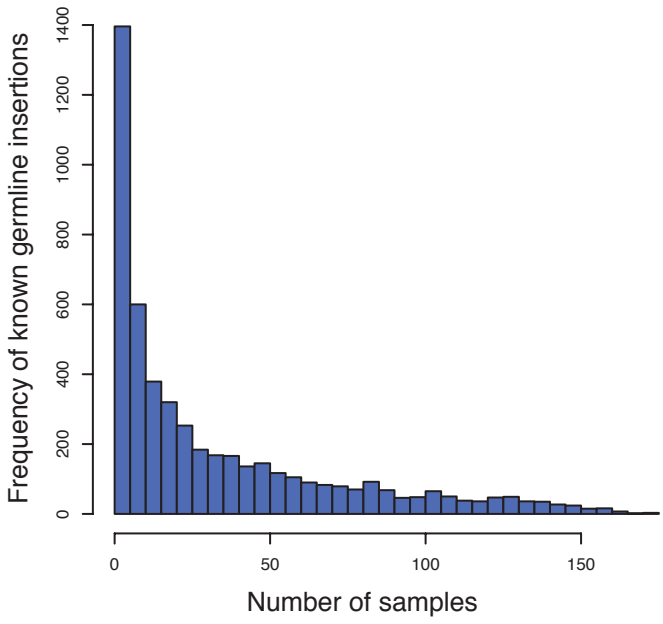
B

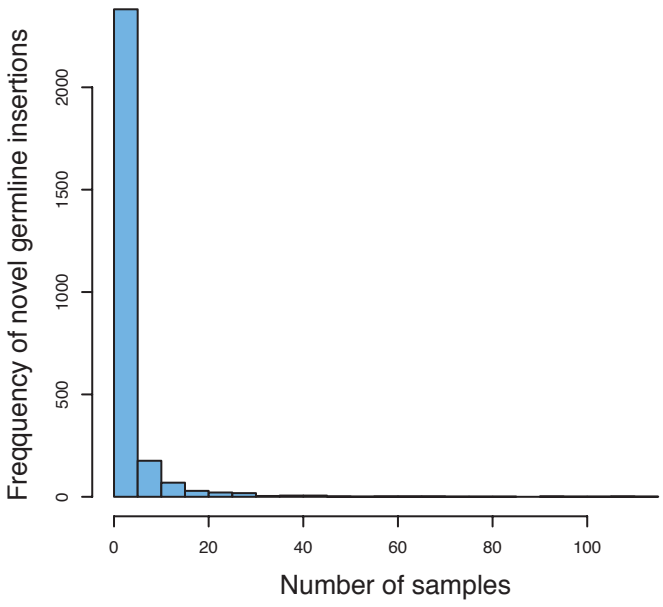


C



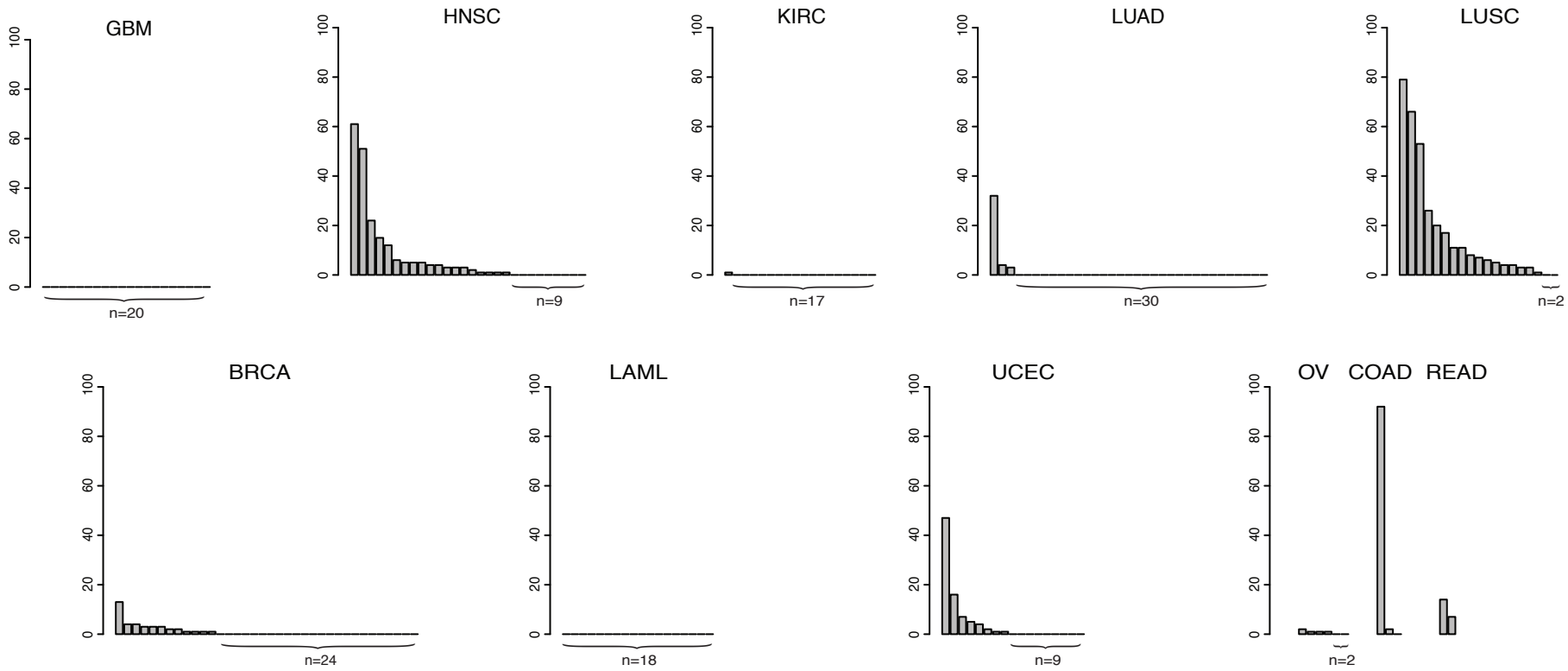
D





Supplemental Figure 4

A

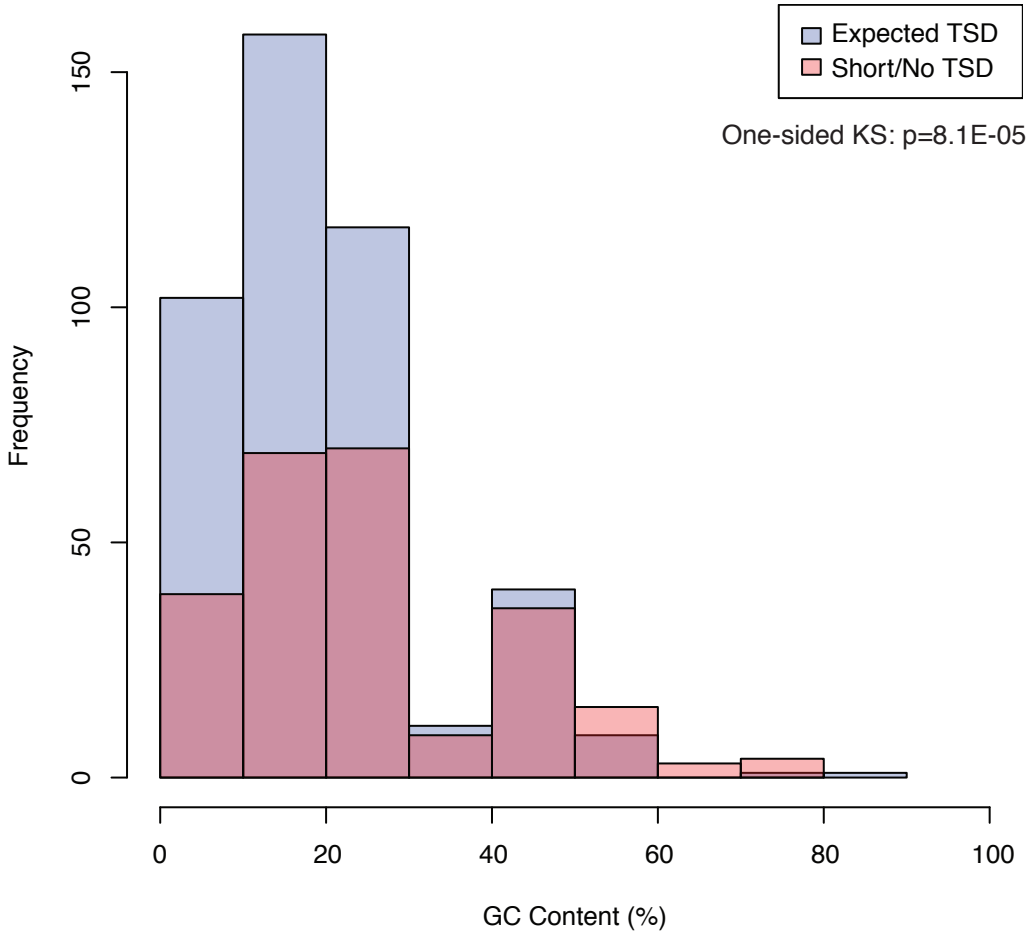


B

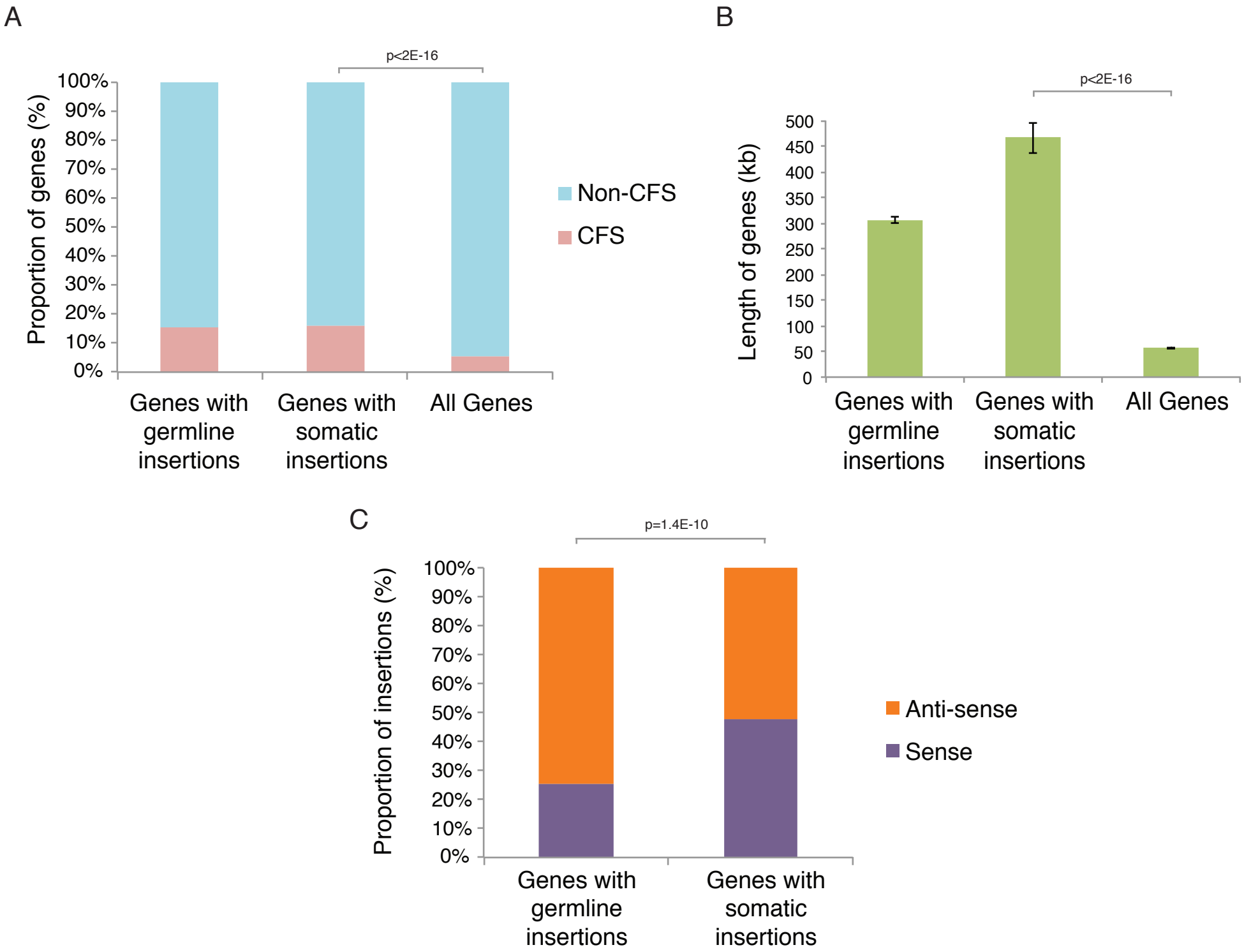
Tumor Type	Mean number of somatic retrotransposon insertions
BRCA	1.06 (n=36)
COAD	31.3 (n=3)
GBM	0.00 (n=20)
HNSC	7.32 (n=28)
KIRC	0.06 (n=18)
LAML	0.00 (n=18)
LUAD	1.18 (n=33)
LUSC	17.1 (n=19)
OV	0.83 (n=6)
READ	10.5 (n=2)
UCEC	4.88 (n=17)

C

Supplementary Figure 5



Supplemental Figure 6



Supplemental Figure 7

GO term	Description	P-value	FDR q-value
GO:0022610	biological adhesion	2.34E-07	2.71E-03
GO:0007155	cell adhesion	2.34E-07	1.36E-03
GO:0016337	cell-cell adhesion	1.15E-05	4.42E-02
GO:0007156	homophilic cell adhesion	2.87E-05	8.29E-02
GO:0007411	axon guidance	3.06E-05	7.09E-02
GO:0044763	single-organism cellular process	3.98E-05	7.67E-02
GO:0008038	neuron recognition	5.78E-05	9.56E-02
GO:0040011	locomotion	7.29E-05	1.05E-01
GO:0006935	chemotaxis	9.75E-05	1.25E-01
GO:0042330	taxis	9.75E-05	1.13E-01
GO:0044699	single-organism process	1.94E-04	2.04E-01
GO:0021942	radial glia guided migration of Purkinje cell	1.99E-04	1.92E-01
GO:0031175	neuron projection development	4.24E-04	3.77E-01
GO:0048010	vascular endothelial growth factor receptor signaling pathway	5.29E-04	4.37E-01
GO:0007158	neuron cell-cell adhesion	5.58E-04	4.31E-01
GO:0061364	apoptotic process involved in luteolysis	5.91E-04	4.28E-01
GO:0021932	hindbrain radial glia guided cell migration	5.91E-04	4.02E-01
GO:0035335	peptidyl-tyrosine dephosphorylation	6.22E-04	4.00E-01
GO:0006198	cAMP catabolic process	9.04E-04	5.51E-01
GO:0007165	signal transduction	9.34E-04	5.41E-01
GO:0007268	synaptic transmission	9.99E-04	5.51E-01

Supplemental Figure 8

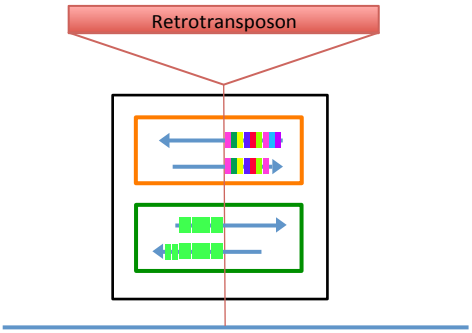
Step 1: Align Reads

1. Obtain BAM

2. Parse out soft-clipped reads

3. Align clipped sequences to consensus retrotransposon database

Step 2: Process Reads



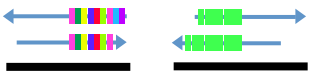
1. Cluster clipped positions
Identify forward and reverse clusters

2. Overlap forward and reverse clusters
Identify candidate insertion regions

3. Filter and annotate candidate regions

Step 3: Assemble Reads

1. Assemble reads *de-novo*



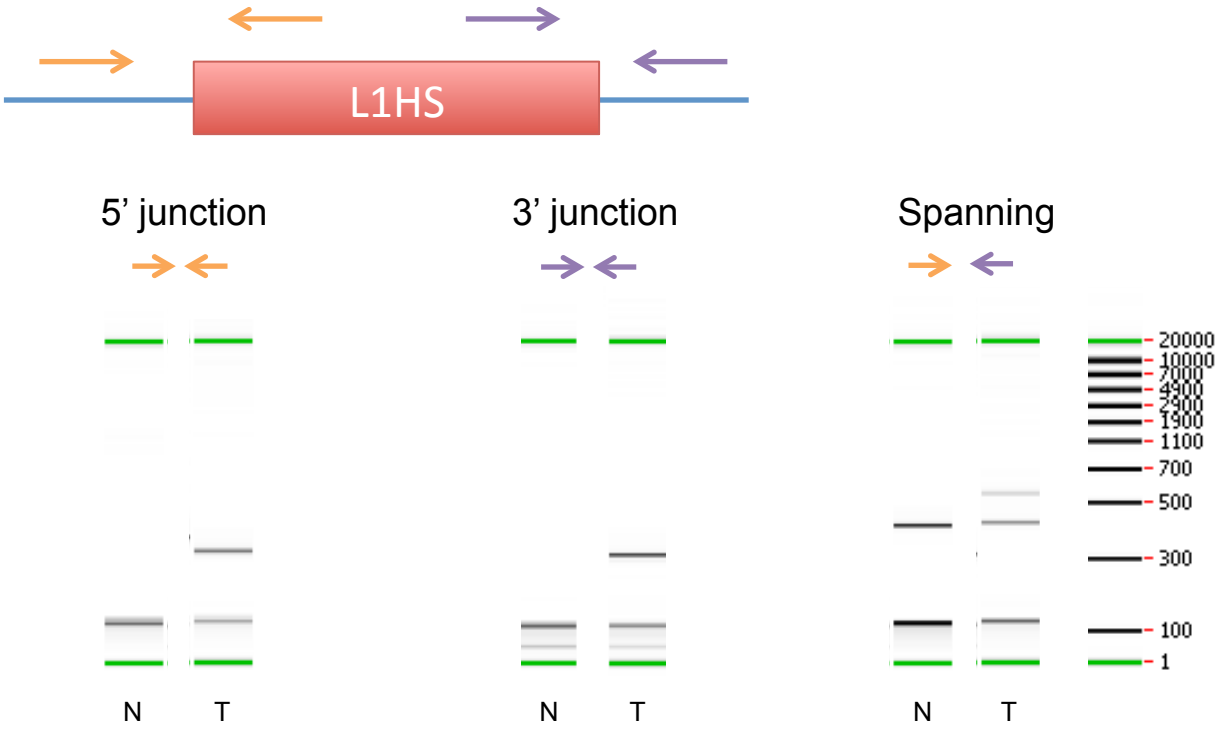
2. Align assembled contigs to retrotransposon database

3. Add assembly information, filter further, and output somatic and germline retrotransposon insertions

somatic one-sided	somatic two-sided	germline one-sided	germline two-sided

Supplemental Figure 9

A



B

GCGCTATGTGTATTATTATAGCTACCTGTTAAAGAATCATCTGGATTATAGACCAGCATGACAAAAG
TATACATATGTAACTAACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAGAAAAAGAATCATCTGGATTATAGACCAGTGGCACTGTTGT
TTCACAAGATGATGTTTGAAACTATTC

112bp 5’-truncated L1HS

TSD