Somatic Retrotransposition in the Cancer Genome

by

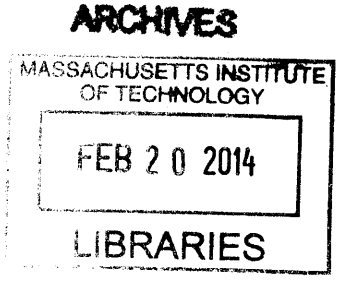
Elena Helman

Sc.B. Computational Biology
Brown University, 2009

SUBMITTED TO THE DIVISION OF HEALTH-SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2013



Signature of Author: _____
Harvard-MIT Program in Health Sciences and Technology
September 25, 2013

Certified by: _____
Matthew Meyerson, MD, PhD
Professor of Pathology and Medical Oncology
Thesis Supervisor

Accepted by: _____
Emery N. Brown, MD, PhD
Professor of Computational Neuroscience and Health Sciences and Technology
Director, Harvard-MIT Program in Health Sciences and Technology

Somatic Retrotransposition in the Cancer Genome

by
Elena Helman

Submitted to the Harvard-MIT Program in Health Sciences and Technology
on September 25, 2013 in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Cancer is a complex disease of the genome exhibiting myriad somatic mutations, from single nucleotide changes to various chromosomal rearrangements. The technological advances of next-generation sequencing enable high-throughput identification and characterization of these events genome-wide using computational algorithms.

Retrotransposons comprise 42% of the human genome and have the capacity to "jump" across the genome in a copy-and-paste manner. Recent studies have identified families of retrotransposable elements that are currently active. In fact, retrotransposons constitute a major source of human genetic variation, and somatic retrotransposon insertions have been implicated in several cancers, including an insertion into the *APC* tumor suppressor in a colorectal tumor. Because of the highly repetitive nature of these elements, however, the full extent of somatic retrotransposon movement across cancer remains largely unexplored.

To this end, we developed TranspoSeq, a computational framework that identifies retrotransposon insertions from paired-end whole genome sequencing data, and TranspoSeq-Exome, a tool that localizes these insertions from whole-exome data. TranspoSeq identifies novel somatic retrotransposon insertions with high sensitivity and specificity in simulated data and with a 94% validation rate via site-specific PCR. Next, we applied these methods to whole-genomes from 200 tumor/normal pairs and whole-exomes from 767 tumor/normal pairs across 11 tumor types as part of The Cancer Genome Atlas (TCGA) Pan-Cancer Project. We discover more than 800 somatic retrotransposon insertions primarily in lung squamous, head and neck, colorectal and endometrial carcinomas, while glioblastoma multiforme and acute myeloid leukemia show no evidence of somatic retrotransposition. Moreover, many somatic retrotransposon insertions occur in known cancer genes. TranspoSeq-Exome uncovers 35 additional somatic retrotransposon insertions into exonic regions, including an insertion into an exon of the *PTEN* tumor suppressor in endometrial cancer. Finally, we integrate orthogonal genomic and clinical data to characterize features of retrotransposon insertion and samples that exhibit extensive somatic retrotransposition.

We present a large-scale, comprehensive analysis of retrotransposon movement across tumor types using next-generation sequencing data. Our results suggest that somatic retrotransposon insertions may represent an important class of tumor-specific structural variation in cancer and future studies should incorporate this form of somatic genome aberration.

Thesis Supervisor: Matthew Meyerson, MD, PhD
Title: Professor of Pathology and Medical Oncology

Acknowledgments

This thesis was completed thanks to the support of my wonderful professors, colleagues, friends and family, and I thank each and every one for the role they played during my years here.

I'd like to express my utmost gratitude toward my supervisor, mentor, and steadfast supporter, Dr. Matthew Meyerson. His intuition for cancer biology, scientific rigor, and the values and collegiality he displays are virtues to which I aspire. Despite a large lab with myriad responsibilities, Matthew found time to meet with me, review my papers, and discuss life goals. He wasn't afraid to fight for what's right and for that I am forever grateful. Thanks for having my back.

I'm greatly appreciative to my thesis committee members, Dr. Isaac Kohane and Dr. Gad Getz, for their scientific input, guidance in my project and in my career. Their computational insights and global perspective were critical to the progression of my research work.

I would like to thank everyone in the Meyerson lab and the Broad Cancer Genome Analysis group: Alice Berger for teaching me how to use a pipette; Mike Lawrence for being the all-knowing resource to whom I could turn with any question; Chip Stewart for the lengthy discussions on the gory details of retrotransposition; Carrie Sougnez for all the help with coordinating validation efforts; Peter Hammerman for answering any question within two minutes of my email; Angela Brooks for all the support in and outside of lab; Ami Bhatt for her clinical insights; the members of the Meyerson band for rocking out with me; and Scott Carter, Aaron McKenna, Alex Ramos, Marcin Imielinski, Chandra Pedhar, Bryan Hernandez, ChengZhong Zhang for providing a fun yet scientifically invigorating work environment.

The support of the HST program – its students, professors, and administrators, has been an invaluable resource to me throughout graduate school. Big thanks to: Alal Eran for being the best "BIG buddy" any girl could ask for; for the unending encouragement and love, and for sitting at the Broad with me for hours on end, with notebook in hand, interpreting my jumbled thoughts about how to improve my algorithm; and Alexandra German for putting up with me in classes, and in our apartment.

Finally, I'd like to thank my parents Vera and Alex Helman, and my older brother and sister-in-law, Igor and Ainsley Helman, for their unconditional love and support. They are my role models and I wouldn't be where I am today without their pushing me to achieve my fullest potential. Thank you for leaving the lives you knew behind and bringing me to this country so that I could have all opportunities available to me. Thanks for the last-minute editing, being there through all of my ups and downs, and the homemade meals and laundry brought to my door while I was writing this thesis!

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

Cancer is a disease of the genome. It is characterized by the accumulation of mutations in a cell's DNA that leads to uncontrolled proliferation, invasion into nearby tissue, and distant metastasis. There are at least 200 known forms of cancer and many more subtypes. In 2008, cancer was the cause of 7.6 million deaths worldwide and 12.7 million new cancer cases (Ferlay J, 2008). As a result of rapid advances in sequencing technology, insight into the cancer genome has amounted at an unprecedented rate. Improvements in elucidating the genomic alterations leading to tumorigenesis will ultimately result in improved targeted cancer therapeutics and diagnostics.

1.1 The Cancer Genome

The cancer genome is enormously complex (Meyerson et al. 2010). Individual cells continuously acquire genetic variation by random mutation, and the cell that acquires mutations that allow it to proliferate autonomously forms the basis for the clonal tumor cell population. Somatic alterations, present in the tumor cells but not in an individual's germline, typically include nucleotide substitutions, small insertions and deletions, copy number alterations and genomic rearrangements. Cancer genomes vary wildly in the number and types of mutations they harbor, with some carrying over 100,000 point mutations and hundreds of somatic rearrangements and others with relatively few (Stratton et al. 2009).

Driver vs. passenger mutations

Comprehensively cataloguing somatic mutations in the cancer cell will allow investigators to tease out "driver" mutations, those that confer selective growth advantage to the cancer cell,

from "passenger" events, those that do not confer growth and merely happened to be present in an ancestral cell when it acquired a driver. Most somatic point mutations in cancer are passengers (Greenman et al. 2007). A central goal of cancer genome analysis is identifying "cancer genes" that harbor a driver mutation and are causally implicated in oncogenesis (Stratton et al. 2009). It has been suggested that adult epithelial cancers like breast, colorectal and prostate require at least 5-7 driver-like events, whereas hematological cancers require fewer (Miller 1980), but recent studies have challenged these findings (Beerenwinkel et al. 2007). Specific and recurrent genomic abnormalities, such as the Philadelphia chromosome (Nowell & Hungerford 1961), are associated with particular tumor types. Moreover, subclasses of cancer can be defined on the basis of certain mutations and can more accurately determine prognosis and course of treatment for the patient's specific tumor.

Next-generation sequencing

The advent of second-generation sequencing technologies has vastly increased our knowledge of the cancer genome. Next-generation sequencing involves the shearing of genomic DNA and parallel sequencing of the resulting short fragments, followed by computational assembly of the overlapping sequences such that each base in the reference genome is covered several times by a sequence fragment. Sequencing both ends of the segment, known as paired-end sequencing, facilitates the accurate alignment of reads to the reference human genome. The number of bases that can be sequenced for a given cost has more than doubled every year, proving twice as fast as Moore's law for semiconductors (Figure 1-1) (Meyerson et al. 2010; Mardis 2012). In 2012, sequencing cost 50¢ per megabase using Illumina's MiSeq machine (Loman et al. 2012).

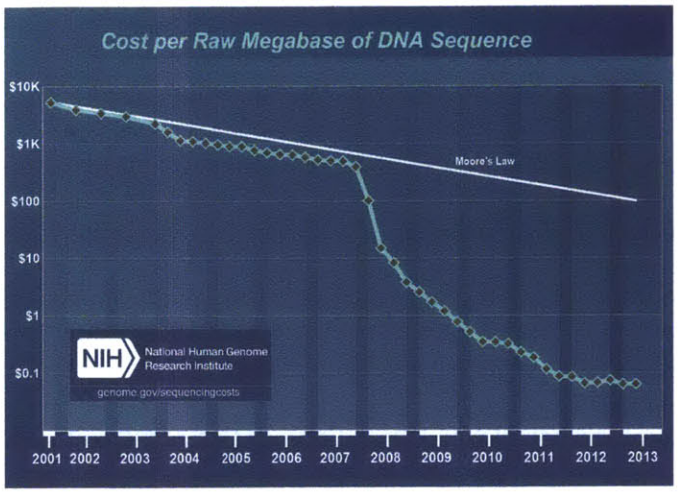


Figure 1-1 Decreasing cost of sequencing.
Cost to sequence a Megabase of DNA through the years (connected points) compared to hypothetical data reflecting Moore's law (solid line). Image from (Wetterstrand, 2013).

Challenges of somatic mutation detection in cancer

Although there is a flood of sequencing data being generated, efficient and accurate analysis of these data leading to relevant biological insights forms a bottleneck. Sequencing errors and artifacts, as well as alignment to an imperfect and incomplete reference genome are just some of the obstacles the computational biologist faces, in addition to pure hardware considerations such as data storage and compute power to maneuver large files. Moreover, challenges specific to detection of somatic mutation include the level of purity of the tumor sample – or how much normal tissue is excised and sequenced with the tumor, as well as tumor heterogeneity – or how many different subclonal populations of cells are present within the tumor sample. Higher coverage allows for the detection of mutations present at smaller fractions in the tumor population and will elucidate some tumor heterogeneity, but regional heterogeneity across different sections of the tumor (Gerlinger et al. 2012) will have to be addressed through enhanced experimental design. Mutations must be evaluated on a sample-specific background,

that is, against a normal sample from the same individual, and preferably from adjacent tissue rather than blood. However, normal samples must also be assessed for field effect, or the existence of histologically and genetically abnormal tissue beyond a neoplastic area (Chai & Brown 2009). Finally, genomic DNA integrity is often poor for tumor samples because biopsies are small and commonly formalin-fixed or paraffin-embedded to optimize the resolution of microscopic histology (Meyerson et al. 2010). However, increased sequencing coverage and sample sizes will help overcome some of these challenges associated with somatic alteration detection (R. K. Thomas et al. 2006) and ultimately enable accurate identification of the myriad genomic aberrations present in the tumor cell.

A form of genome alteration that remains relatively understudied in the context of cancer is the insertion of a retrotransposable element into a novel position in the genome.

1.2 Retrotransposons

Retrotransposons are mobile genomic elements that "jump" via an RNA intermediate in a copy-and-paste mechanism across the genome. Regarded as "drivers of genome evolution", retrotransposons comprise nearly half of the human genome and are important vehicles of genomic diversity (Lander et al. 2001; Kazazian 2004). The majority of these elements are ancient insertions, which have significantly diverged in the last 100 million years and lost the capability to retrotranspose (Lander et al. 2001), however some 80-100 elements are still mobile (Brouha et al. 2003). The three most active retrotransposon families known are the Long INterspersed Element (LINE-1 or L1), Alu, and SVA (SINE/VNTR/Alu) (Figure 1-2) (Xing et al. 2009).

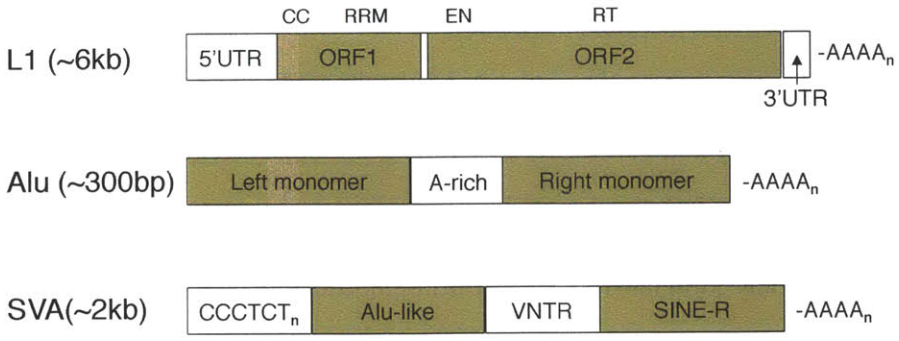


Figure 1-2 Diagram of active retrotransposon elements.
L1 is comprised of two open reading frames, ORF1 and ORF2, flanked by a 5' and 3' UTR and concluding with a poly(A) tail. ORF1 has a coiled coil (CC) motif and a RNA recognition motif (RRM), while ORF2 had an endonuclease (EN) and reverse transcriptase (RT) domains. Alu contains two monomeric regions separated by a short, A-rich sequence. SVA is a hybrid element combining a hexamer repeat ($CCTCT_n$) with two SINE elements separated by a variable-number-of-tandem-repeats (VNTR) region. Figure adapted from (Faulkner 2011).

L1

L1s are 6kb autonomous elements encoding their own retrotransposition enzymes. The structure of an L1 consists of a 5'UTR containing an internal RNA polymerase II promoter (Swergold 1990), two open reading frames (ORF1 and ORF2) and a 3' UTR containing a poly-adenylation signal ending with an oligo(dA)-rich tail of variable length (Babushok & Kazazian 2007). ORF1 encodes a 40 kDa protein that functions as a nucleic acid chaperone (Martin & Bushman 2001). ORF2 encodes a 150 kDa protein that contains an endonuclease (Feng et al. 1996) and reverse transcriptase (C et al. 1991). Both ORF1 and ORF2 are required for retrotransposition (J. V. Moran et al. 1996).

There are over 500,000 L1s annotated in the human reference genome (Lander et al. 2001; Venter 2001), consisting of more than 50 different families and subfamilies (Smit et al. 1995). Throughout the last ~40 million years of primate evolution, one actively mobilizing subfamily of

L1s has been replaced by another, so that only one subfamily is active at any time (Boissinot & Furano 2001; Khan et al. 2005). The reason for this is largely unknown, although it is speculated that competition for the same host factors between subfamilies prevents their coexistence (Khan et al. 2005). An interesting correlation exists between L1 family and hominid evolution (Figure 1-3). The divergence of Old-world monkeys and ancestral apes occurred around the same time as L1PA5 family of L1s took over from L1PA6 (Gibbs et al. 2007). At that time, ancestors of hominids were diverse but restricted to the tropical forests and woodlands of Africa and the Arabian Peninsula (Reed 1997). Later, human/chimpanzee separated from the gorilla during the arrival of L1PA2 (~8 Mya) likely due to the cooling and drying of Africa, reducing ecological diversity and causing hominids to become dominant. Within the human lineage, the arrival of L1 pre-Ta (~3 Mya) and L1 Ta-1 (~2 Mya) subfamilies corresponds to the speciation of *Australopthecus africanus* and *Homo ergaster*, respectively (J. Lee et al. 2007). It is thought that these species made great advances in human cognition, which affected both behavior and intelligence, associated with cranial size expansion and the use of complex tools, and primitive language, and hunter/gatherer society (Stout et al. 2008). Although these are correlations, the concordance of the timing of these evolutionary events is striking and brings to question L1's role in hominid evolution.

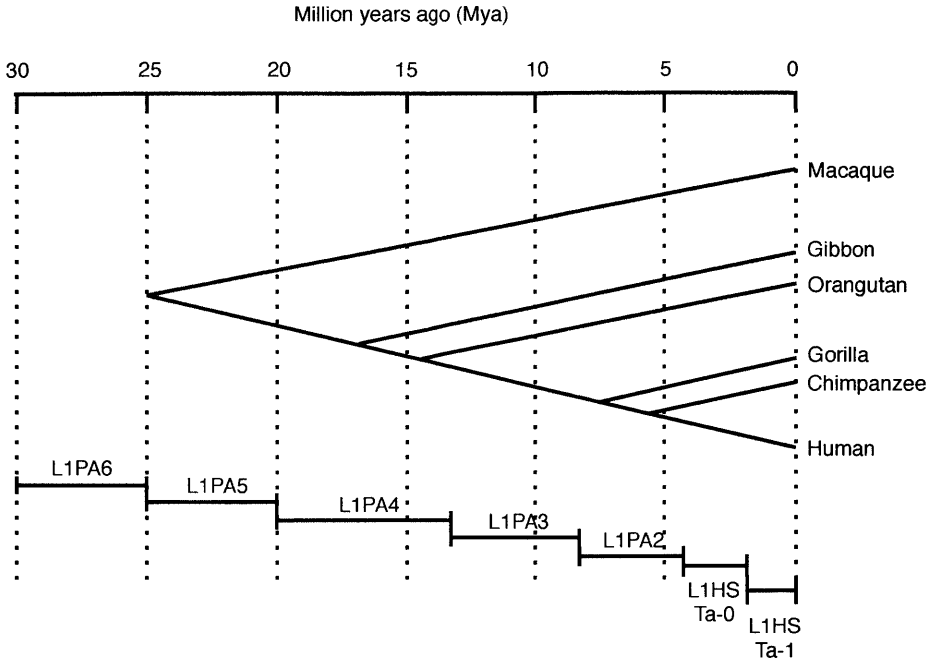


Figure 1-3 Linear evolution of L1.
L1 family and hominid evolution over the past 30 million years. Figure adapted from (C. A. Thomas et al. 2012).

Currently, the active L1 family is the L1HS (human specific) element, which is subdivided into pre-Ta and Ta (Transcribed group a) subfamilies (Salem et al. 2003). Ta is further subdivided into Ta-0 and Ta-1 based on diagnostic nucleotides scattered throughout practically identical sequences (Boissinot et al. 2000; Ovchinnikov 2001; Brouha et al. 2003). Ta-0 is older than Ta-1, and although Ta-0 retains some active elements, Ta-1 now accounts for about one half of the Ta family. There are some 80-100 such Ta elements described as "hot", or those that retain retrotransposition capacity, and these are responsible for the majority of current L1 retrotransposition (Brouha et al. 2003; Beck et al. 2010).

Alu

Alus are 300bp-long elements that take advantage of the L1 retrotransposition machinery to mobilize. They comprise 11% of the genome, with over 1 million elements. The structure of an Alu consists of the fusion of two monomers derived from the *7SL* RNA gene, the RNA scaffold of the signal recognition protein (SRP), separated by an A-rich linker region. The 5' region contains an internal RNA polymerase III promoter and the element ends with an oligo(dA)-rich tail of variable length. Alus are thought to localize to the ribosome via binding with the SRP9/14 protein complex and that is where they are thought to interact with the nascent L1 ORF2 protein (Boeke 1997). Like L1s, Alus can be divided into many subfamilies (Deininger et al. 1992). The AluY subfamily, most notably AluYa5 and AluYb8, account for the majority of disease-producing insertions in humans (Carroll et al. 2001).

SVA

SVAs are heterogeneous, non-autonomous elements ranging in size from 700bp to 4kb-long, composed of a hybrid of other repeat elements, and present at about 3,000 copies in the human genome (H. Wang et al. 2005; Ostertag & Kazazian 2001). They are also mobilized *in trans* by the L1 machinery. Canonical SVAs contain a variable number of CCCTCT repeats at their 5' end, followed by an Alu-like domain, a GC-rich variable number of tandem repeats (VNTR) domain, and a SIN-R domain which is derived from the envelope (env) gene and right LTR of an extinct HERV-K10 element (Hancks & Kazazian 2012). SVAs contain a poly(A) signal and variable length poly(A) tail.

Although L1 displays a strong *cis* preference to retrotranspose its own mRNA (Wei et al. 2001), other RNAs in addition to Alu and SVA can also hijack the L1 retrotransposition machinery and

insert into new locations within the genome to create new pseudogenes and regulatory sites (Esnault et al. 2000).

Target Primed Reverse Transcription

The first step in L1 retrotransposition is the transcription of genomic L1 from an internal promoter; the L1 RNA is exported to the cytoplasm, in which ORF1 and ORF2 are translated. Both proteins preferentially associate with the L1 RNA transcript that encoded them to produce a ribonucleoprotein (RNP) particle. The RNP is then transported back into the nucleus (the mechanism for which remains unclear). The canonical mechanism by which retrotransposons have been shown to insert into the genome is known as target primed reverse transcription (TPRT) (Luan et al. 1993; Cost et al. 2002) (see Figure 1-4 for a schematic). In this process, the L1 endonuclease creates a nick in a DNA strand and the L1 reverse transcriptase extends the free-hanging 3'-OH which serves as a primer for newly synthesized cDNA. The second DNA strand is staggered such that a short ~15bp sequence of the target site is duplicated flanking the insertion (termed a target site duplication, or TSD). This mechanism was first proposed for the *Bombyx mori* R2 retrotransposon based on the observation that R2 endonuclease activity was coupled with initiation of reverse transcription (Luan et al. 1993; Luan & Eickbush 1995). TPRT often results in inversions and truncations of 5' L1 sequence (Boissinot & Furano 2001; Szak et al. 2002). It remains unclear how exactly the integration is completed; host DNA repair proteins such as *ATM*, likely recognize and process L1 integration intermediates (Gasior et al. 2008).

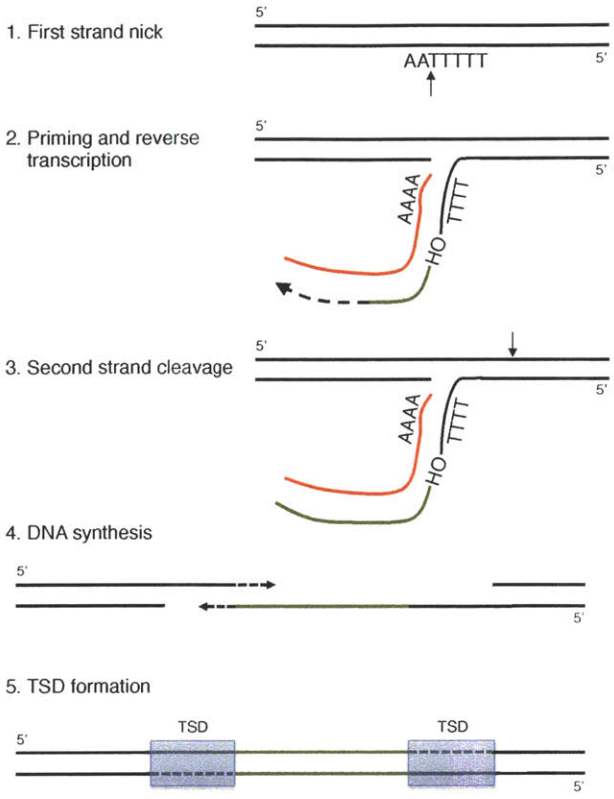


Figure 1-4 Schematic of target primed reverse transcription.
In canonical target primed reverse transcription (TPRT), 1. the L1 endonuclease creates a nick in one strand of the DNA, 2. the free –OH is then used as a primer for the L1 reverse transcriptase to convert L1 RNA (red) to DNA (green), 3. the second strand of DNA is cleaved typically some 15bp downstream of the initial nick, 4. DNA synthesis proceeds to repair the two nicks in a process that remains poorly understood, and 5. TSDs surrounding the insertion are created as a result. Figure adapted from (Cordaux & Batzer 2009).

1.3 Methods for studying retrotransposition

Retrotransposition Assay

Activity of L1s has been studied extensively with the use of L1 retrotransposition assays in culture (J. V. Moran et al. 1996; Rangwala & Kazazian 2009; Freeman et al. 1994). Briefly, this technique relies on a reporter gene signaling *de-novo* retrotransposition via splicing of a disruptive intron when the transcript is reverse transcribed, integrated into chromosomal DNA

and expressed from its own promoter. Using these assays in cultured human cell lines, several L1 subfamily elements have been shown to be capable of high frequency autonomous retrotransposition (J. V. Moran et al. 1996; Sassaman et al. 1997; Wei et al. 2000).

Hybrid capture assay (RC-seq)

In order to interrogate retrotransposition in a more high-throughput manner, custom sequence capture arrays are employed. Retrotransposon capture sequencing (RC-seq) (Baillie et al. 2011) and L1-seq (Ewing & Kazazian 2010) methods first enrich for the 5' and 3' termini of full-length L1 (L1-seq) and Alu and SVA retrotransposons (RC-seq) using targeted hybridization arrays. Captured DNA fragments are then sequenced using paired-end massively parallel sequencing and reads spanning the insertion junctions between reference genome and retrotransposon are computationally analyzed.

Mining sequencing data

The abundance of tumor sequencing data becoming available provides a unique opportunity to interrogate hundreds of tumor and matched normal samples for retrotransposon movement using existing data. Since these data are produced for other purposes and cover a vast portion of the genome (rife with repeat elements and reference retrotransposons) at relatively low coverage, they require specialized algorithms to accurately discover true novel retrotransposon insertions. Several methods exist for discovery of germline non-reference retrotransposons in whole-genome sequencing data (Ewing & Kazazian 2010; Stewart et al. 2011; Keane et al. 2013). Somatic retrotransposon insertion identification, however, requires additional considerations due to the complexity of these events (see Section 1.5).

1.4 Germline variation

Retrotransposon insertions have recently been described as a major source of genetic variation. The rate of Alu retrotransposition is approximately 1 insertion for every 20 births, based on both the frequency of disease-causing de novo insertions compared with normal nucleotide substitutions and on comparative genomic studies between human and chimpanzee genomes (Cordaux & Batzer 2009) and between human genomes (Xing et al. 2009). The current rate of L1 retrotransposition is approximately 1 insertion for every 200 births based on genome comparisons (Kazazian 1999). And the rate of SVA retrotransposition is tentatively estimated at 1 in every 900 due to smaller data sets (Xing et al. 2009; Cordaux & Batzer 2009).

Amplification rates of retrotransposons have not been uniform over time, with the most prolific period of L1 insertion about 12-40 million years ago and that of Alu insertions ~40 million years ago when there was a new Alu insertion in every birth (Cordaux & Batzer 2009). The impact of transposon mutagenesis was likely greatest in humans during the past ~6 million years, since the split from chimpanzee lineage. The human genome has supported more L1, Alu, and SVA retrotransposition events than chimpanzees; specifically, humans harbor an additional 5,000 transposon insertions compared with chimpanzees (Mills et al. 2006).

Recent studies of human genetic variation, such as the 1000 Genomes Project, have led to the discovery of thousands of polymorphic retrotransposon sites within and across human populations. The database of retrotransposon insertion polymorphisms (dbRIP) (J. Wang et al. 2006) contains 2,761 known polymorphic insertion sites, and studies from the 1000 Genomes Project (Stewart et al. 2011; Ewing & Kazazian 2011) have reported 5,291 additional L1, Alu

and SVA insertions in normal, healthy individuals. A pair of individuals of European origin are estimated to differ by approximately 500-800 retrotransposon insertion polymorphisms (Stewart et al. 2011). The prevalence of retrotransposon polymorphisms indicates that active retrotransposition is an ongoing feature of human population variation.

Retrotransposition in the Brain

L1 is capable of retrotransposition in germ cells (Ostertag et al. 2002) as well as in neuronal progenitor cells (Muotri et al. 2005); however, the extent of somatic L1 retrotransposition in the brain is currently under debate. Within an individual, neuronal genomes are diverse and brains form "somatic mosaics"; this neuronal diversity is vital for neural plasticity, cognition and behavior (Singer et al. 2010). The genetic mechanisms contributing to this diversity include aneuploidy (Rehen et al. 2001), copy number variations (Bruder et al. 2008), and possible L1 insertions (Muotri et al. 2005). L1 elements are mobilized early in development during the formation of the central nervous system (CNS) and later during adult neurogenesis. Since this mobilization process appears to occur frequently and independently in individual cells, the result is potentially a substantial number of newly transposed L1 elements in differentiated neurons (Singer et al. 2010). In fact, studies suggest that a surprisingly large number of somatic retrotransposon insertions specific to neurons from the hippocampus and several other areas of the brain (Coufal et al. 2009; Baillie et al. 2011), and thus L1 retrotransposition is a *main* contributor to neuronal cell diversity. The candidate somatic insertions they report, however, are low-coverage events with high false-positive rates, and multiple rounds of site-specific PCR validation of 30 putative insertions carried out on the retrotransposon-junction enriched library (not the original DNA) may also have produced artifacts that inflate the validation rate (Xing et

al. 2013). Another study used single-neuron sequencing from three individuals to reveal far less L1 retrotransposition in the cerebral cortex and caudate nucleus (Evrony et al. 2012). All of their candidate somatic insertions were subject to PCR validation on the original genomic DNA and only five insertions were validated in 300 cells. This equates to a rate of one L1 insertion per 25 cells, consistent with the rate of insertions per cell division in the germline (Xing et al. 2013). These results dispute L1's role as a major generator of neuronal diversity, at least in the cortex and caudate nucleus. L1 retrotransposition in the brain, though potentially beneficial in terms of diversity, may also have negative implications for neuronal genomes, such as increased disease risk. Rett Syndrome, a rare neurodevelopmental disorder caused by mutations in the methyl-CpG-binding protein 2, *MeCP2,* gene, which is thought to control L1 retrotransposition, is associated with overly active L1 retrotransposons (Muotri et al. 2010). Thus, somatic retrotransposition in the brain remains a controversial topic that will require more accurate (both sensitive and specific) identification methods and more focused sequencing efforts.

1.4 Retrotransposons in Disease

In addition to neurological diseases, retrotransposon insertions have been implicated in almost one hundred other single-gene human diseases. Retrotransposon insertions can have various effects on the genome depending on where they land (Figure 1-5). If a retrotransposon inserts in an intragenic region, it can affect gene expression via a variety of mechanisms. Most intuitively, insertion into a coding region can disrupt the codon code and create missense or nonsense mutations (Kazazian et al. 1988). An insertion can also change a gene's splicing pattern with alternate splice sites (Mülhardt et al. 1994), by exon skipping (Takahara et al. 1996), or by altering a regulatory sequence (Shukla et al. 2013). Additionally, the L1 5'UTR has both sense

and antisense promoter activity, so an insertion can create new transcription start sites in both directions (Speek 2001; Wolff et al. 2010). Outside of intragenic regions, insertion of a retrotransposon can lead to further genomic rearrangement due to nonallelic homologous recombination (Robberecht et al. 2013), and in general, catalyze a large amount of genomic instability on the cell (Symer et al. 2002). These retrotransposon-mediated deletions and rearrangements have been demonstrated in transformed cell lines, in certain spontaneous germline diseases, and during evolution (Burwinkel & Kilimann 1998; Gilbert et al. 2002; Han 2005).

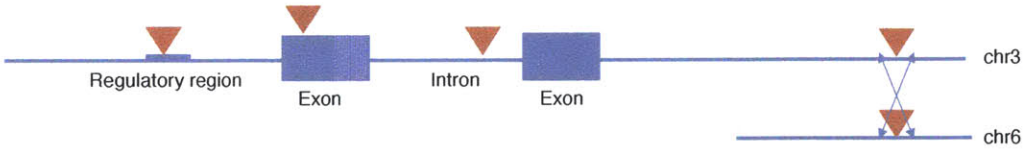


Figure 1-5 Effects of retrotransposon insertion on the genome.
Possible effects of retrotransposon insertion (red triangle) in various genomic contexts: insertion into a regulatory region such as an enhancer or repressor may affect gene expression nearby, insertion into exon may cause gene dysfunction, truncation or exon-skipping, insertion into an intron may also cause alternative splicing and variant isoforms or affect gene regulation, and insertion into even intergenic regions may lead to homologous recombination and further genomic instability.

Over ninety human diseases are known to be caused by heritable or *de novo* retrotransposition events (Cordaux & Batzer 2009; Hancks & Kazazian 2012), including hemophilia A caused by L1 insertions into an exon of the Factor VIII gene (Kazazian et al. 1988) and hemophilia B due to an Alu insertion into the coding region of the Factor IX gene (Vidaud et al. 1993). Several instances of Duchenne muscular dystrophy have been revealed to be caused by an L1 insertion into the dystrophin (*DMD*) gene (Narita et al. 1993; Holmes et al. 1994; Musova et al. 2006; Awano et al. n.d.; Solyom et al. 2011), often resulting in exon skipping, with some insertion sites in independent patients within 87 bp from one another. Similarly, the Neurofibromatosis Type 1

(*NF1*) gene contains hotspots for *de novo* retrotransposon insertion, with three integration sites each used twice in independent insertion events, and six insertions clustering in a 1.5-kb region (Wimmer et al. 2011). Episodic evidence such as these suggest non-random retrotransposon integration into the genome. Finally, the Fibroblast growth factor receptor 2 (*FGFR2*) gene harbors multiple causal Alu insertions in Apert syndrome (Oldridge et al. 1999; Bochukova et al. 2009). Mutations in *FGFR2* are associated with abnormal bone development diseases, such as Apert syndrome, but also gastric, breast, endometrial and lung cancer (Xie et al. 2013; Reintjes et al. 2013; Dutt et al. 2008; Liao et al. 2013).

1.5 Retrotransposons in Cancer

Early insertion into APC

The first record of a *bona-fide* tumor-related retrotransposon insertion came in 1992 when an exon of the tumor suppressor, adenomatous polyposis coli (*APC)* was reported to be disrupted by the somatic insertion of an L1 element in a colorectal tumor (Miki et al. 1992). This event was discovered during the search for somatic mutations in the *APC* gene specifically across 150 cases of colorectal cancer, and the L1 insertion was further characterized to exhibit several hallmarks of TPRT, such as 5' truncation and an 8bp duplication at the insertion site. Because *APC* mutation is an early event in the colorectal tumorigenesis, this account remains to our knowledge the only likely causal somatic L1 insertion in human cancer.

Colorectal, lung, and liver cancers from RC-seq

Using an RC-seq assay (described above), Iskow et al. (2010) was the first study to systematically examine tumor samples for retrotransposon movement. They found nine somatic

L1 (and zero Alu) insertions in six of 20 primary non-small cell lung tumors (Iskow et al. 2010), and no retrotransposon movement in any of the five glioblastoma and five medulloblastoma samples they analyzed. Using a similar RC-seq method, Solyom et al. (2012) later revealed a high rate of L1 retrotransposition in certain colorectal cancer genomes. They found 67 tumor-specific L1 insertions in 16 primary colorectal samples (Solyom et al. 2012). In liver cancer, Shukla et al. (2013) found 12 somatic L1 insertions in 19 tumor and matched normal samples using RC-Seq (Shukla et al. 2013). Notably, they describe activation of the transcriptional repressor suppression of tumorigenicity 18 (*ST18*) gene by somatic insertion of a 410bp L1 element into an intronic binding site motif. This represents the first example of a somatic retrotransposon insertion upregulating a gene in cancer, suggesting that, in addition to the repression of tumor suppressors, proto-oncogenes may be activated via this mechanism in tumors.

Lee et al. (2012)

In 2011, we publically presented our technique for identifying novel somatic retrotransposon insertions in whole-genome sequencing data and described the first account of multiple tumor-specific L1 insertions in nine colorectal carcinoma genomes at The Cancer Genome Atlas (TCGA) Annual Symposium (Helman & Meyerson, 2011). In 2012, Lee and colleagues published a similar method (E. Lee et al. 2012) applied to whole-genome sequencing data from 45 tumor and matched normal samples across five cancer types in TCGA. They confirmed the prevalence of somatic L1 insertions in colorectal cancer in five samples, with one outlier colorectal tumor exhibiting more than 100 such events. Glioblastoma and multiple myeloma did not show any signs of somatic retrotransposition, while ovarian and prostate tumors each had a

few cases of somatic L1 insertions. Although the study found that some genes with L1 insertions are frequently mutated via alternative mutations in cancer, Lee et al. (2012) did not find any evidence of causal retrotransposon insertions important to tumorigenesis. This represents the first published manuscript describing somatic retrotransposon insertions in cancer from 45 whole-genome sequences.

Technical challenges

Although these studies present a start to the examination of retrotransposon insertions in tumors, the full extent of somatic retrotransposition in human cancer remains largely unexplored. This may be because, biologically, somatic retrotransposon movement is perhaps rare and its effect size is small. Other reasons for this discrepancy may be more technical. In general, investigations of genetic changes in cancer avoid dealing with repeat sequences and focus on protein-coding genes (Schulz 2006). Repetitive DNA is often deemed "junk DNA" with no functional consequence on oncogenesis. Due to their sheer abundance in the genome, repetitive elements are difficult to study with current methodologies that rely on unique genomic sequence. Computational algorithms as well as experimental protocols must be optimized for repetitive sequence and biological interpretation must look beyond protein-coding genes.

Coupled with proper scientific research methods and attitudes, investigating the importance of retrotransposition across human cancers requires the power of numbers. Studies of other somatic genome alterations have amassed thousands of samples in order to identify function, localize targets, and stratify patients. Large sequencing efforts, like The Cancer Genome Atlas (NHGRI

2009), are enabling the systematic interrogation of the cancer genome in numbers that are approaching those needed to elucidate relevant somatic mutations.

1.6 Overall Objective

The **overall goal** of this thesis is to examine the extent of somatic retrotransposition across cancer through next-generation sequencing and to comprehensively characterize the genomic attributes associated with these events in tumors.

Chapter 2. Tools for interrogating next-generation sequencing data for novel retrotransposon insertions

Paired-end sequencing data provides the opportunity to search for somatic retrotransposon movement genome-wide across cancers. To leverage this rich data source, an algorithm that parses high-dimensional data accurately and efficiently is crucial. Since retrotransposons comprise almost half of the genome, locating novel insertions has been described as "finding a new straw of hay placed in the middle of a haystack" (C. A. Thomas et al. 2012). To find this straw, *in silico* computational methods for cataloging repeat insertions can now be used where computational pipelines rather than sequencing methods are tailored for repeat discovery (Burns & Boeke 2012). Although many methods for genomic rearrangement identification existed, tools to localize somatic retrotransposon insertion were lacking. We decided to create a tool to search for tumor-associated instances of retrotransposition within the compendium of paired-end sequencing data available through TCGA and other large sequencing projects. To this end, we created TranspoSeq (Figure 2-1).

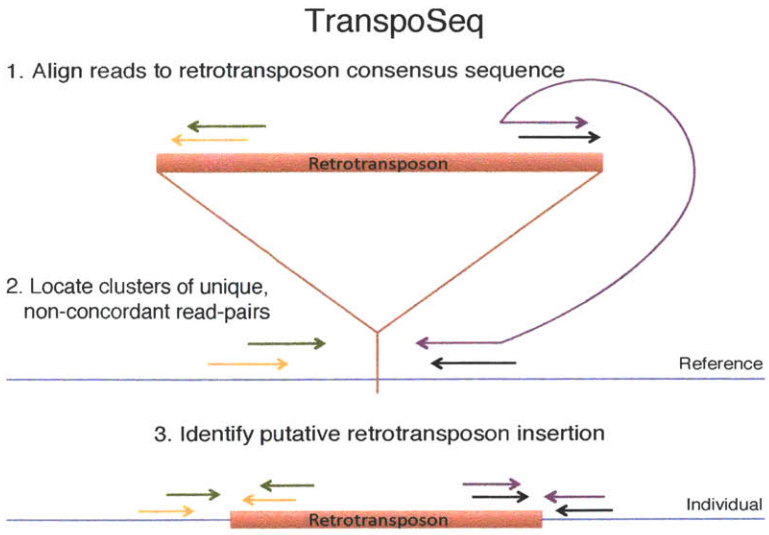


Figure 2-1 Outline of TranspoSeq algorithm.
TranspoSeq identifies clusters of unique sequencing reads whose discordant pair-mates align to a database of consensus retrotransposon sequences to localize a putative non-reference retrotransposon insertion at a specific genomic site.

2.1 TranspoSeq methodology

TranspoSeq was first presented in 2011 as RetroSeq (Helman & Meyerson 2011; Helman & Meyerson 2011; Helman & Meyerson 2012) (http://cancergenome.nih.gov/newsevents/multimedialibrary/videos/retroseqhelman). It uses both paired and split read information to identify and characterize non-reference retrotransposon insertion events from tumor and matched normal BAM files. TranspoSeq consists of three main steps: 1) Get Reads, 2) Process Reads, and 3) Assemble Reads. See Figure 2-2 for a detailed schematic of the process.

1. Get Reads:

Beginning with the input BAM file, TranspoSeq parses out all discordant read-pairs, defined as pair-mates whose aligned positions are non-concordant with the fragment length distribution. We use a threshold of 1kb to call a non-concordant read-pair, in order balance the desired sensitivity and specificity given an average fragment length of about 400 basepairs. These read-pairs are then aligned to a database of consensus retrotransposon sequences using NCBI's blastn algorithm. Reads that align with either a predefined minimal percent identity and number of consecutive bases, or a predefined maximal BLAST e-value are kept for further processing. In this analysis, we use a BLAST e-value threshold of 2E-07, which is equivalent to approximately 30 consecutive nucleotides with 85% identity to the consensus retrotransposon sequence. For each read that successfully aligns, we locate its pair-mate: if this mate also aligns to the retrotransposon database, the pair is discarded; if not, and the mate aligns to the genome with adequate mapping quality (MAPQ>0), the pair is collected for further processing.

2. Process Reads:

Unique reads whose pair-mates align to a retrotransposon consensus sequence are grouped by read orientation (forward or reverse) and each set is clustered separately. Clusters are defined by the distance between the start positions of two adjacent reads as no larger than 200bp. Forward and reverse clusters are then overlapped – allowing for an overlap of up to 60bp and a gap of up to 500bp between a forward and reverse cluster, in order to account for target sequence duplications (TSDs) and variable coverage. Parameter values were chosen based on prior knowledge as well as empirically, and tested on simulated datasets. One-sided events, clusters without an overlapping cluster in the opposing orientation are set aside for future investigation.

Events supported by clusters in both directions are annotated based on: presence in matched normal sample, proximity (within a 200bp window) to a reference retrotransposon, known RIP (dbRIP (J. Wang et al. 2006) and 1000Genomes (Thibodeau et al. 1993; Stewart et al. 2011; Ewing & Kazazian 2011)), known gene (RefSeq track of UCSC Genome Browser (Fujita et al. 2010)), and known CNV (Beroukhim et al. 2007). Events are also annotated with information pertaining to alignment to the retrotransposon database: identity, inferred length, and inversion status of inserted retrotransposon element.

3. Assemble Reads:

Read-pairs supporting a candidate insertion as well as split reads spanning the putative insertion breakpoint are then assembled *de novo* using INCHWORM (Grabherr et al. 2011) to form contigs in the forward and reverse directions separately. Contigs in each direction are aligned back to the database of retrotransposon consensus sequences with BLAST (blastn) and to the reference genome using BLAT. The longest contig containing a retrotransposon-aligned region and a reference-aligned region with minimal overlap is returned along with the specific retrotransposon subfamily and alignment properties. If such a contig cannot be constructed, TranspoSeq uses the alignment properties of the discordant reads themselves. Split reads are used, when available, to determine the forward and reverse breakpoints as well as the putative TSD sequence defined as the region between these forward and reverse breakpoints.

Filtering

Post-processing filtering is performed to remove regions with greater than 30% poor quality

reads (MAPQ=0), less than 0.005 allelic fraction, and greater than 25 discordant reads within the candidate region in the normal sample, as well as regions that did not produce at least one substantial contig (>14bp) from *de-novo* assembly. Allelic fraction is calculated by (number of split reads supporting insertion from both sides/2)/(number of total reads spanning breakpoint). To increase sensitivity and prevent filtering out almost half of the genome, we do not discard insertions that fall into all reference retrotransposons, but only those that land in reference elements within the same subfamily (i.e., L1PA, L1PB, etc.). Only events with at least 10 read-pairs, including at least two in each direction, supporting the insertion were maintained. Events consistent with microsatellite instability or ancient retrotransposons were filtered out. Finally, we manually reviewed each putative somatic insertion region using the Broad Institute's Integrative Genome Viewer(Robinson et al. 2011) and only those events that passed manual inspection were retained for further analysis.

Consensus retrotransposon sequences were downloaded from GIRI RepBase (www.girinst.org/repbase/). All elements in the L1 (n=117) and SINE1/7SL (n=55) families, as well as SVA were included in this analysis. Reference retrotransposon identities were downloaded from RepeatMasker on January 12, 2013 (repeatmasker.org).

When comparing putative retrotransposon insertions to annotated polymorphisms, we used the largest database of known retrotransposon insertion polymorphisms (dbRIP) (J. Wang et al. 2006), accessed on May 22, 2012, at which point it contained 2086 Alu, 598 L1, and 77 SVA annotated elements, and data from ten other previous studies reporting germline retrotransposon insertions (E. Lee et al. 2012; Beck et al. 2010; Huang et al. 2010; Hormozdiari et al. 2010; Xing

et al. 2009; Iskow et al. 2010; Witherspoon et al. 2010; Stewart et al. 2011; Ewing & Kazazian 2010; Ewing & Kazazian 2011).

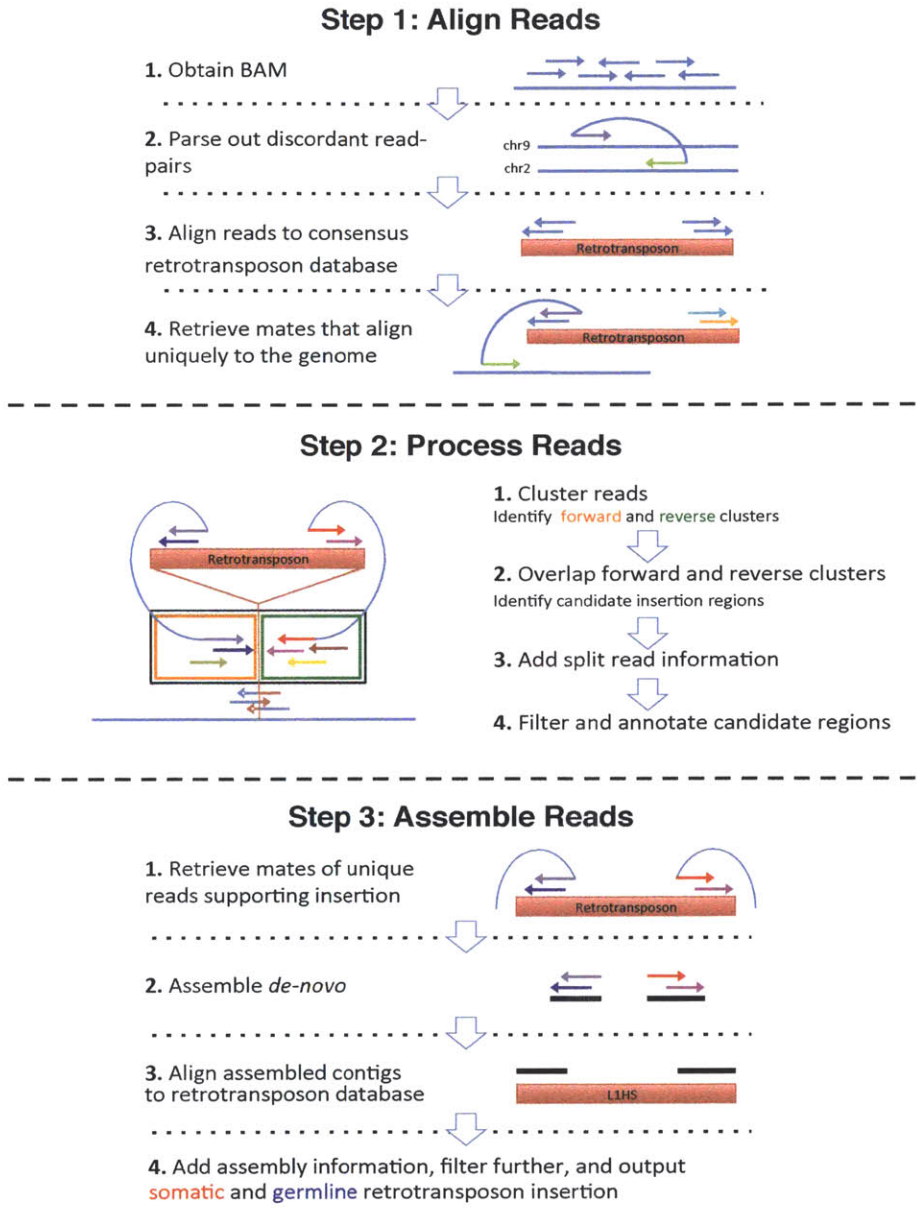


Figure 2-2 Detailed schematic of TranspoSeq pipeline.
TranspoSeq is a computational framework that takes in paired-end sequencing data and produces a list of annotated putative somatic retrotransposon insertion sites. First, input BAMs are parsed for discordant read-pairs; these pairs are then aligned 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.

Structure

The steps involved in TranspoSeq are computationally intensive given that the whole-genome BAM files are each on the order of 200 gigabytes (assuming a coverage of approximately 40X). To reduce run-time, the current implementation of TranspoSeq depends on the Broad Institute's load sharing farm to run parallel processing on each chromosome arm. Using a method called 'scatter-gather', TranspoSeq splits up the steps of the pipeline that can be run in parallel (scatter) and then collates resulting outputs (gather) before scattering again for the next step (Figure 2-3). TranspoSeq uses the Picard Samtools netsf java toolkit to parse BAM files and R for data processing (http://picard.sourceforge.net/javadoc/net/sf/samtools/package-summary.html). Pipelines are run with the reference assembly corresponding to the input BAMs, and the resulting calls are then converted to Hg19 when necessary using UCSC Genome Browser Database liftOver (Meyer et al. 2012). The alignment parameters discussed above were used in this study; however, they are input parameters that are easily modifiable for future runs of the pipeline.

TranspoSeq will be available at www.broadinstitute.org/cancer/cga/transposeq.

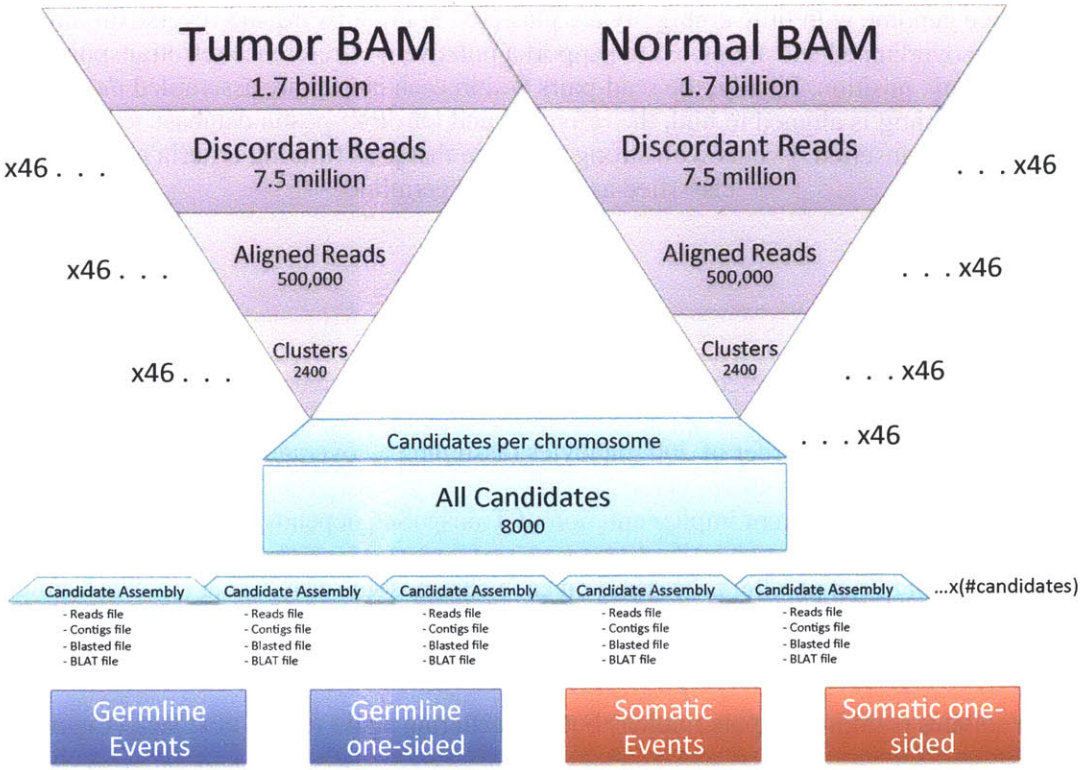


Figure 2-3 Infrastructure of TranspoSeq tool.
Schematic describing the organization of the TranspoSeq tool, implementing a scatter-gather algorithm on the Broad Institute's load sharing farm. Tumor and normal BAM files are required as input, each containing approximately 1.7 billion reads. These are then filtered for only discordant read pairs and split across 46 files according to chromosome arm (7.5 million reads each). These are aligned to the retrotransposon database resulting in 46 alignment files containing 500,000 reads. These files are then gathered back, split across files by unique pair-mate per chromosome arm, and each file is clustered in parallel, resulting in ~2400 clusters. Finally, clusters are filtered and gathered to one candidate file. This file is then split by candidate and each candidate region undergoes assembly and alignment, before they are gathered back, filtered, and parsed to output germline and somatic one- and two-sided putative insertion events. The entire procedure, assuming infinite cluster node availability, takes about 300 CPU hours.

Manual review

Somatic events that passed filtering criteria were subject to strict manual review using IGV. Figure 2-4 shows an example of a homozygous L1HS insertion where the colored bars represent sequencing reads whose pair-mates align to L1HS reference elements on different chromosomes. Importantly, using the clipped basepairs (indicated by the colored nucleotides), we are able to

identify the precise locations of the insertion breakpoints. The space between clipped reads in forward and reverse directions represents the duplicated sequence at the point of insertion (the TSD). In Figure 2-4, the poly(A) tail of the inserted L1HS is visible as clipped reads of adjacent adenosines (red), indicating an inverted L1HS insertion. Reasons that a putative insertion failed to pass manual review included: misclipped reads, spurious chimeric read-pairs in the region, and general poor alignment and read quality. Furthermore, in order to pass review as a tumor-specific event, the genomic location must not contain any supporting reads or clipped reads in the matched normal genome. The somatic status of putative events with low coverage in the matched normal could not be determined. Finally, if an insertion event supports a genomic rearrangement rather than a sole retrotransposition, that is, the reads have pair-mates that all align to the same region in the reference, this event is marked as a likely rearrangement associated with a retrotransposon and is discarded from further analysis.

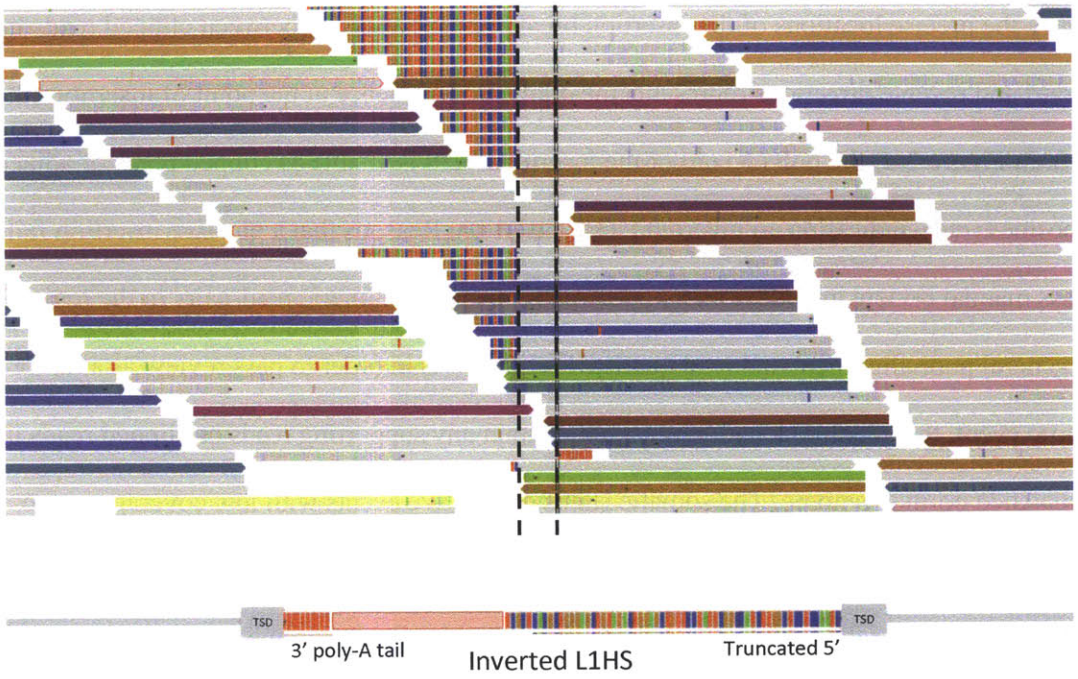


Figure 2-4 Visualization of retrotransposon insertion from paired-end sequencing reads. Colored bars represent reads with pair-mates aligning to a different chromosome (depending on color). Clipped nucleotides are represented by colored basepairs. The space between the two dashed vertical lines is the TSD. The bottom panel diagrams the insertion revealed by the sequencing reads in the top panel.

Retrotransposon insertion versus rearrangement

In this analysis, we do not identify retrotransposon insertions that lead to chromosomal rearrangements or rearrangements that occur as a result of existing retrotransposons (Gilbert et al. 2002; Gilbert et al. 2005). Translocations that are due to somatic retrotransposon insertions will likely appear as a "one-sided" retrotransposon insertion event in close proximity to a one-sided genomic rearrangement event. In other words, sequencing reads in one direction will have pair-mates aligning to retrotransposon elements, while sequencing reads in the other direction will have pair-mates aligning to one region in the reference genome. We exclude these events from our downstream analyses, although TranspoSeq does identify and retain them.

Limitations

There are several limitations that must be considered in repeat element localization due to the high homology between elements and their sheer abundance in the human genome. Transposon subfamily identification in particular is limited by these factors. The ability to distinguish between inserted Alu sequences will be especially restricted because of their small size (~300bp) and extreme sequence homology. Identifying an inserted element is confounded by reference elements in the same family nearby. To remove risk of false positives from misalignment due to a reference element, we could discard all pairs of reads where both reads align to any element in the retrotransposon database, but this would mean disregarding almost half of the genome. As a compromise, we only discard read-pairs where both reads align to the same element subfamily. If an element is inserted directly inside or within 700bp from a reference retrotransposon, assembly through the junction results in a contig that aligns to the reference element instead of the newly inserted element. Additionally, tumor heterogeneity and purity will affect the sensitivity of TranpoSeq. Currently, events of extremely low allelic fraction are discarded.

2.2 TranspoSeq performance metrics

Simulation

To assess TranspoSeq's ability to identify novel retrotransposon insertions, we created simulated genome alignment files and randomly inserted retrotransposon sequences *in-silico* (Figure 2-5). Simulated alignment data were created by computationally inserting 226 full length L1HS and 772 AluY 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 (Stratton et al. 2009; Li et al. 2009) was used to create a simulated BAM file with read length 100bp, fragment length 500bp, 20x coverage and default values for all other parameters. A second simulated dataset was made using the same method, but inserting 100 elements each of 5' truncated L1HS of lengths 40bp through 6000bp for a total of 1000 inserted elements. Although these simulations inherently have a higher signal-to-noise ratio than real heterogeneous and potentially contaminated tumor data, we attempted to recreate a true insertion event as realistically as possible. We add 15bp TSDs surrounding the insert and include *wgsim*'s baseline mutation and sequencing error rate.

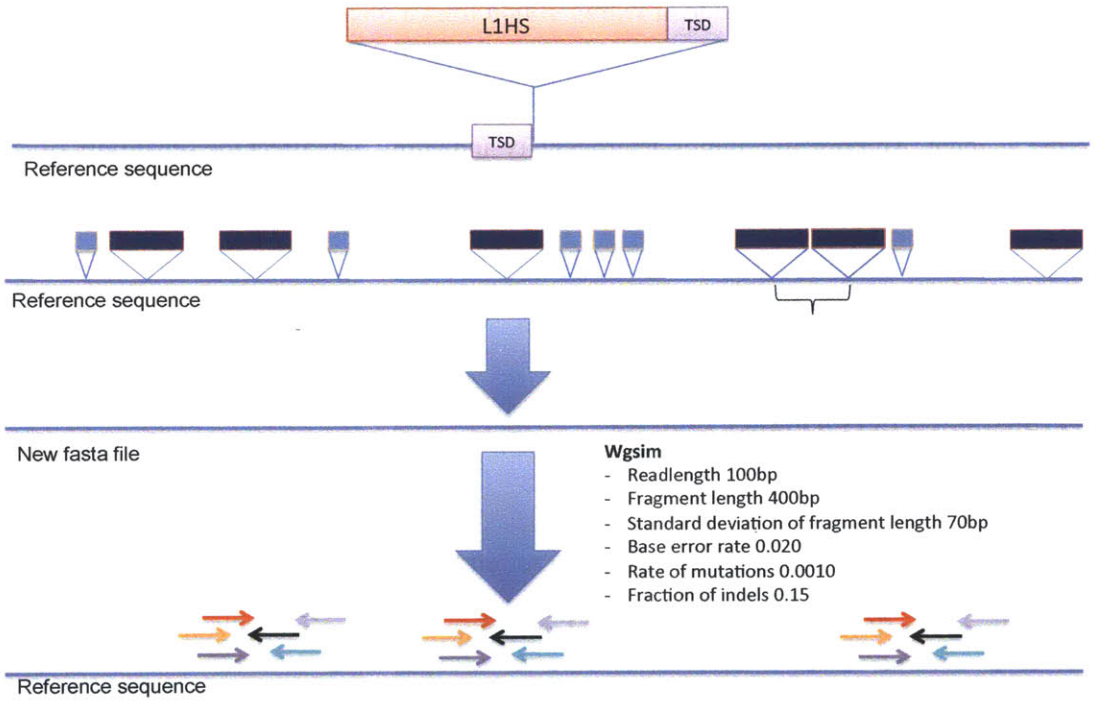


Figure 2-5 Schematic of simulated data generation.
Simulated data was generated by inserting consensus retrotransposon sequences into a reference genome and creating artificial TSDs surrounding the insertion site. The new genome was then converted to paired-end BAM file using Wgsim with the parameters listed.

We ran TranspoSeq on a simulated alignment file with 226 L1 and 730 Alu elements computationally inserted; TranspoSeq was able to correctly identify 225/226 L1 and 730/732 Alu elements with no false positive calls. Next, we created a simulated file with 5'-truncated L1s of varying lengths, 100 instances of element lengths ranging from 40bp to 6kb, and found that TranspoSeq's performance begins decreasing at around 150bp for both germline and somatic calls (see Figure 2-6). This may influence the tool's ability to detect severely truncated L1 elements, but should not impede performance on Alus because these are ~300bp and well within the sensitivity limit.

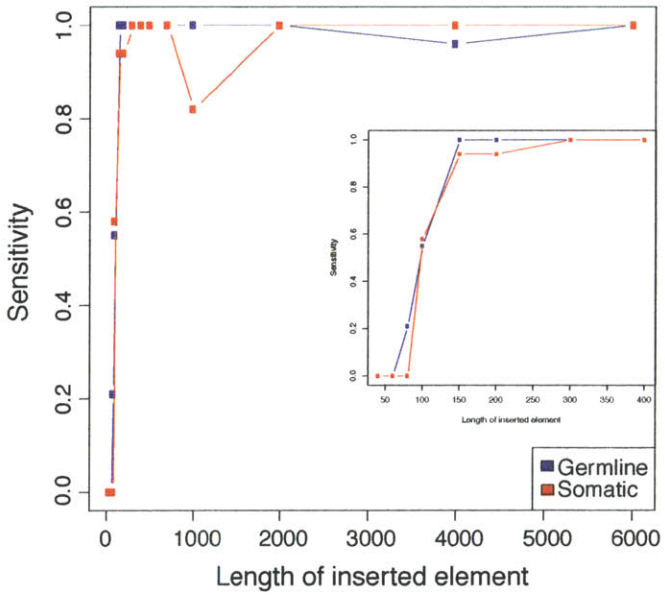


Figure 2-6 Sensitivity of TranspoSeq to insertion length.
Fraction of total L1 insertions of varying lengths identified by TranspoSeq. Inset shows sensitivity at lengths below 400bp.

Comparison to similar methods

TranspoSeq is functionally similar to other recently reported read-anchored and split-read mobile element insertion tools such as Tea (E. Lee et al. 2012) and the Sanger Institute's RetroSeq

(Keane et al. 2013), but includes additional *de-novo* assembly and contig alignment procedures. To computationally assess the performance of TranspoSeq, we compared our findings to those of Lee et al. (2012) (E. Lee et al. 2012) on the colorectal sample TCGA-AA-3518. Of the 146 high-confidence somatic retrotransposon insertions we identify, 91 insertions are common to both studies (see Figure 2-7). Fifteen events are missed by TranspoSeq; upon manual review, these events do not pass TranspoSeq's stringent filtering criteria including coverage and mappability of region required to call a somatic event. Xing et al. (2013) found that retrotransposon observed in only one individual or a sample and supported by only by a few sequencing reads (ten or less) have a validation rate of approximately 20% (by locus-specific PCR) (Xing et al. 2013). Events supported by few reads are likely to be false positives, reflecting chimeras generated during library preparation. Tea uses a threshold of 6 discordant reads (plus at least two clipped reads) to call insertions, while TranspoSeq requires 10 supporting discordant reads. TranspoSeq identifies 55 additional events that Tea misses. These events pass manual review and a subset was chosen for validation.

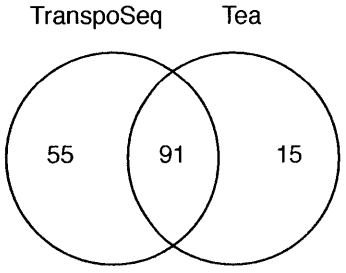


Figure 2-7 Comparison to other methods.
Number of somatic retrotransposon insertions identified in a TCGA colorectal cancer using TranspoSeq (left) and Tea (E. Lee et al. 2012) (right).

Somatic specificity

To determine whether these events are truly somatic, or tumor-associated, and not just random

noise and variation one would expect by comparing two samples from the same individual, we swapped tumor and normal BAM files 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.

Retrotransposon subfamily calling

The aim of the additional assembly step TranspoSeq performs is to more accurately identify the exact retrotransposon element inserted at a given position. To gauge our element family and subfamily calling performance, we took the set of germline insertions identified as known polymorphisms in dbRIP, and compared TranspoSeq's subfamily call with the annotated subfamily call according to dbRIP. A conservative estimate for TranspoSeq's calling performance is thus 93%. This is a lower bound on subcalling accuracy because many dbRIP annotations are non-standardized, which leads to ambiguity, and because we only determine a subfamily call as 'correct' if it is an exact match to the dbRIP annotation, i.e., AluYa4b is not the same thing as AluYa4 in our assessment.

2.4 TranspoSeq-Exome

Hybrid-capture sequencing entails the selective hybridization of DNA fragments to a given set of 'baits', known sequences from the reference genome that are to be re-sequenced in a particular sample. The TruSeq Exome Enrichment Kit, for example, covers 200,000 exons from 20,800 genes and 62 Mb of total sequence including 5' UTR, 3' UTR, microRNA, and other non-coding RNA. Exome data therefore covers 20 times less of the genome than whole-genome sequencing,

but the regions that it does capture, it typically sequences with a mean coverage of about 120X (compared with whole-genome coverage of 40X).

The territory covered by exome sequencing, though much smaller than whole-genome, is more likely to be functionally relevant (as far is currently known); that is, if a gene exon is altered, it is more likely to have direct functional implications on gene expression and the cell than if an intergenic region is altered. Thus, we decided to analyze exome data for somatic retrotransposon movement in the hopes of finding retrotransposon insertions that disrupt exons and exon borders. For this purpose, we developed TranspoSeq-Exome. We present the first, to our knowledge, tool that interrogates whole-exome sequencing data for novel insertion of retrotransposons.

We modified TranspoSeq to search for novel junctions between retrotransposons and unique genomic sequence using split reads. Instead of aligning all discordant read-pairs to the database of consensus retrotransposon sequences, TranspoSeq-Exome first parses out all clipped reads identified by BWA and aligns the clipped sequence to the database of retrotransposons. Split reads that have >10bp aligning to a retrotransposon with an E-value of 2E-07 or lower are then clustered, processed, and annotated as in TranspoSeq (see Figure 2-8 for a schematic of the method).

We assessed TranspoSeq-Exome's performance by analyzing samples with both whole-genome and whole-exome data available. Of the six somatic retrotransposon events found in exonic regions of three LUSC whole-genomes, four are recapitulated in the exome data. The reasons why TranspoSeq-Exome did not identify some of the whole-genome events include: i. insertion

occurred too far from exon and was not covered by exome sequencing, ii. only poly(A) portion of retrotransposon was captured rendering the event undiscoverable (see Limitations below).

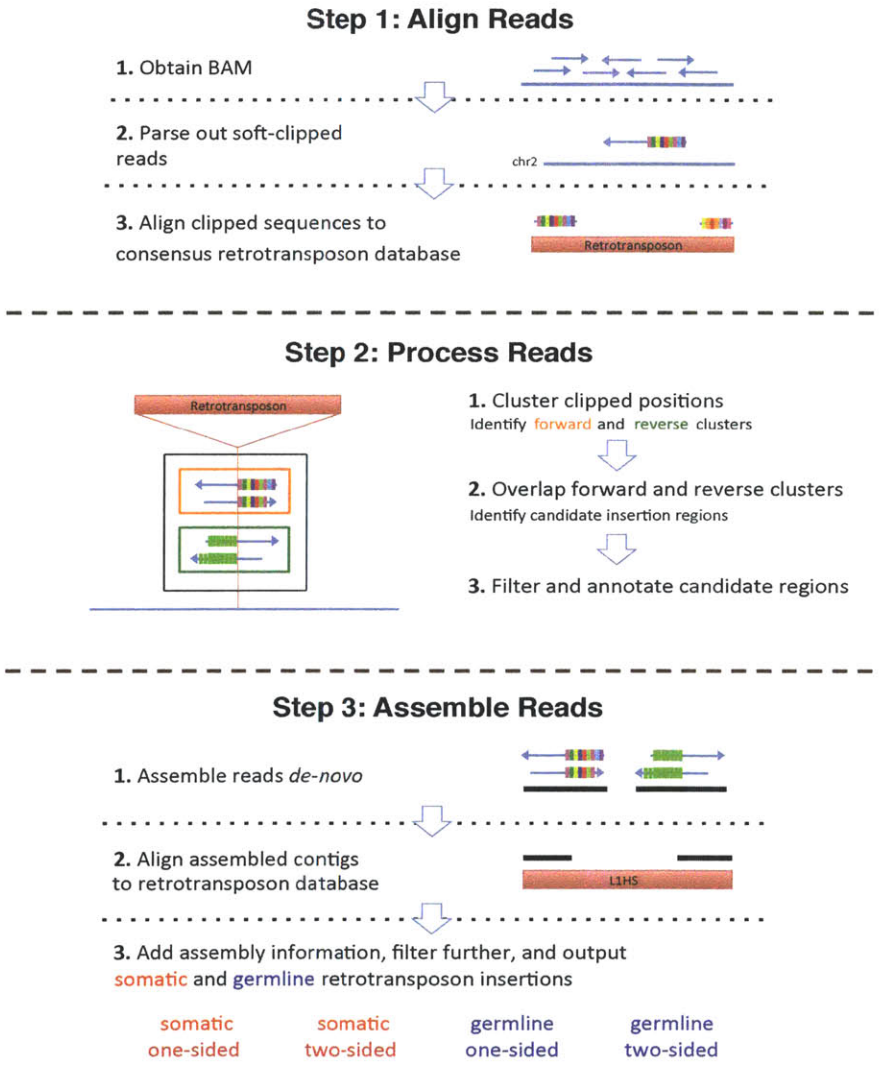


Figure 2-8 Schematic of the TranspoSeq-Exome pipeline.
TranspoSeq-Exome consists of three steps. Get Reads parses tumor and normal BAM files for split reads identified by BWA that are at least 10bp in length. These portions of the read are aligned to the database of consensus retrotransposon sequences 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.

Limitations

One limitation of this technique is that we are only able to identify inserted L1s where the 5' end (even if truncated) of the L1HS is captured, because the poly(A)-containing 3' end does not align significantly to the database. Additionally, the exact base-pair location of a clip can be misidentified by BWA, leading to reduced evidence for an insertion breakpoint. For this analysis, we focus on L1 insertion in exome data.

2.3 Experimental validation

Putative somatic retrotransposon insertions identified by TranspoSeq were first validated in an independent cohort of 9 colorectal tumor/normal samples (Bass et al. 2011). To experimentally validate our candidate somatic retrotransposon insertions, we performed long-range targeted PCR across the putative insertion breakpoint (Figure 2-9). Two primer pairs for each candidate insertion were designed using Primer3 (Rozen & Skaletsky 2000) to hybridize on either side of the putative insertion breakpoint so that the entire inserted element is amplified. Primers were first tested on human cell line (J-82) to determine sufficient hybridization on the reference (non-insertion) locus. If both primer pairs did not produce the expected product (short sequence between primers without an insertion) in the cell line, this candidate site was excluded from further analysis. The passing primers were then used to amplify the region across the insertion breakpoint for both the tumor and matched normal DNA. Following the protocol in Stewart et al. (2011), we used 25 ng of template DNA, 200 uM dNTPs, 5 ng of each primer, 2.5 ul of 10X La PCR Buffer II (Mg^{2+} plus), and 0.5 uL of LA Taq in a 25 ul reaction. PCR was performed on a PTC-225 Peltier Thermal Cycler under the following conditions: initial denaturation at 94°C for

90 seconds, followed by 35 cycles of denaturation at 94°C for 20 sec, annealing at 58°C for all primers for 20 sec, and extension at 68°C for 8 min 30 sec, followed by a final extension step at 68°C for 10 min. 8ul of each PCR product was size-fractioned in a 1.5% agarose gel containing 2ul of 10mg/ml ethidium bromide for approximately 30 minutes at 250V. UV-fluorescence was used to visualize the DNA fragments and images were saved on the AlphaView AlphaImager system.

Samples that passed validation showed a PCR product of approximately the predicted insertion size in the tumor while the matched normal showed only a single band of the size of the amplicon without an insertion (the genomic distance between the two primers). Of the 33 insertion candidates we tested, 31 displayed heterozygous insertion with an insertion allele and an 'empty' allele. Several gels contained a third, faded band above the insertion allele suggestive of heteroduplex formation. The two events that did not pass validation did not produce product in the matched normal sample, leaving us unable to determine whether or not these events were truly somatic.

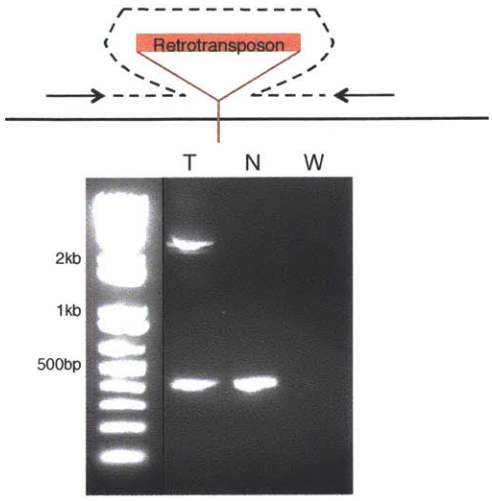


Figure 2-9 Experimental validation.
Top, schematic of long-range PCR validation technique. Bottom, example somatic retrotransposon insertion, wherein a PCR product is present at both 400bp and ~2.5kb in the tumor sample (T), and only at both 400bp in the matched normal sample (N), with a water control (W).

Sequencing

Sanger sequencing was performed on three candidate events through either cloning or direct extraction of the PCR product from the gel. Sequencing was performed using two different techniques to purify the desired PCR product: gel extraction and purification followed by TOPO TA cloning, and Life Technologies E-Gel Size-Select. Gel extraction was performed following the protocol in the Qiagen Quik Gel Purification Kit (Qiagen) and DNA from each band was then used for Life Technologies TOPO TA cloning to further purify one clone of the long PCR product and sent to Beckman Coulter Genomics (www.beckmangenomics.com) for sequencing. For one sample, the Life Sciences E-Gel Size Select CloneWell system was used to purify the larger PCR products directly and sent for Sanger sequencing at GeneWiz (www.genewiz.com).

Resulting forward and reverse sequences were aligned to the retrotransposon consensus database. Sequences from matched normal samples aligned to the reference genome, whereas the tumor sequences aligned to the L1HS element. The three tumor sequences aligned to the L1HS element (E=0.0) and displayed the predicted target site duplications (TSDs) associated with target-primed reverse transcription. See Figure 2-10 for an example sequenced somatic insertion. The size of the insertion, as well as the aligned positions on the retrotransposon and the inversion status, were computed using resulting sequence information and compared with the computationally predicted values. We find that the predicted alignment positions on the retrotransposon from short contigs do differ from those obtained from longer sequences, as is to be expected, but our computational predictions are all within 1kb of the 'true' size, with one prediction only off by 77bp.

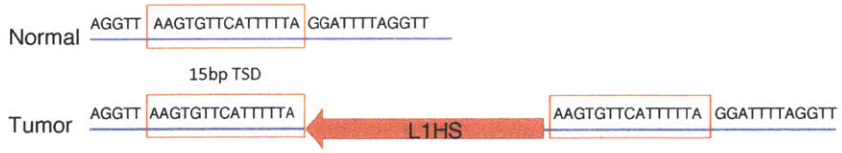


Figure 2-10 Sequencing of validated insertion.
Example of Sanger sequencing result of an inverted L1HS insertion in the tumor sample, but not in normal. The somatic insertion exhibits canonical 15bp TSDs surrounding insertion site (boxes) and the L1 endonuclease motif.

Second round of validations on TCGA samples

We carried out additional validation experiments on the TCGA samples analyzed in this thesis. We performed site-specific PCR on 48 putative retrotransposon insertions across lung squamous cell carcinoma, lung adenocarcinoma, endometrial carcinoma, and head and neck squamous cell carcinoma, including two germline events, one putative somatic SVA insertion, three full-length

somatic L1 insertions, and five somatic L1 insertions identified through TranspoSeq-Exome, one of which is an insertion into the *PTEN* tumor suppressor. These rounds of validation consisted of a slightly modified protocol than the initial validations. Here, primer sets were designed to target both the 5' and 3' junctions (spanning unique reference and inserted retrotransposon sequence) of each putative event (Figure 2-11). Amplification products were then sequenced via Illumina paired-end sequencing to produce the exact sequences that span the insertion breakpoints. This enabled us to verify not only the presence of the insertion at the predicted location, but also the TSD sequences, poly(A) tracts, and precise 3' retrotransposon sequence (and potentially element subfamily).

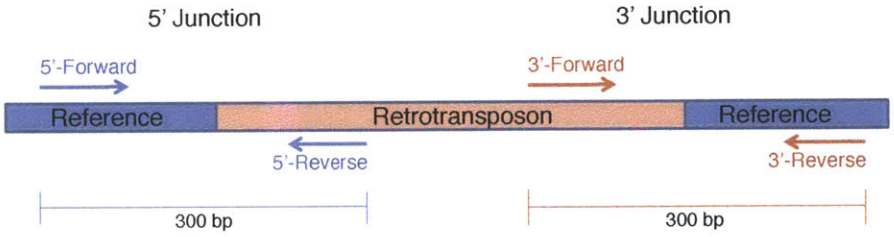


Figure 2-11 Schematic for second round of experimental validations.
5' and 3' junctions of putative retrotransposon insertions will be targeted via PCR surrounding each breakpoint, producing an expected size fragment of 300bp.

A pilot study examining this method of validation was first conducted on 5 candidate insertions – two germline and three somatic. Figure 2-12 shows the PCR results of one germline and somatic insertion, where the expected amplification products are produced for both the 5' and 3' junctions. Namely, in the candidate somatic event, the normal sample does not produce any PCR product for either 5' or 3' reference-retrotransposon junctions, confirming this event as likely tumor-specific.

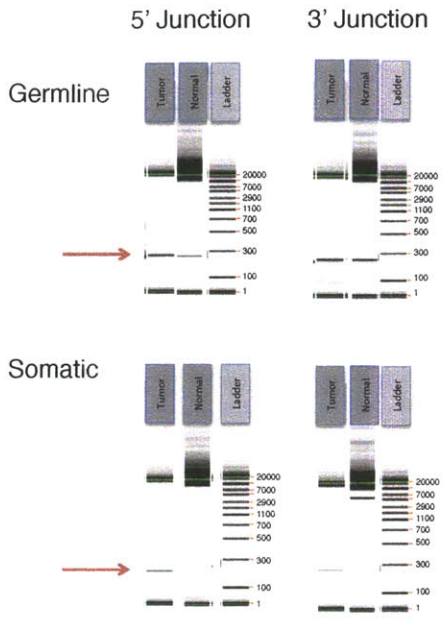
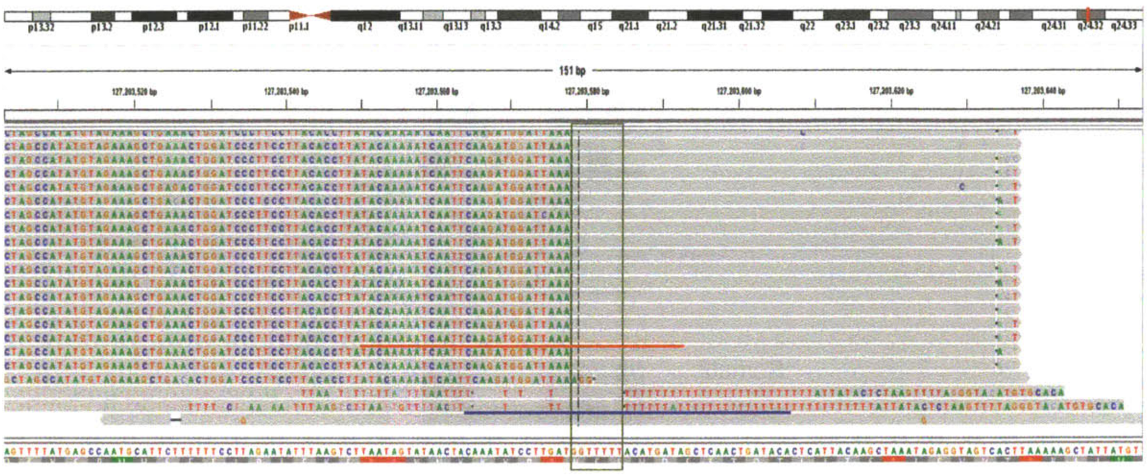


Figure 2-12 Pilot validation results.
Results for pilot PCR validation of one germline (top) and one somatic (bottom) event. In the top panel, the red arrow points to bands present in both tumor and matched normal at approximately 300bp for 5' and 3' junctions. In the bottom panel, the red arrow points to a ~300bp band present in the tumor but not in matched normal for both junctions.

Subsequent sequencing across these junctions in the pilot study confirmed TSD sequences and the computationally predicted length of the inserted element; although we still cannot rule out internal inversion or insertion events within the inserted retroelement. Additionally, although targeted sequencing can give a better idea of the length of poly(A) tracts, these may still remain ambiguous due to low-complexity alignments, such as in Figure 2-13.



5' Junction:

AGCTAATATTTTTGACCACTTACTATATTCTAGCCACTTTACTAAGCATTGTACATACATGATCTACTGAATCCCCACAATCATTTTGTGAGGTATGCATTTTTAA
CATTTCCATTTTAAAGACGAGAAAGGGAAGGATTAGAGGGACTAAATACTTACAGAAGAAAATAAATCAATATGCAAAGAGAAACAAAATTTGTAGATGGG
GGAAGAGAACAACCAAGATAGAGAAGAGATCATTTGATTTGACTCCTAGATGAAACAATACCTGAAGCTATCATCATCCCTAAACCTCAGTTTTATGAGCCAA
TGCATTCTTTTTTCCTTAGAATATTTAAGTCTTAATAGTATAACTACAAATATCCTTGAT|GGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTATTATACTCTAAGTTTTAGGGTACATGTGCACATTGTGCAGGTTAGTTACATATGTATACATGTGCCATGCTGGTGCACTGCACCCACTAATGTGTCAT
CTAGCATTAGGTATATCTCCCAATGCTATCCCTCCCCCCTCCCCGACCCCACCACAGTCCCCAGAGTGTGATATTCCCCTTCCTGGGTCCATGTGATCTCATTGT
TCAATTCCCACCTATGAGT

3' Junction:

ATCACACTACCTGACTTCAAACTATACTACAAGGCTACAGTAACCAAAACAGCATGGTACTGGTACCAAAACAGAGATATAGATCAATGGAACAGAACAGAG
CCCTCAGAAATAATGCCGCATATCTACAACTATCTGATCTTTGACAAACCTGAGAAAAACAAGCAATGGGGAAAGGATTCCCTATTTAATAAATGGTGCTGGG
AAAACTGGCTAGCCATATGTAGAAAGCTGAAACTGGATCCCTTCCTTACACCTTATACAAAAATCAATTCAAGATGGATTAAAGGTTTT|TACATGATAGCTCA
ACTGATACACTCATTACAAGCTAAATAGAGGTAGTCACTTAAAAAGCTATTATGTTTCCCCACTGCCCAGTGTCAATACACAGAAGTAATGTCAATCACTTTTTC
TCCCCCAGGTCTCAACACGTGGCAGTAGGCAGCAGAATTAATTCCCTGTGCCTCCTGCAATTCCCTGAGCCTCCCTTTCTCAAAACGCCAGCTTTCTCACAAGG
ATAGCCTGAGACATACGGAAGTGGGAATGATGTTCTGTTCTCAGGTGAAGTGTTTGTGCATATGTGTGTGTCGGGGGAGAGGAAGGT

Figure 2-13 Next-generation sequencing across putative insertion breakpoints. Top, IGV view of sequencing alignments from validation sequencing. Bottom, resulting sequence across 5' and 3' junctions. Pink text represents retrotransposon sequence, while black represents reference genome. Green boxes indicate TSD sequence, and blue and red lines indicate 5' and 3' junctions, respectively.

With the successful completion of the pilot study, validations were carried out on a set of 47 putative somatic retrotransposon insertions, across 21 individuals and 4 tumor types, including 5 somatic insertions identified from exome data, as well as 4 predicted germline transpositions. Primers for site-specific PCR were designed using Primer3 (Rozen and Skaletsky 2000) to span the 5' and 3' junctions of candidate insertions for tumor and matched normal samples. PCRs were performed with 3ul of 2.5ng/ul DNA, 5 ul of 1uM mixed primers, 0.08 ul of 100mM dNTPs, 0.04 ul Hotstart Taq, 0.4 ul of 25mM MgCL2 and 1ul of 10X buffer, with 1.47 ul of

dH20 for a total reaction volume of 11 ul. The reactions were run with a hot start of 95°C for 5m, then 30 cycles of 94° for 30s, 60° for 30s and 72° for 1m, followed by a final cool-down at 72° for 3min. 2ul of each PCR reaction was run on a caliper to visualize PCR amplicons. Initial PCRs underwent 8 cycles of a tailing reaction to add adapters and indexes for sequencing and run on Illumina MiSeq with single 8 bp index and standard Illumina sequencing primers, resulting in 250bp paired-end reads and insert size approximately 320bp and a coverage of ~200X.

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.

Importantly, we are able to validate the insertion of an L1 element into a PTEN exon identified in an endometrial carcinoma sample, but not in the matched normal through exome sequencing (Figure 2-14).

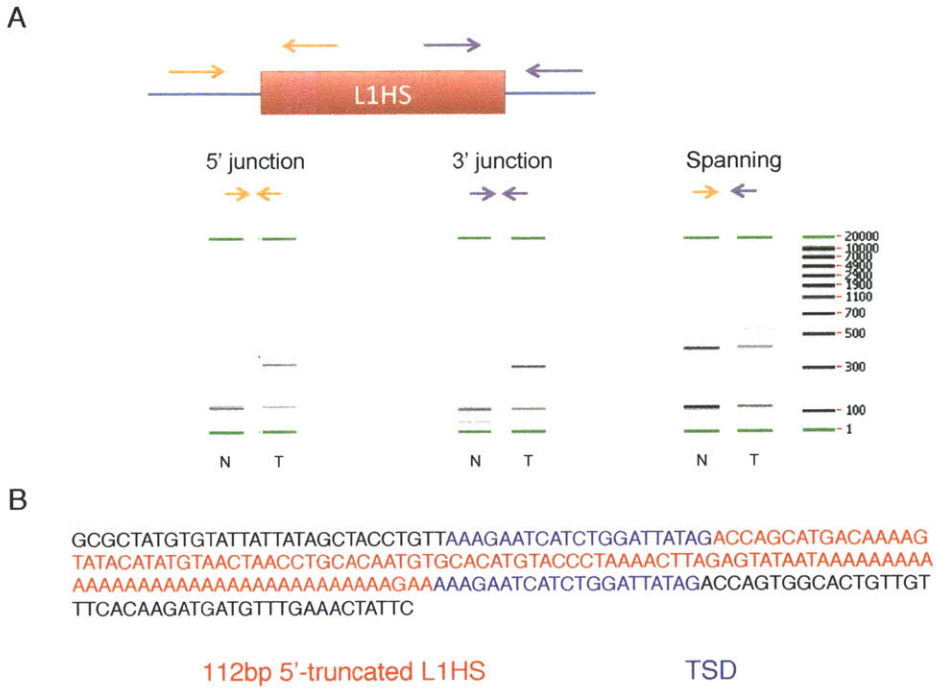


Figure 2-14 Site-specific PCR confirms presence of retrotransposon insertion in *PTEN* exon.

(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.

2.4 Summary

We developed TranspoSeq to computationally mine paired-end whole genome sequencing data across cancers for evidence of non-reference retrotransposon insertions. This tool utilizes multiple alignment approaches and clustering heuristics, as well as parallelized computing to produce annotated lists of putative germine and somatic retrotransposon insertions. We next modified the TranspoSeq framework to split-read information from whole-exome sequencing data and identify exonic retrotransposon insertions. We assessed TranspoSeq's performance on simulated data and compared it to similar methods such as Tea (E. Lee et al. 2012). Furthermore, we experimentally validated TranspoSeq on an independent cohort of 9 colorectal tumors to produce a 94% validation rate. Finally, a second round of validations is currently underway to experimentally verify somatic events discussed in Chapter 3.

Chapter 3. Landscape of retrotransposon insertions across human cancer

By leveraging large-scale sequencing datasets and the necessary computational tools developed in Chapter 2, the landscape of retrotransposon insertion events across tumor samples may now be assessed for the first time. Characterizing the differences in retrotransposition rates and features between tumor types has the potential to reveal insights into the etiology of the disparate diseases and elucidate aspects of tumor biology previously unknown. It will also help tailor future studies of retrotransposition to those cancer types that display active retrotransposition.

3.1 Data

To determine the extent of somatic retrotransposon activity across cancer, we applied TranspoSeq to whole-genome sequencing data from 200 tumor and matched normal samples collected and sequenced through The Cancer Genome Atlas across 11 tumor types: 40 lung adenocarcinoma (LUAD), 19 lung squamous cell carcinoma (LUSC), ovarian carcinoma (OV), 2 rectal adenocarcinoma (READ), 3 colon adenocarcinoma (COAD), 20 kidney clear cell carcinoma (KIRC), 17 uterine corpus endometrioid carcinoma (UCEC), 28 head and neck squamous cell carcinoma (HNSC), 36 breast carcinoma (BRCA), 18 acute myeloid leukemia (LAML), and 20 glioblastoma multiforme (GBM).

Binary alignment (BAM) files were downloaded from The Cancer Genome Atlas CGhub repository at https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp. Sequencing was performed on the Illumina Genome Analyzer IIx

(http://www.illumina.com/systems/genome_analyzer_iix.ilmn). These files were then put through quality control and cleaning and uploaded using the Broad Institute's Firehose pipeline (www.broadinstitute.org/cancer/cga/firehose). BAM files had an average of 40X and 50X coverage, respectively, for tumor and matched normal samples. Mean sequencing fragment lengths were ~375bp with a standard deviation of 150bp. See Figure 3-1 for an example fragment length distribution.

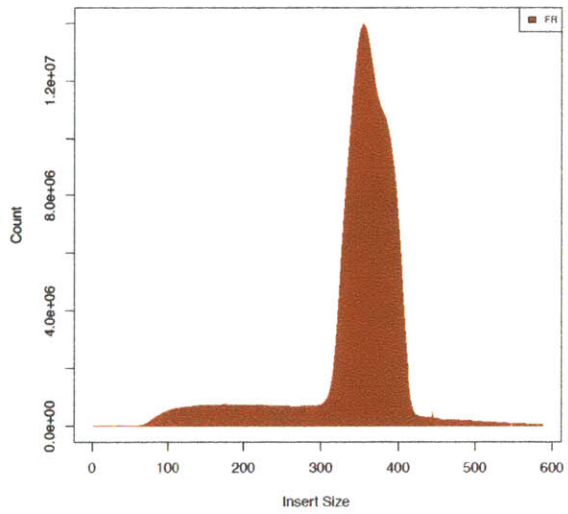


Figure 3-1 Sample fragment length distribution for whole-genome sequencing data. Number of sequencing read-pairs with a given insert size (bp).

Exome data was collected from 767 tumor and matched normal pairs from the three tumor types with the highest rates of retrotransposition in the WGS studies: 199 LUSC, 327 HNSC, 241 UCEC. These BAM files have an average coverage of 120 in tumor and normal samples, and a mean sequencing fragment length of ~125bp. See Figure 3-2 for an example fragment length distribution of exome data.

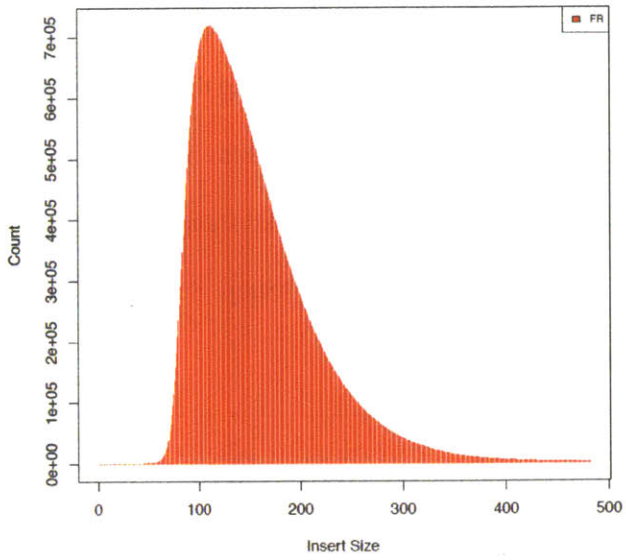


Figure 3-2 Sample fragment length distribution for whole-exome sequencing data. Number of sequencing read-pairs with a given insert size (bp).

3.2 Germline retrotransposon insertions across individuals

Normal genetic variation

We identified 7,724 unique, non-reference germline insertion sites seen in both tumor and matched normal samples. The number of non-reference germline retrotransposon insertions per individual was on average 880 +/- 275, consistent with previous population and cancer studies (Stewart et al. 2011; E. Lee et al. 2012). All putative retrotransposon insertion events were assessed for presence of target site duplications (TSDs) and endonuclease consensus motifs. Figure 3-3 shows the distribution of lengths of duplications or small deletions at the insertion target site. A negative length here represents the number of bases that are deleted at the point of insertion. There is a distinct peak at 15bp, which is expected from the typical TPRT, indicating that TPRT is likely the predominant mechanism of retrotransposition in the germline.

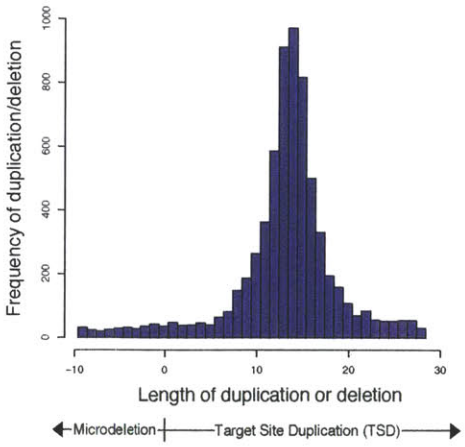


Figure 3-3 Germline TSD lengths.
Distribution of sequence lengths at target site of germline retrotransposon insertions.

We next asked whether germline retrotransposons tend to insert into a preferred sequence pattern. Insertion motifs were determined from assembled contig sequences for both strand directions. Four basepairs on either side surrounding the putative insertion breakpoint across all germline breakpoints were used as input to the MEME motif finder (T. L. Bailey et al. 2009) with an optimal motif width of between 2 and 6bp long (inclusive). Figure 3-4 depicts the resulting motif discovered. The 'TTAAAA' pattern, with some degeneracy in the second and fourth positions, represents the canonical L1 endonuclease target motif, again providing evidence for traditional TPRT as the mechanism behind recent germline retrotransposition.

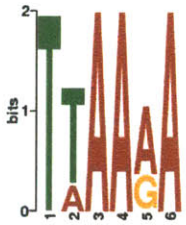


Figure 3-4 Germline insertion motif.
Enriched sequence motif at insertion breakpoints of germline retrotransposon insertions.

Length of germline L1 insertions

An additional hallmark of TPRT is the prevalence of L1 truncation at the 5' end. The L1 reverse transcriptase, in addition to functioning with very low fidelity, also tends to detach before completion of the transcript, resulting in the insertion of a 5'-truncated L1 element. L1s that are significantly truncated cannot catalyze additional retrotransposition because their retrotransposition machinery encoded in ORF1 and ORF2 is dysfunctional. Full-length, but not 5' truncated, L1 insertions are deleterious and subject to negative selection (Boissinot et al. 2006).

We can computationally assess the length of L1 elements inserted in the genome by determining where along the consensus retrotransposon sequence the supporting reads align. Specifically, we can determine the start and end positions of the inserted retrotransposon element; we cannot determine the entirety of the sequence, however, so if there is an insertion or inversion inside the retrotransposon, paired-end sequencing data will not reveal this. From the 200 individuals analyzed here, it appears that recent germline L1 insertions in the population are often 5'-truncated; however, many do represent full-length L1 insertions, retaining the potential to retrotranspose (Figure 3-5).

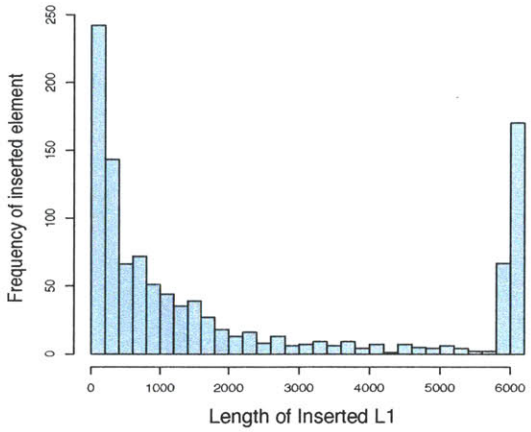


Figure 3-5 Distribution of germline L1 insertion lengths.
Length (bp) of L1 elements inserted in both tumor and matched normal samples.

Known polymorphisms

To determine whether the germline retrotransposon insertions we identify are known retrotransposon insertion polymorphisms (RIPs), we collected annotated data from dbRIP (J. Wang et al. 2006) and nine other germline and somatic retrotransposon studies: Xing et al. (2009), Beck et al. (2010), Huang et al. (2010), Hormozdiari et al. (2010), Witherspoon et al. (2010), Iskow et al. (2010), Ewing et al. (2010), Ewing and Kazazian (2011), Stewart et al. (2011), Lee et al. (2012) (E. Lee et al. 2012; Beck et al. 2010; Huang et al. 2010; Hormozdiari et al. 2010; Xing et al. 2009; Iskow et al. 2010; Witherspoon et al. 2010; Stewart et al. 2011; Ewing & Kazazian 2010; Ewing & Kazazian 2011). Of the 7,724 non-reference insertions we identified across 200 samples, 65% are known retrotransposon insertion polymorphisms annotated previously. The fact that TranspoSeq is able to identify known retrotransposon insertions (of various elements including short Alus) serves to validate the methodology. Many of the 2,703 novel germline retrotransposon insertions identified here represent previously unannotated *common* polymorphisms, present in as many as 114 individuals (Figure 3-6), though the majority

are rare events seen only in a few individuals. Additional, large-scale studies of retrotransposon insertions in individuals are needed to comprehensively annotate the diversity of these events across populations.

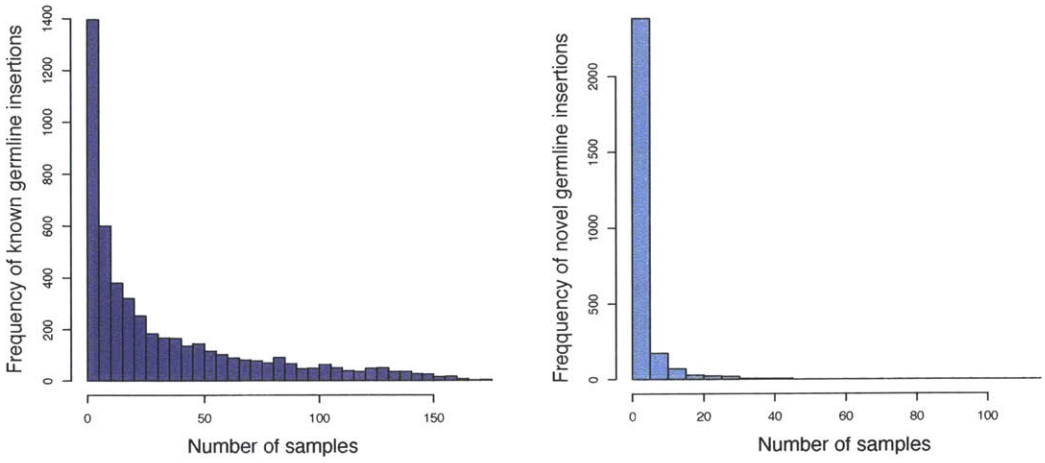


Figure 3-6 Germline retrotransposon insertion polymorphisms.
Number of individuals with specific germline retrotransposon insertion; known (left panel) and novel (right panel).

The L1, Alu and SVA families of retrotransposons are known to retain the capacity to retrotranspose. Though the Alu element hijacks retrotransposition machinery encoded by L1, the rate of Alu retrotransposition, predicted from previous population studies as well as comparative genomics studies, is approximately ten times greater than the rate of L1 (see Section 1.3). And the rate of SVA insertion, although understudied compared to the other two active families of retrotransposons, is approximately five times less than L1 (Xing et al. 2013). When we examine the proportions of the retrotransposon elements that mobilize in tumor and matched normal samples within individuals in our cohort, we see a consistent distribution with the known rates of retrotransposition (Figure 3-7).

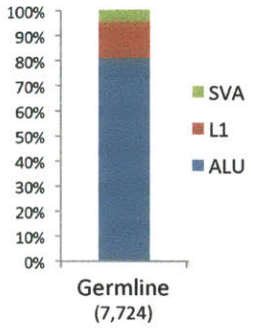


Figure 3-7 Germline element distribution.
Proportion of non-reference germline retrotransposons within each element family.

Genomic distribution

It is unknown whether retrotransposons are limited as to where they are able to insert. Mills et al. (2006) found that human-specific retrotransposon insertions in the reference genome are generally distributed evenly according to the amount of DNA present on each chromosome (Mills et al. 2006). We looked at the genomic distribution of non-reference germline retrotransposon insertions across our 200 samples. Insertions are spread across autosomal and the X-chromosome (Figure 3-8), however, it appears that the X-chromosome actually contains fewer non-reference germline retrotransposon insertions than would be expected from its length (Figure 3-9).

Figure 3-8 Genomic distribution of germline retrotransposon insertions.
Each blue line represents a non-reference retrotransposon insertions seen in at least one of 200 individuals sequenced.

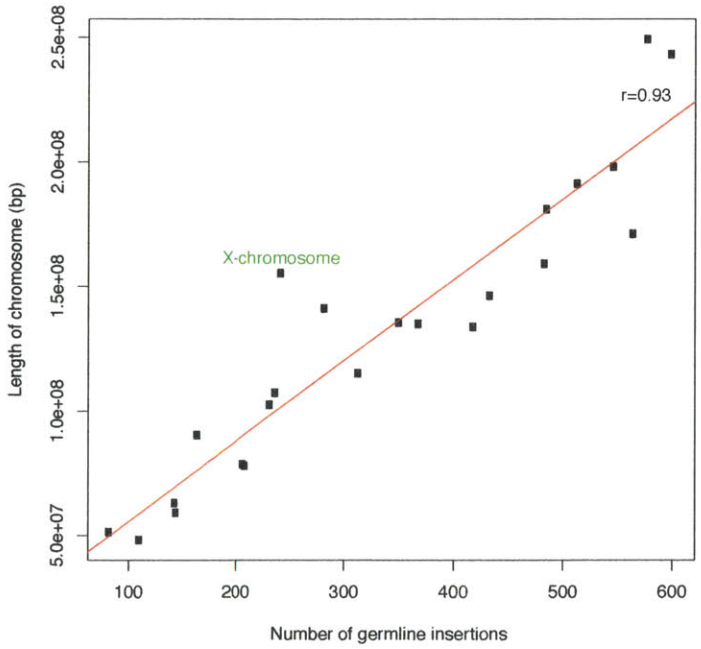


Figure 3-9 Number of germline retrotransposon insertions by chromosome length.
Length of chromosome (bp) is plotted against number of non-reference germline retrotransposon insertions identified in that chromosome. Best-fit line is shown in red, and the outlier X-chromosome, containing fewer germline insertions than would be expected from its length, is highlighted in green.

Genic distribution

Approximately 35% of the germline insertions identified across individuals are located within known gene regions (including 1kb upstream and downstream of Refseq genes). Similarly, this proportion was found in recent, human-specific retrotransposon insertions in the reference (Mills et al. 2006). The proportion of the human genome that comprises genic regions is approximately 35% as well, implying that germline retrotransposons land in gene regions as would be expected by chance if these integrations occurred randomly. However, only ~11% of all human RefSeq genes contain a retrotransposon insertion; many genes thus contain multiple instances of retrotransposon integration, suggesting either non-random insertion or purifying selection.

Recurrent germline genes

Non-reference retrotransposon insertions are recent, human-specific, events not present in even our closest ancestors, like chimpanzees. Germline retrotransposon insertions into genes that we identify are non-lethal, and it is plausible that the common insertion events are well tolerated by the cell. Deleterious insertions with large effects are not likely to survive in the human population (Boissinot et al. 2001); this depends on how common the insertion is in the human population. If an insertion is sufficiently rare, specific to only one individual for example, it may not yet have been subject to selection and thus is more likely to have a non-neutral effect on the cell.

Many of the genes with recurrent germline retrotransposon insertions have been implicated in genetic disorders and cancer. The putative tumor suppressor, low-density lipoprotein receptor-related protein 1B (*LRP1B*) contains 13 different sites of germline retrotransposon insertion, with

a total of 145 individuals in our cohort exhibiting at least one of these insertions. Originally deemed *LRP-DIT*, Deleted In Tumors, this gene has been implicated in ovarian cancer (Cowin et al. 2012), urotherial cancer (Langbein et al. 2002), and lung cancer cell lines (C. X. Liu et al. 2000) and is important in normal cell function and development. However, *LRP1B* is unusually large, spanning 1.9 Mb (600 kDa) and containing 91 exons.

Gene	Number of insertion sites	Total number of individuals with insertion in gene
LRP1B	13	145
CTNNA3	13	124
EYS	11	108
PRIM2	8	176
PCDH15	8	109
GPC5	8	183
ERBB4	8	126
PARK2	7	124
CSMD1	7	116

Table 3-1 Genes with recurrent germline retrotransposon insertions.
Top nine genes containing multiple retrotransposon insertions across individuals.

Catenin Alpha 3 (*CTNNA3*) is another commonly altered gene in germlines, with a total of 13 insertion sites present in 124 individuals. *CTNNA3* stabilizes cellular adherence, a feature that is often compromised in cancer, and has been implicated in urothelial carcinoma of the bladder (Meehan et al. 2007). Similarly, *ERBB4* is commonly deleted in breast cancer (Sundvall et al. 2008). Germline mutations in the *EYS* gene account for some 5% of autosomal recessive retinitis pigmentosa, a degenerative eye disease (Littink et al. 2010). The parkin (*PARK2*) gene encodes an E3 ubiquitin ligase that is the most commonly mutated gene in autosomal recessive Parkinson's Disease and is also a putative tumor suppressor mutated in ovarian, glioblastoma, colon and lung cancer (Veeriah et al. 2009). One recent study implicates germline heterozygous mutations of *PARK2* as predisposing events in lung adenocarcinoma (Iwakawa et al. 2012);

however, it is also a very large (1.4 Mb) gene prone to deletions and mutations (Plun-Favreau et al. 2010), so no claims can be made regarding the prevalence of germline retrotransposon insertions in this gene or many of the other large genes (see Section 4.3).

3.3 Somatic retrotransposons across cancer from whole-genome sequencing

We find a total of 810 retrotransposon insertions present in a tumor and not in the matched normal sample amongst the 200 individuals with whole-genome sequencing data available. These somatic retrotransposition events exhibit the hallmarks of target-primed reverse transcription, such as target site duplications approximately 15bp in length, and a canonical L1-endonuclease motif (Feng et al. 1996; Morrish et al. 2002) at the site of insertion.

Somatic target site duplications and sequence motifs

In contrast to the distribution of germline TSD lengths, there is an additional class of somatic events lacking a TSD, or exhibiting a deletion of several basepairs at the site of insertion (Figure 3-10). These events are most likely distinct from genomic rearrangements because they involve the few active subfamilies of L1s and exhibit 5' truncation and poly-adenylation characteristic of true retrotranspositions. The abundance of events lacking the canonical TSD may suggest a possible alternative mechanism for somatic retrotransposon insertion, in addition to traditional TPRT. One alternate mechanism described previously in Chinese hamster ovary (CHO) cells is termed L1 endonuclease-independent insertion (Morrish et al. 2002; Sen et al. 2007), which involves L1-mediated double-strand break repair. The increased number of somatic retrotransposition events with no TSD or microdeletion was also seen in Lee et al. (2012) and

appears to be a cancer-specific phenomenon. Differential analysis of these two seemingly distinct groups of events (those with canonical TSDs and those without) can be found in Section 4.1.

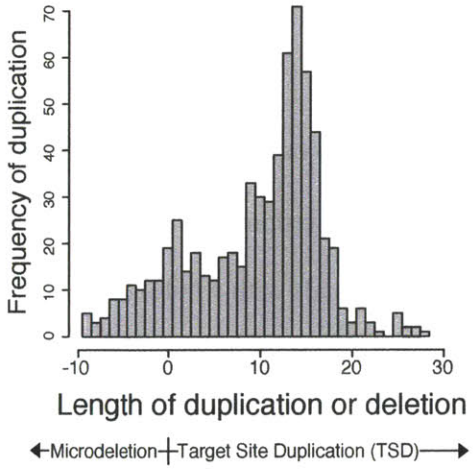


Figure 3-10 Somatic TSD lengths.
Length of sequence at target site of somatic retrotransposon insertion (duplications are shown as events with target site length >0 and microdeletions as those with <=0).

The sequence pattern surrounding the integration site of L1s that are inserted somatically is consistent with the canonical L1 endonuclease motif (Figure 3-11). We obtained this motif as described in Section 3.2 for germline insertions. Although this sequence contains some degeneracy, it retains the established recognition site for L1 endonuclease.

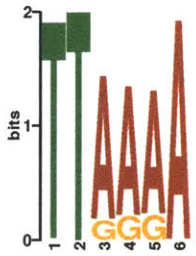


Figure 3-11 Somatic insertion motif.
Enriched sequence motif at site of somatic retrotransposon insertions.

Length of somatically inserted L1s

We calculated the length of somatic L1 insertions computationally using the method described in Section 3.2. In contrast to germline insertions, somatic L1 insertions exhibit a bias toward severe 5'-truncation (Figure 3-12), with most inserted elements less than 400bp in length, consistent with previous studies (Solyom et al. 2012; E. Lee et al. 2012). The extreme truncation may be indicative of the fast pace of replication in the cancer cell, causing the reverse transcriptase to fall off and disengage from the nascent DNA strand before completion of the entire L1. Alternatively, the instability of the cancer cell and altered kinetics of DNA damage repair (Wallace et al. 2010) may impact the final steps of retrotransposition; that is, repair of the DSB that occurs during TPRT begins before reverse transcription is completed, displacing the reverse transcriptase machinery. Since full-length L1 elements are subject to negative selection, this suggests that the cancer cell, which undergoes accelerated somatic evolution, may have selected against these deleterious insertions and retained only 5'-truncated elements in its genome. We do however find several examples of full-length (>=6000bp) L1 insertions in the tumor but not matched normal sample. These elements may retain the capacity to retrotranspose and potentially contribute to the perpetuation of somatic retrotransposition activity.

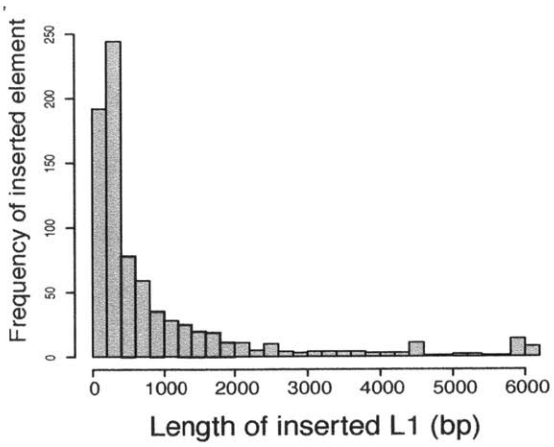


Figure 3-12 Distribution of somatic L1 insertion lengths.
Length (bp) of inserted L1 elements present in tumor but not matched normal.

Allelic fraction

The allelic fraction of a genomic alteration can elucidate its homozygosity as well as its clonality, or the proportion of tumor cells in which it is present. This can provide information about the specific mutation's role in tumorigenesis – alterations occurring earlier in tumor evolution may be important for cell death evasion whereas later events may be responsible for tissue invasion and metastasis.

We compared the allelic fraction of germline retrotransposon insertions to that of somatic insertions (Figure 3-13). As expected, the majority of germline insertions are homozygous events, whereas somatic insertions are heterozygous, at an allelic fraction centered on 0.5.

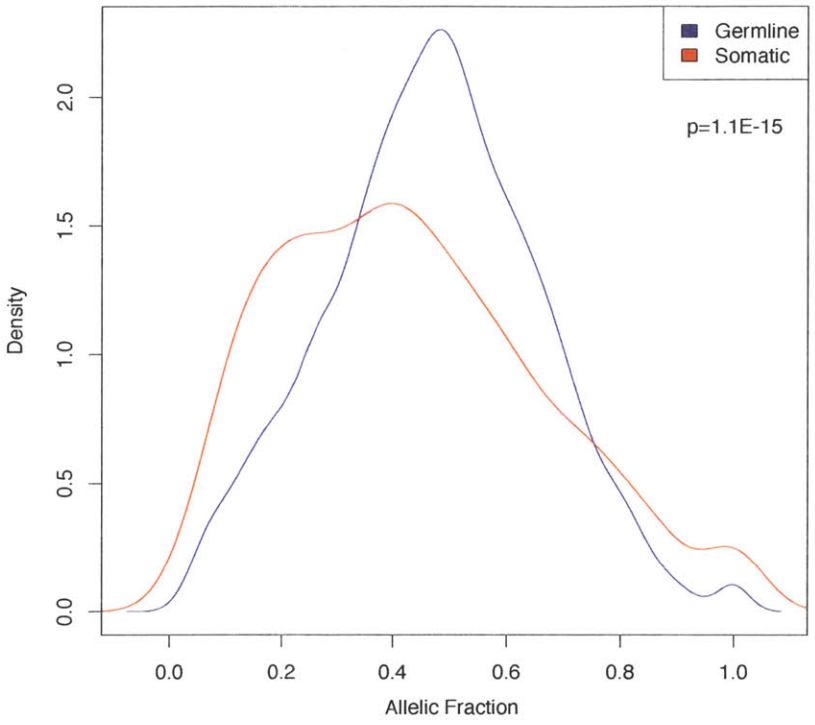


Figure 3-13 Allelic fraction of germline versus somatic retroransposon insertions. Allelic fraction of each insertion event was measured by the ratio of split reads supporting insertion to number of total reads spanning breakpoint.

We asked whether the full-length L1 insertions are relatively early or late events in tumor evolution. To do this, we compared the allelic fraction distribution of somatic L1 insertions (>6000bp) to the distribution of truncated L1 insertions (<6000bp) and find that full-length insertions are skewed toward higher allelic fractions, though this trend is not statistically significant, likely due to the small sample size of full-length events (KS-test p=0.10, Figure 3-14). These events are expected to have larger effect sizes in terms of impact on genomic stability; thus, it appears they are earlier, more clonal events in the progression of cancer. In general, inserted element length does not correlate with allelic fraction (r^2=0.005). Additionally, events present at higher allelic fraction (>0.8) are not enriched for genic insertions compared to those present at lower fractions (Fisher's exact p=0.074).

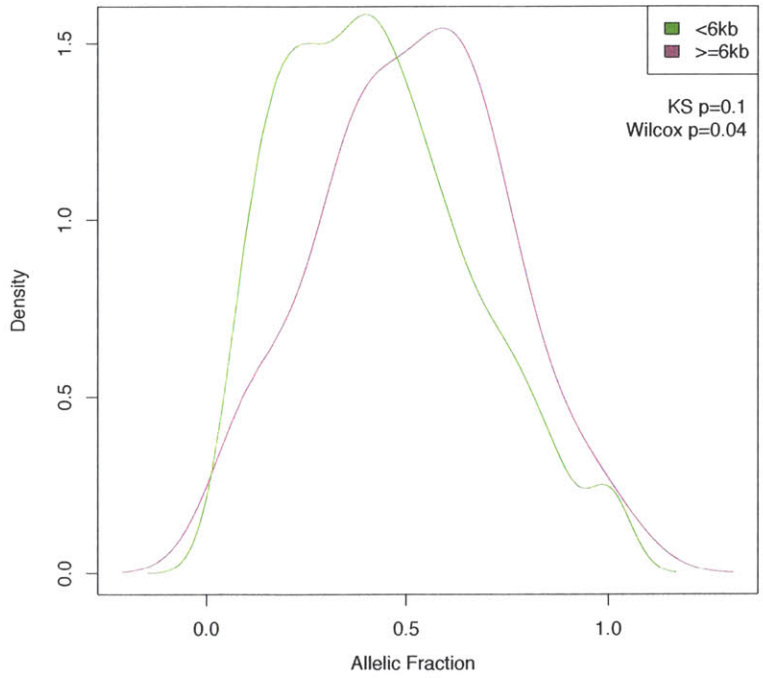


Figure 3-14 Allelic fraction of somatic full-length L1 insertions.
Comparison of allelic fraction of somatic insertions of full-length L1 elements (6000bp, magenta) and distribution of truncated L1s (<6000bp, green).

There are several caveats to this analysis that must be taken into account. Notably, we do not factor in tumor ploidy or purity, so it is difficult to assess cell fraction and subclonality without this normalizing information. Tools such as ABSOLUTE (Carter et al. 2012) and ASCAT (Van Loo et al. 2010) quantify allele-specific copy number alterations from allelic fractions using tumor purity and ploidy estimates; these tools are based largely on SNP array data, although exome sequencing is beginning to be used as input. Methods to normalize genomic rearrangement data in order to assess the cell fraction of these complex events have yet to be developed, but will be important in understanding the role of these events in tumorigenesis.

Somatic element distribution

Consistent with previous reports (E. Lee et al. 2012; Solyom et al. 2012; Iskow et al. 2010), we find that somatic insertions are composed primarily (97%) of L1HS elements, differing significantly from the distribution of germline insertions (Figure 3-15). Despite the rate of Alu insertions far exceeding that of L1s by almost 10 times, we find very few instances of Alu retrotransposition in cancer. The reason for this discrepancy remains an open question. It is unlikely due to an identification bias on the part of TranspoSeq because we are able to detect germline Alu insertions at the expected proportion. Furthermore, every study of somatic retrotransposition in cancer, to date, has confirmed this phenomenon. Since Alu retrotransposition relies on L1 transcription, it is logical that, when it comes to re-activation of mobile elements, L1 would come first. One possibility is that the regulation of L1 elements breaks down, while that of Alu remains intact. Although it is believed that both Alu and L1 elements are commonly hypomethylated in cancer (Choi et al. 2007; Cho et al. 2007), one study found that methylation is inversely correlated with the age of Alu elements, that is, the younger the Alu subfamily, the less likely it is to be demethylated (Rodriguez et al. 2007). Whether this is true for L1 subfamilies is not known, to our knowledge, although it has been shown that certain subsets of L1HS elements are differentially hypomethylated in cancer cell lines (Alves et al. 1996). The mechanism behind L1 reactivation while Alu remains stationary in cancer should thus be investigated further using both genomic and epigenetic data. Notably, we find evidence for one somatic insertion of an SVA element – the first tumor-associated somatic retrotransposition of an SVA to our knowledge.

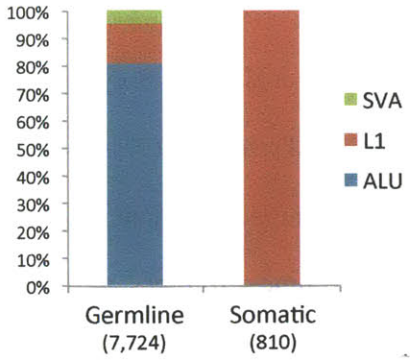


Figure 3-15 Somatic element distribution.
Distribution of retrotransposon elements inserted somatically (right) versus in the germline (left).

The L1 family of retrotransposons consists of many subfamilies, with varying levels of activity (reviewed in Section 1.3). The active subfamily, L1 Ta-1, is distinct from other subfamilies by several diagnostic nucleotides in its 3' UTR. Since TranspoSeq includes the assembly of the 3' end of inserted L1s (and this end is not typically truncated), we are able to assess subfamily specification. Sequence analysis of the somatically inserted L1 elements reveals that they are all in the Ta-1 subfamily; see Figure 3-16 for an example of sequence homology between the 3' ends of six somatic L1 insertions and a reference L1 Ta-1 element.

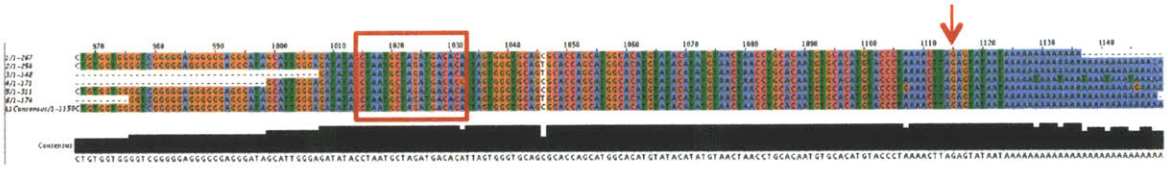


Figure 3-16 Sequence homology of inserted elements.
Sequence homology between reference active L1 Ta-1 sequence (bottom track) and six somatic L1HS insertion sequences. Red box and arrow point to key nucleotides distinguishing transcriptionally active subfamily of L1HS from older, inactive subfamily.

Landscape across cancers

Somatic retrotransposon insertions display a tumor-specific pattern. While GBM, LAML, BRCA, KIRC, OV, and LUAD samples exhibit little or no detected somatic retrotransposition,

LUSC, COAD/READ, HNSC, and UCEC show active mobilization of retrotransposons (Figure 3-17). These findings are in accordance with other studies where L1 insertions were seen in epithelial cancers but not in glioblastomas or blood cancers (Iskow et al. 2010; E. Lee et al. 2012; Solyom et al. 2012). The reason for the tumor-specific pattern of retrotransposition remains unclear. The tumor types of squamous cell origin – HNSC and LUSC – exhibit higher rates of retrotransposition. In fact, there is a stark contrast between the two non-small cell lung cancer types: wherein adenocarcinoma shows almost no retrotransposition compared to lung squamous cell carcinoma's extreme rate. The one outlier LUAD sample with >30 somatic retrotransposon insertions, upon further histological review, appears to be closer to a large-cell neuroendocrine carcinoma (LCNEC) diagnosis. LCNEC is an aggressive form of non-small cell lung cancer (Battafarano et al. 2005), with neuroendocrine differentiation and lacks features of small cell carcinoma, adenocarcinoma, and squamous cell carcinoma (C. A. Moran et al. 2009). There is also wide variation of somatic events amongst individuals within tumor types, with LUSC samples ranging from zero somatic insertions to up to 79 somatic insertions per sample. In Chapter 4, we investigate several possible explanations for these differences.

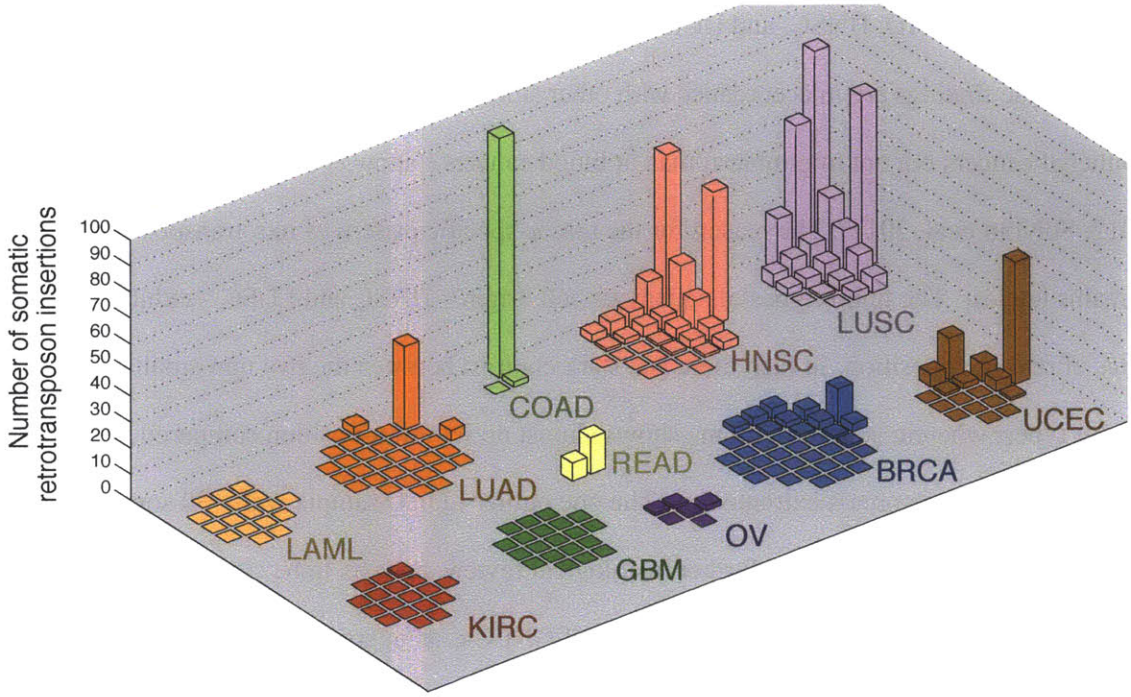


Figure 3-17 Landscape of somatic retrotransposon insertions across cancers. The y-axis represents the number of somatic retrotransposon insertions per individual; samples are arranged by tumor type in order to facilitate visualization.

Genomic Distribution

Earlier studies found enrichment of disease-causing retrotransposon insertions on the X chromosome (Deininger & Batzer 1999; J.-M. Chen et al. 2005; Belancio et al. 2008), possibly due to an ascertainment bias from X-linked disorders. It is also possible that L1s preferentially insert into the X chromosome perhaps because of a proposed involvement in X inactivation (J. A. Bailey et al. 2000; Chow et al. 2010) where L1 elements help spread silencing signals. We find cancer-associated somatic events to be evenly distributed across the autosomal and X chromosomes (Figure 3-18). The distribution of retrotransposon insertions across chromosomal arms significantly differs between germline and somatic events (Wilcoxon p=3.706e-08). Specifically, the short arm of chromosome 4 has a 1.6-fold enrichment compared to a null

distribution of somatic retrotransposon insertions, differing from germline insertions in that arm (Fisher's p=0.0087). Chromosome 4p loss has previously been associated in various cancer types (Arribas et al. 1999; Shivapurkar et al. 1999; Polascik et al. 1995).

Figure 3-18 Genomic distribution of somatic retrotransposon insertions.
Red marks represent precise loci of non-reference somatic retrotransposon insertions across all chromosomes examined (autosomes and X-chromosome).

Genic distribution

Retrotransposons have the capacity to mobilize into genes and surrounding regulatory regions to affect gene expression and disrupt protein function; these insertions have previously been implicated in cancer. Most recently, Shukla et al. (2013) (Shukla et al. 2013) discovered an L1 insertion into *ST18* in hepatocellular carcinoma that resulted in overexpression of the gene. We find that the proportion of somatic retrotransposon insertions into genes is similar to that of germline events, where approximately 35% of events falling in genic regions, including 1kb upstream and downstream of the gene. Again, this proportion is expected given the genic

composition of the genome (see Section 3.2); however, only ~1% of RefSeq genes contain a retrotransposon insertion, strongly suggesting either non-random insertion or purifying selection for or against some genes.

Recurrent genes

Recurrent mutations in a gene across multiple tumors may be suggestive of either a role in tumorigenesis or a systematic artifact due to myriad possible reasons. Here, we find several genes that are recurrently disrupted by retrotransposon insertions in multiple samples across tumors and tumor types (Table 3-2). Contactin-associated protein-like 2 (*CNTNAP2*) is the most highly recurrent, with a somatic L1 insertion in 4 individuals – three lung squamous cell carcinomas and one uterine carcinoma. Variants in *CNTNAP2* have been associated with autism spectrum disorder, schizophrenia, and other neurological disorders (Rodenas-Cuadrado et al. 2013) but it is also 2.3Mb long and mutated in 3.6% of COSMIC samples across cancers (D et al. 2010). Some genes that have multiple instance of germline retrotransposon insertion also have recurrent somatic insertions in cancer, e.g., *EYS* possesses two instances of somatic insertions, in addition to the 13 germline insertions. Similarly, CTNNA3 was the most commonly altered (through germline retrotransposition) gene, while closely related, catenin alpha 2 (*CTNNA2*) harbors multiple somatic retrotransposon insertions across LUSC, HNSC, and COAD.

Gene	Samples with somatic insertion
CNTNAP2	LUSC-43-3920; LUSC-60-2711; LUSC-66-2766; UCEC-A5-A0GA
CTNNA2	LUSC-21-1076; HNSC-BA-4076; COAD-AA-3518
MDGA2	LUSC-60-2711; LUSC-66-2766; HNSC-BA-4076
AGMO	LUSC-43-3394; LUSC-60-2711
ARHGAP15	LUSC-60-2698; HNSC-CR-6487
BBS9	LUSC-60-2713; UCEC-A5-A0GA
CSMD1	HNSC-CV-7180 (2)
DLG2	LUSC-60-2713; HNSC-BA-4076
EYS	LUSC-60-2698; COAD-AA-3518
FAM19A2	LUSC-43-3920; UCEC-A5-A0GA
LRRTM4	UCEC-A5-A0GA; HNSC-CR-6472
MAGI2	LUSC-60-2724; HNSC-CV-5442
PDE4B	LUSC-60-2711; UCEC-AP-A052
RIMS1	HNSC-CN-5374; COAD-AA-3518
SEMA3E	LUSC-60-2711; LUSC-66-2766
DAB1	LUSC-34-2600; LUAD-38-4630
GRID2	HNSC-CV-7255; OV-25-1319

Table 3-2 Genes with somatic retrotransposon insertions in more than one sample. Samples containing a retrotransposon are colored according to tumor type.

Cancer genes

A closer look at the specific genes that contain somatic insertions reveals several known cancer genes, such as *RUNX1*, a putative tumor suppressor in gastric carcinoma (Silva et al. 2003) that is subject to recurrent loss-of-function inactivation in breast cancer and esophageal adenocarcinoma (Banerji et al. 2012; Dulak et al. 2012; Koboldt et al. 2012), as well as in the exon of *REV3L*, which has been implicated as a novel tumor suppressor in colorectal and lung cancers, and is involved in maintenance of genomic stability (Zhang et al. 2012; Brondello et al. 2008). One UCEC sample contains an intronic somatic L1 insertion in the *ESR1* gene, an important hormone receptor often overexpressed in endometrial and breast cancers (Lebeau et al. 2008). Ankyrin repeat domain 18A (*ANKRD18A*) has recently been found to be specifically inactivated by promoter hypermethylation in lung cancer (W.-B. Liu et al. 2012). And

Phosphodiesterase 4B (*PDE4B*) is another tumor suppressor candidate gene that has multiple somatic retrotransposon insertions; its downregulation activates protein kinase A and may contributes to the progression of prostate cancer (Kashiwagi et al. 2012). The fact that we see retrotransposons inserted somatically into candidate tumor suppressor genes implicated in cancer indicates that retrotransposition may be a possible mechanism for recessive cancer gene inactivation.

While previous studies found somatic insertion only in intronic regions, we identify 21 somatic events in or within 200bp of exons of genes such as *CYR61* and *HSF2*, with seven falling in the protein coding sequence itself (Table 1-3). Thus, the cancer cell does not necessarily select against exonic insertions.

Sample	Gene	Region
LUAD-38-4630	*CYR61*	Exon 4
LUSC-43-3920	*REV3L*	Exon 12
LUSC-60-2726	*ZNF267*	Exon 4
LUSC-66-2766	*HSF2*	Exon 10
LUSC-66-2766	*PBLD*	Exon 3
HNSC-CR-6470	*ANKRD18A*	Exon 15
HNSC-CV-6433	*GUCY1B2*	Exon 4
COAD-AA-3518	*GPATCH2*	3' UTR
LUSC-60-2698	*C20orf107*	3' UTR
HNSC-CV-5442	*TRDMT1*	3' UTR
LUSC-60-2698	*DHRS7B*	13bp before exon3
HNSC-BA-6873	*TNIP3*	22bp before 5'UTR
LUSC-66-2766	*C3orf33*	59bp before exon4
HNSC-BA-6873	*ERO1L*	94bp after exon10

Table 1-3: Somatic retrotransposon insertions into exonic regions revealed by whole-genome sequencing.
Table of somatic retrotransposon insertions into exonic regions of genes, including up to 100bp before or after coding exon. These were identified from whole-genomes across tumor types.

3.4 Somatic retrotransposon insertions from whole-exome data

Landscape in exome data

Since whole-genome sequencing revealed several somatic L1 insertions into coding regions, we examined somatic retrotransposition in whole-exome data from the three tumor types with high retrotransposition activity in whole-genomes (LUSC, UCEC and HNSC) using TranpoSeq-Exome. Figure 3-19 shows the distribution of somatic retrotransposon insertions discovered in exome data. The LUSC cohort has several samples with multiple exonic retrotransposon insertions, whereas HNSC and UCEC have only a few samples with single events.

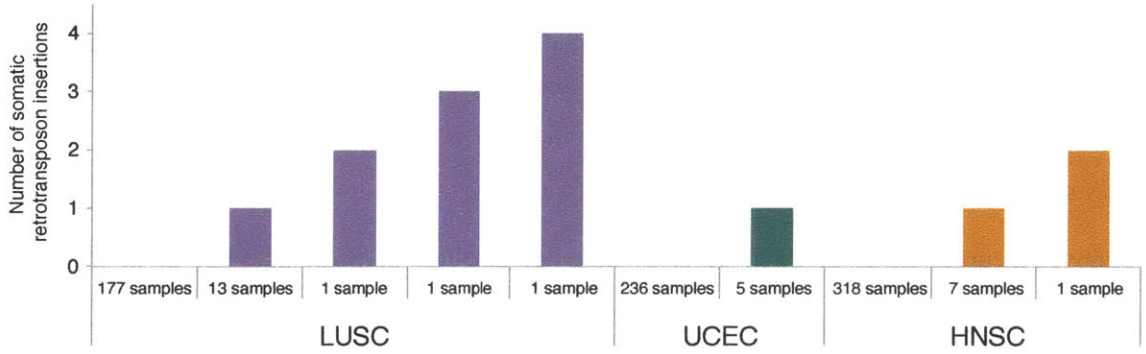


Figure 3-19 Landscape of somatic retrotransposon insertions in whole-exome sequencing data.
Number of somatic retrotransposon insertions found in LUSC, UCEC and HNSC whole-exomes.

Recurrent genes in exome data

When we add somatic events revealed through exome sequencing to those discovered by whole-genome, we discover several genes that have somatic retrotransposon insertions across multiple samples (Table 1-4). Genes such as Crumbs homolog 1 (*CRB1*), mutations of which are known to cause Leber congenital amaurosis (Lotery et al. 2001), appear as recurrently affected by retrotransposition with the additional power of the exome data.

Gene	Sample	Region
PPFIA2	LUSC-60-2711(WGS); UCEC-AP-A0LF(Capture)	14kb after exon29 (WGS); 121bp before exon5 (Capture)
PCNX	LUSC-66-2766(WGS); LUSC-66-2758(Capture)	1kb after exon29 (WGS); 48bp before exon14 (Capture)
CRB1	LUSC-60-2698(WGS); LUSC-22-4593(Capture)	9kb after exon9 (WGS); Exon 7 (Capture)
PTEN	UCEC-BG-A0VV	Exon 6
FAP	UCEC-BG-A0M9	22bp before exon9
CP10	LUSC-43-2578	Exon 11
CABLES1	LUSC-60-2698	87bp after exon2
BCHE	LUSC-60-2708	71bp after exon1
DPF3	LUSC-66-2777	87bp after exon1
PLD1	HNSC-CQ-5332	80bp before exon7
APOL2	HNSC-DQ-5629	3bp after exon1

Table 1-4: Select genes with somatic retrotransposon insertions in exonic regions identified using TranspoSeq-Exome.

Table of a selection of genes that contain somatic retrotransposon insertions. The top three genes each harbor two somatic retrotransposon insertions – one identified from a whole-genome sample and the other identified through exome sequencing of a different sample.

PTEN event

Notably, the tumor suppressor gene, phosphotase and tensin homolog (*PTEN*), is disrupted by a somatic L1HS insertion in a UCEC sample. *PTEN*, which is mutated at a high frequency in cancer, is an important regulator of the PI3K-AKT signaling pathway and double-strand break repair, and is implicated in the pathogenesis of endometrial cancer (Djordjevic et al. 2012). Moreover, loss of function of *PTEN* is the most common genetic aberration in endometrioid carcinomas, seen in up to 80% of cases (Mutter et al. 2000; Dedes et al. 2010). This loss of function is known to be caused by a variety of mechanisms, including point mutation, deletion, promoter methylation, and microRNA regulation. Our findings suggest that retrotransposition may be another mechanism of *PTEN* alteration in endometrial cancer.

3' microhomology

While the L1HS element's 3'-end is inserted at the canonical L1-endonucelase cleavage motif, this retrotransposition is likely the result of a 5' microhomology-mediated end-joining (Zingler 2005), with a 12bp overlap between reference sequence at the 5'-end integration site (just 3' of the TSD) and the 5'-truncated L1HS element. Microhomology is significantly more frequent in non-inverted, 5'-truncated L1 insertions, such as this event (Symer et al. 2002). This may represent a cellular nonhomologous DNA end-joining pathway that resolved the intermediate L1 retrotransposition before its completion, thus eliciting the 5' truncation.

RNAseq evidence

To assess whether the inserted L1 element is expressed, we examined RNAseq data from the uterine carcinoma sample that contains the event. First, we created references sequences for both forward and reverse genome-transposon junctions using the assembled contigs determined in TranspoSeq-Exome. Then we aligned all RNAseq reads to these junction sequences using Bowtie2. Three RNAseq reads (two in the forward and one in the reverse direction) support the expression of the putative L1 insertion at the predicted position (Figure 3-20). As a control, we repeated this procedure on three other samples with no evidence of an L1 insertion in *PTEN* and did not produce any reads aligning to the junctions. Whether the chimeric expression of *PTEN* and this inserted L1HS sequence renders the RNA less stable, and hence leads to the downregulation of functional protein, remains an open question.

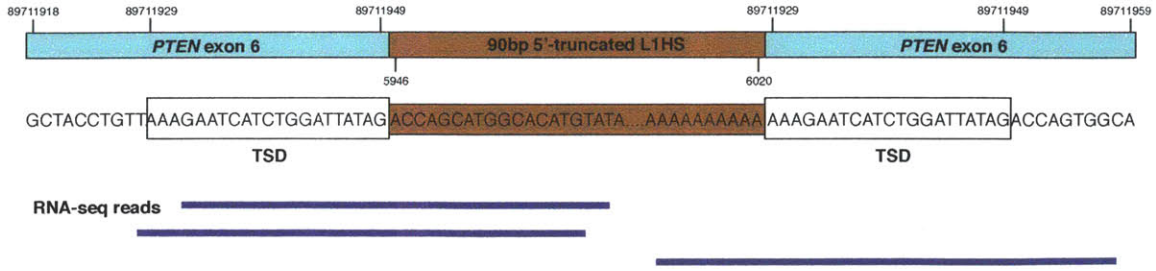


Figure 3-20 Schematic of somatic L1HS insertion into *PTEN* exon.
Somatic L1HS element inserted into exon 6 of the *PTEN* tumor suppressor gene, with canonical TSD sequence surrounding insertion (boxes). Blue lines represent RNA-seq reads that span insertion breakpoints.

3.5 Summary

We analyzed 200 whole-genomes and 767 whole-exomes from tumor and matched normal samples across 11 tumor types using TranspoSeq and TranspoSeq-Exome to reveal the landscape of somatic retrotransposition in human cancer. We find that certain tumor types, such as lung squamous cell carcinoma, head and neck squamous cell carcinoma, and endometrial carcinoma exhibit high rates of somatic retrotransposon insertions, while others such as lung adenocarcinoma, acute myeloid leukemia, and glioblastoma multiforme have little to no retrotransposon activity. The insertion events we identify are largely consistent with known mechanisms of retrotransposition, including the presence of TSD, canonical endonuclease motifs at the point of insertion, and 5' truncation of L1s; however, we do reveal some attributes specific to somatic retrotransposition. Namely, L1 element insertions are highly enriched in tumors relative to Alu insertions, despite their lower rate of germline activity. We find many genes that have somatic retrotransposon insertions in multiple samples across tumor types, implying possible hotspots of retrotransposition. Furthermore, we find several exons disrupted by somatic

retrotransposon insertions in whole-genomes and whole-exomes within genes that have previously been implicated in tumorigenesis.

Chapter 4. Genomic features of somatic retrotransposon insertions in cancer

Given the landscape of somatic retrotransposition across tumor types, it is now possible to comprehensively characterize these events in terms of their genomic and sample-specific context. We look at where retrotransposons tend to integrate, which tumor samples they tend to reactivate in, and we even assess which reference elements are active in several cases. By integrating data from other mutation types, RNAseq, and genomic information, we are able to form a more complete picture of somatic retrotransposition in cancer.

4.1 Genomic rearrangement and mutation versus retrotransposition

Rearrangements

The correlation between genomic rearrangements and retrotransposition, although intuitively strong, has never been formally described. From whole-genome data and our retrotransposon analysis, we are able to characterize the association between these two genomic events for the first time. We binned individuals by number of somatic retrotransposon insertions: Retrotransposon-High (RTI-H) tumors have greater than 10 somatic insertions and Retrotransposon-Low (RTI-L) have 10 or fewer insertions. Samples in the high somatic retrotransposition cluster have more complex genomes in terms of somatic rearrangements (Wilcoxon p=0.0097, Figure 4-1).

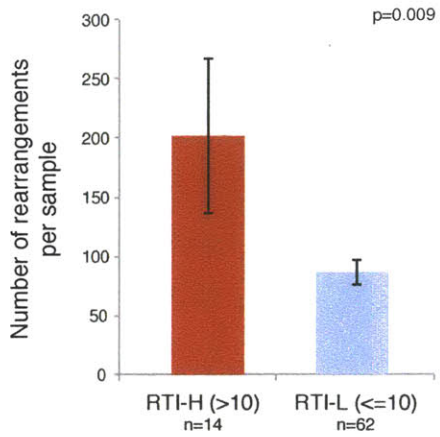


Figure 4-1 Rate of genomic rearrangements versus retrotransposition.
Mean number of somatic genome rearrangements per sample in samples with high rates of retrotransposition (RTI-H) and with low (RTI-L).

Distance from rearrangements

Since samples with high retrotransposon load also have greater numbers of rearrangements, we asked whether the breakpoints for these two events were physically correlated along the DNA sequence. This might imply mutagenic, rearrangement-inducing forces at play preferentially in certain hotspots of the genome. For each LUSC sample, we plotted the genomic sequence distance between each retrotransposon insertion site and all rearrangement breakpoints on that chromosome (Figure 4-2). There does not appear to be a significant difference between rearrangement-retrotransposon distances in RTI-H and RTI-L samples.

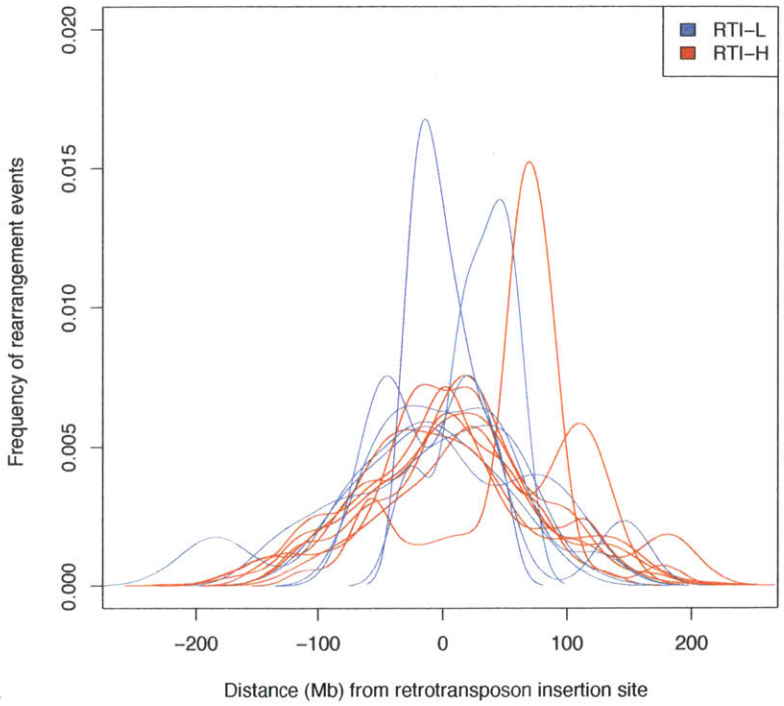


Figure 4-2 Proximity of rearrangement breakpoints to retrotransposon integration sites. Distribution of distances (Mb) between retrotransposon insertion site and genomic rearrangement breakpoints within LUSC samples are plotted; samples with high and low retrotransposon rates are colored red and blue, respectively.

In order to determine whether the general distribution of distances is significant or merely by chance, we created a random set of chromosome and genomic positions for all of the retrotransposon insertions and plotted the distance between these and the real set of genomic rearrangements. Figure 4-3 shows the two distributions plotted on the same coordinate system. We repeated this procedure 1000 times and found that the true set of LUSC retrotransposon insertions points lie, on average, farther from rearrangement breakpoints than would be expected by chance (Wilcoxon and Bonferroni corrected p=4.1E-5). Thus, we do not find evidence for the co-localization of rearrangement insertion points and somatic retrotransposon integration sites.

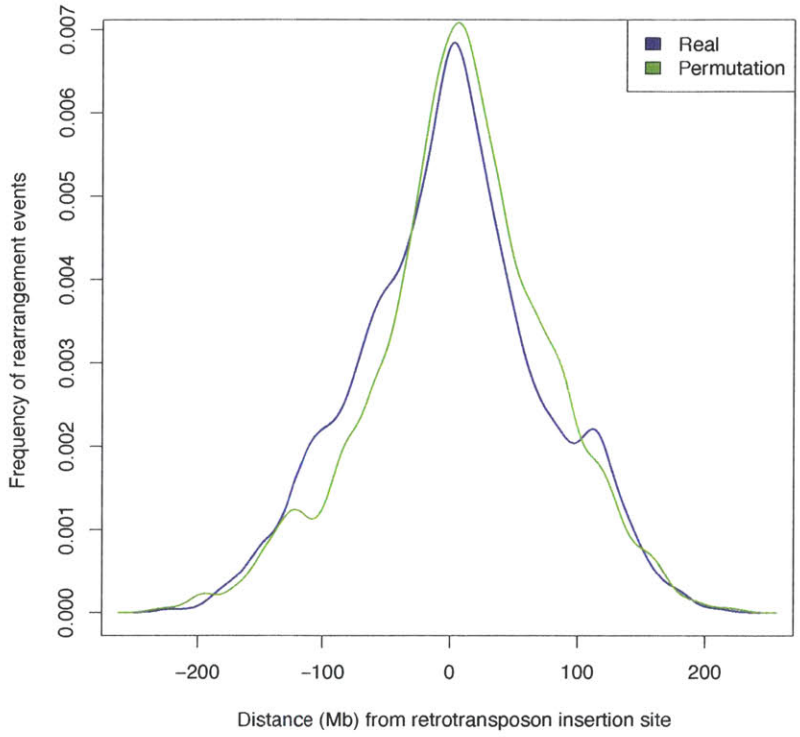


Figure 4-3 Permutation analysis on proximity between rearrangements and retrotransposition sites.
Distance (Mb) between retrotransposon insertions and genomic rearrangement breakpoints across LUSC samples (blue) and when permutated (green).

Mutations

Samples with many retrotransposon insertions have increased rearrangements; does this imply that these samples have higher mutation rates in general? For each sample in the LUSC, HNSC, and UCEC cohorts, we collected somatic nucleotide substitution and short insertion-deletion (indel) data. Retrotransposon-high samples also have greater numbers of total somatic substitution mutations per sample than do retrotransposon-low samples (Wilcoxon p=2.8E-04, Figure 4-4). This may imply that RTI-H samples are more tolerant of mutations in general, of both the rearrangement and single nucleotide variety.

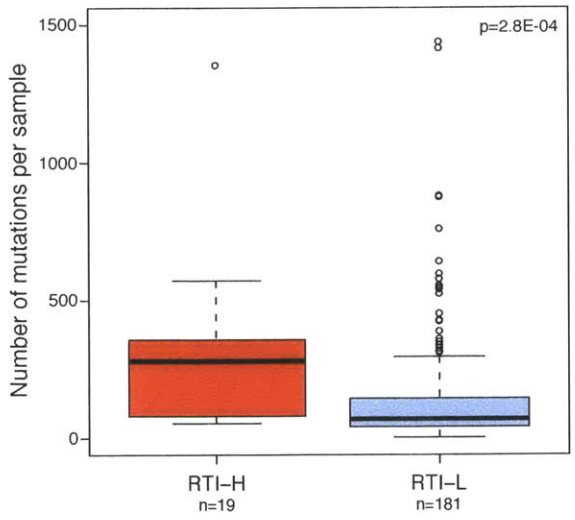


Figure 4-4 Rate of point mutations versus retrotransposition.
Mean number of somatic point mutations per sample in samples with high rates of retrotransposition (RTI-H) and with low (RTI-L).

MSI

It has been suggested previously that samples exhibiting microsatellite instability (MSI) have greater rates of retrotransposition (E. Lee et al. 2012). In our analysis, 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 MSI, as indicated by MSI markers measured through TCGA. MSI status, however, does not predict somatic retrotransposon insertion load, as many MSI-high tumors do not have any somatic insertions. This is in concordance with an extensive study of colorectal tumors where the few samples with the highest number of somatic L1 insertions were MSI positive, but, similarly, MSI status did not correlate in general with the number of somatic retrotransposition events (Solyom et al. 2012).

Clinical variates

Patient age has been associated with somatic retrotransposition rate in 16 colorectal cancers (Solyom et al. 2012). We do not see a statistically significant association between these factors across the 200 patients we analyzed (R^2=0.03). Next, using TCGA and the Broad Institute's GDAC framework, we performed correlations between retrotransposon clusters (RTI-H and RTI-L) versus 9 clinical features for LUSS, HNSC and UCEC: time to death, age, gender, tumor stage, radiation regimen indication, number of pack years smoked, tobacco smoking history indicator, year of tobacco smoking onset, number of lymph nodes. Again, we do not find any statistically significant association that would imply any clinical impact of high somatic retrotransposition rates. We perform these correlations within tumor type, however, and are limited to the tumor types that show a large range of retrotransposon activity. Future studies targeting RTI-H tumor types will provide greater power to search for clinical associations.

4.2 Which came first: L1 endonuclease or double-strand breaks?

Alternate mechanism of somatic insertion

The L1-encoded endonuclease (L1 EN) creates two nicks in DNA during TPRT, resulting in the transient creation of a double-strand break at the site of integration. It has been suggested that L1 EN creates more double-strand breaks than it can repair with L1 insertions (Morrish et al. 2002; Gasior et al. 2006). This may contribute to genomic instability because the repair of double-strand breaks often leads to rearrangements, such as translocations and inversions in cancer (Belgnaoui et al. 2006; Lin et al. 2009). L1 expression can induce formation of gamma-H2AX foci, a marker of double-strand breaks, in an L1 endonuclease-dependent manner, demonstrating

that host DNA repair proteins recognize and process L1-induced lesions in DNA (Gasior et al. 2008). Furthermore, the double-strand break repair proteins, *ERCC1/XPF* and *ATM* is required for L1 integration (Gasior et al. 2006).

Conversely, L1s are also capable of repairing existing double-strand breaks by inserting into the genome via an endonuclease-independent pathway (Morrish et al. 2002). In cell lines that are deficient in nonhomologous end-joining (NHEJ), these integration events occur at near-wildtype levels. L1 EN-independent retrotransposition is distinct from canonical TPRT retrotransposition because it allows for L1 integration at atypical target sequences, exhibits 3' truncation, and an absence of TSDs. See Figure 4-5 for a schematic of both scenarios.

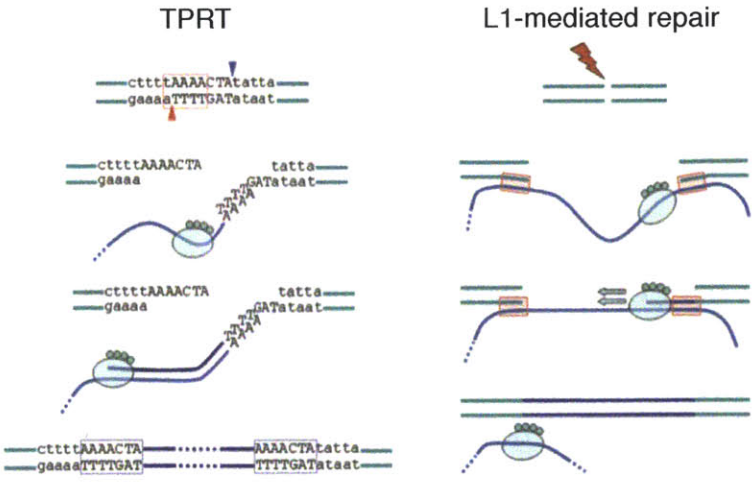


Figure 4-5 Mechanisms of retrotransposon insertion.
Diagram of L1 retrotransposition via target-primed reverse transcription (left) and L1-mediated double-strand break repair (right). In TPRT, L1 endonuclease actively nicks the DNA and creates double-strand breaks that are then resolved through insertion of L1, while in L1-mediated repair, pre-existing double-strand breaks are resolved through L1 insertion. Figure adapted from (Hancks & Kazazian 2012).

We compared the two distributions of somatic retrotransposon insertion events – those with mean TSD length centering around 15bp (>=9bp), typical of traditional TPRT, and those with

mean TSD length around 0bp (<2bp), possibly representative of an alternative insertion mechanism. We sought to determine a sequence motif enrichment separately in the two groups of somatic events. We find that the set of candidate insertions lacking TSDs does not display the canonical L1 endonuclease target sequence, or any enriched sequence motif. These target sites do contain a slight GC bias (Figure 4-6) 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), where they describe a similar peak around a TSD length of 0-2bp, consistent with L1 endonuclease-independent somatic insertion. This implies that a separate insertion mechanism may exists for somatic retrotransposition which produces short, or nonexistent TSDs and does not utilize the canonical L1 endonuclease target sequence.

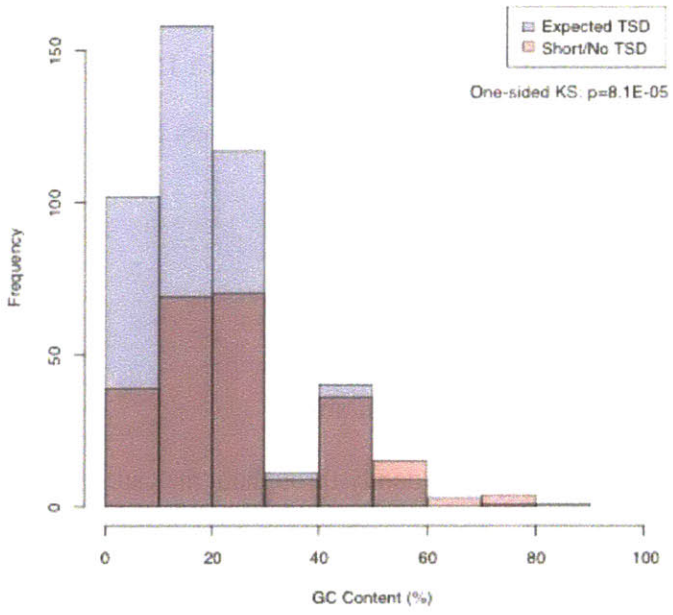


Figure 4-6 GC content of somatic target sites with differing TSD lengths.
GC content (%) at site of somatic insertion of events with TSD length of expected length (>9bp, blue) and those lacking TSDs (<2bp, red).

Insert motif of rearrangements

To assess whether other forms of genomic rearrangement occur via L1-endonuclease induced double-strand breaks, we ran the rearrangement detection tool, dRanger (Bass et al. 2011; Chapman et al. 2011), on the whole-genome sequencing data from LUSC, LUAD and HNSC samples. These genomic rearrangements encompass chromosomal events ranging from tandem duplications and inversions to translocations. Nucleotides surrounding the breakpoints of these events, however, are not enriched for any specific motif, implying that the L1-endonuclease was not the cause of these double-strand breaks.

Thus, it appears that although there is a correlation between rates of retrotransposition and genomic rearrangement, the two mechanisms are distinct – L1 neither repairs existing double-strand breaks, nor creates them for genomic rearrangements to utilize. The alternate mechanism for retrotransposon insertion, represented here as those with a TSD around 0bp and seen as well in Lee et al. (2012) remains to be elucidated, but it is likely not L1-EN independent retrotransposition.

4.3 Features of retrotransposon insertion sites

Whether retrotransposons are inserted randomly throughout the cancer genome or whether there exists selection for or against insertion in certain regions is an open question. Although assays in artificial cell-culture systems do not point to any discriminatory forces at play during retrotransposon integration, it is possible that there exists an integration targeting mechanism that is active in the human organism. Alternatively, negative selection against L1 insertion could

affect integration distribution. To assess any biases toward integration in the context of particular genomic features, we looked at which genes tend to have retrotransposon insertions (and how this affects expression) and which chromatin conformation is most prone to somatic retrotransposition.

Gene expression changes

First, we asked whether somatic retrotransposon insertion into a gene impacts the gene's expression. To assess overall gene expression changes across all tumor types: we compared gene expression in the sample in which the insertion is present to the distribution of RSEM across all other samples investigated. We used a two-tailed Wilcoxon-Mann Whitney test in R to test for the hypothesis that a gene with a retrotransposon insertion is transcribed at a significantly lower level in samples with this insertion. Using available RNAseq data across the eight tumor types with retrotransposon insertions in genes (LUSC, LUAD, HNSC, UCEC, BRCA, OV, COAD, and READ), we find that genes in samples with a retrotransposon insertion tend to be expressed at a lower level than in samples without an insertion (KS-test $p=0.006$, Figure 4-7).

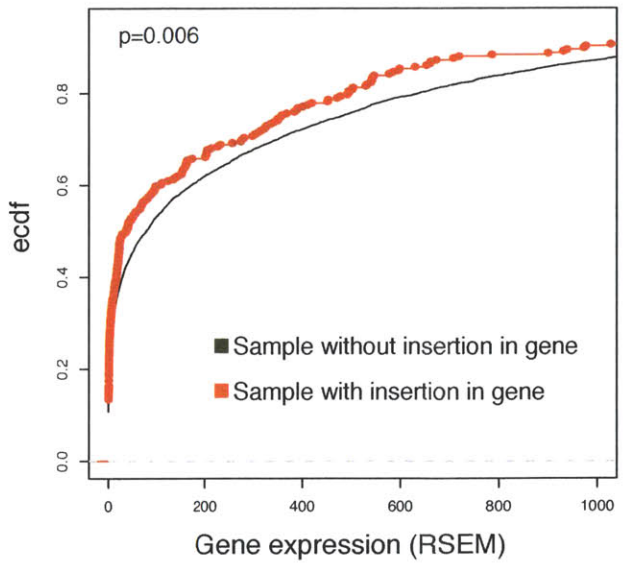


Figure 4-7 Expression of genes with retrotransposon insertions.
Empirical cumulative distribution function (ecdf) of gene expression, quantified by RNAseq by Expectation Maximization (RSEM) values, of genes that contain somatic retrotransposon insertions in a specific sample (red) versus the ecdf of gene expression in genes that do not contain retrotransposon insertions across all other samples (black).

To assess individual gene expression changes independently: for each gene containing a retrotransposon insertion, we compared the RSEM for the sample in which the insertion is present to the empirical cumulative distribution of the RSEM values of that gene across all samples within that tumor type. We used a one-sample, two-sided Kolmogorov-Smirnov test in R (ks.test) to assess the hypothesis that a gene with a retrotransposon insertion is expressed at a significantly different level than in samples without this insertion. P-values were corrected for multiple testing using Bonferroni correction. When examined individually, several genes with retrotransposition insertions show extreme expression relative to all other samples, in either direction (Figure 4-8). Thus, although in general the genes with somatic retrotransposon insertions tend to be expressed at a lower level than those without an insertion, many individual genes are actually expressed at a much higher level than normal when they contain an insertion.

Together with the example of *ST18* upregulation upon retrotransposon insertion (Shukla et al. 2013), this demonstrates that no conclusions can be drawn as to gene expression level based on presence of a somatic retrotransposon insertion.

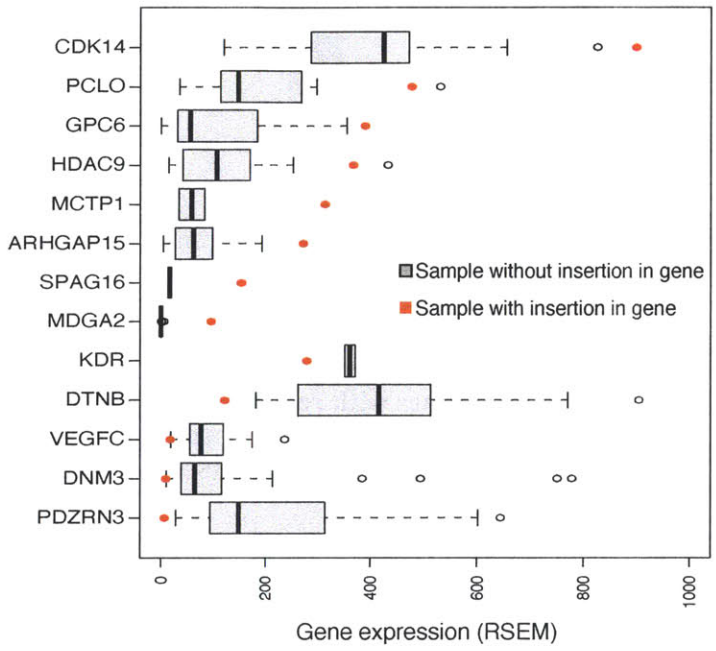


Figure 4-8 Expression of a selection of genes in samples with retrotransposon insertions relative to samples lacking an insertion.
The red dot shows the RSEM value in the particular tumor sample that contained the retrotransposon insertion in that gene, while the grey represents the gene's expression across all other samples within that tumor type that do not contain a retrotransposon insertion.

Large, common fragile site analysis

We looked at which genes tend to have somatic retrotransposon insertions. Longer genes are more likely to harbor mutations merely by chance and have been shown to have higher mutation rates in cancer (Lawrence et al. 2013). We sought to determine whether longer genes have a greater 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 to be tolerated) in longer genes. We find that somatic insertions tend to land in longer genes, as compared to both genes with germline insertions and all genes (Figure 4-9).

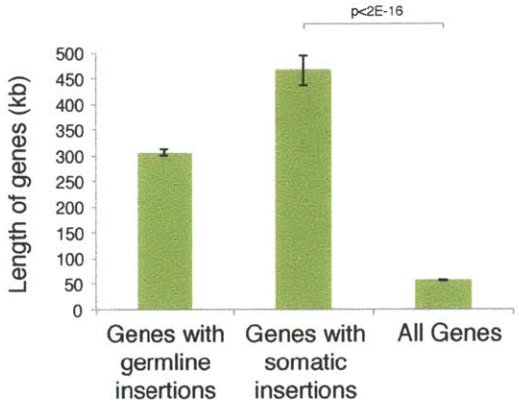


Figure 4-9 Length of genes with retrotransposon insertions.
Mean length (kb) of genes containing germline and somatic retrotransposon insertions, compared to mean length of all RefSeq genes.

Late-replicating

Germline mutation rates are correlated with DNA replication time, with late-replicating regions having much higher mutation rates. This may potentially be due to the depletion of the pool of free nucleotides available toward the end of replication (Stamatoyannopoulos et al. 2009). A recent report found that somatically mutated genes are biased toward later replication time (Lawrence et al. 2013). We find that retrotransposons tend to insert somatically in late-replicating genes, as compared to germline insertions (Wilcoxon p=1.1E-04) and the null distribution of genic replication times (Wilcoxon p<2E-16, Figure 4-10). Replication timing was measured in HeLa cells (C. L. Chen et al. 2010) and highly correlated with blood cell lines (Koren et al. 2012).

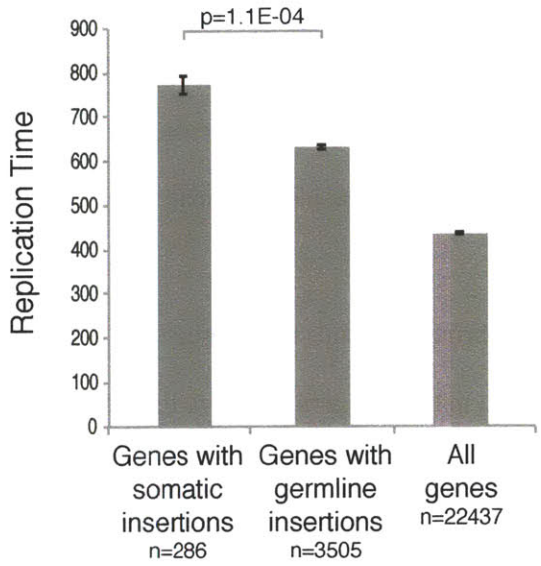


Figure 4-10 Replication timing of genes with retrotransposon insertions.
Replication time of genes with somatic retrotransposon insertions versus those with germline insertions and all RefSeq genes. The greater the replication time value, the later in transcription.

Chromatin conformation

Interestingly, chromatin conformation as assessed by Hi-C long-range interaction data (Lieberman-Aiden et al. 2009), shows that somatic retrotransposon insertions are targeted at regions of the genome that have a more closed conformation (Wilcoxon p=5E-04, Figure 4-11). The distribution of retrotransposon insertions may depend on the accessibility of the chromosome to the transposition machinery. L1-endonuclease, however, shows preference for supercoiled DNA (Feng et al. 1996), and although L1-endonuclease nicking of histone-bound DNA was found to be repressed, some sites were enhanced for L1 nicking when nucleosomal (Cost et al. 2001). We find a disproportionate amount of somatic retrotransposon insertions occurring in closed chromatin regions of the genome. Although we used chromatin open/closed states derived from a normal human lymphoblastoid cell line, Lieberman-Aiden et al. (2009) (Lieberman-Aiden et al. 2009) found high reproducibility between cell lines of different origin

and tissue type. It is tempting to consider that perhaps it is the insertion of retrotransposon elements that *causes* chromatin to change conformation (Chow et al. 2010), but the reader must remember that chromatin conformation states were measured in independent cell lines, not in the primary tumors, so causality here cannot be bidirectional.

Because genes within closed chromatin states are expressed at lower rates, it is conceivable that somatic insertions are tolerated in these regions, despite the difficulty in access. Furthermore, it has previously been suggested that there are lower rates of DNA damage (or enhanced DNA repair) and somatic mutation in open chromatin (Prendergast et al. 2007; Schuster-Böckler & Ben Lehner 2013). Our findings are in agreement with this in that more heterochromatin-like domains are more likely to harbor somatic retrotransposon insertions.

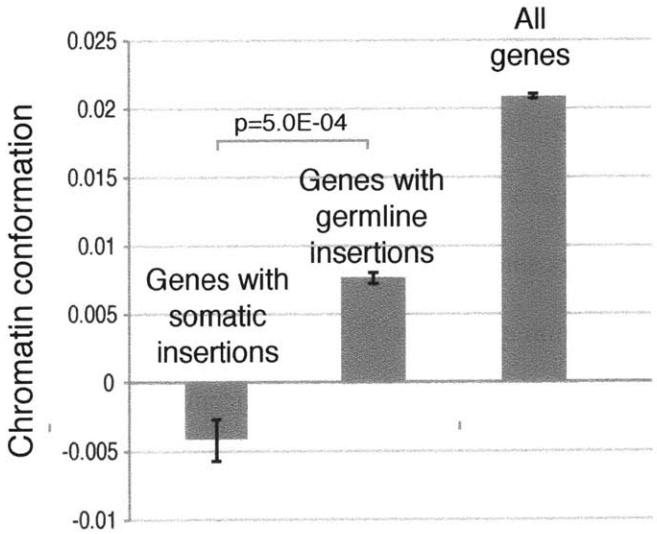


Figure 4-11 Chromatin conformation of genes with retrotransposon insertions.
Mean chromatin conformation of genes with somatic and germline retrotransposon insertions, and all RefSeq genes. Lower values represent more closed conformation.

Common Fragile Site genes

Chromosomal fragile sites are genomic loci that are susceptible to gaps or breaks within metaphase chromosomes. Common fragile sites are observed in all humans and comprise a component of normal chromosome structure (Durkin et al. 2008; Freudenreich 2007; Fungtammasan et al. 2012). These loci are thought to play an important role in chromosomal instability, as they form sites of deletion, amplification and translocation in various cancers (Arlt et al. 2006; Durkin et al. 2008; Burrow et al. 2009). Notably, they are also sites of viral integration (Bester et al. 2006; Dall et al. 2008). Whether these fragile regions are also home to retrotransposon insertion has not to our knowledge been assessed.

We compared somatic retrotransposition sites with the 73 annotated common fragile sites across the genome from Fungtammasan et al. (2012) (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 genes. Similarly, 15% of germline retrotransposon insertions fall in common fragile sites and 18% of genes with germline insertions are common fragile site genes. However, both germline and somatic insertion genes contain more common fragile sites than expected from all RefSeq genes (Fisher's exact $p<2.16E-16$, Figure 4-12). Hence, retrotransposons prefer to insert in common regions of chromosomal breakage in both germline and cancer, much like exogenous viruses. Since L1 encodes its own endonuclease that actively creates a nick in DNA at a specific sequence motif, it would follow that retrotransposons do not require fragile sites of DNA for integration. Why retrotransposon insertions are more prevalent at common fragile sites is then an open question; perhaps these regions are more

accessible to mutagens and forces of chromosomal breakage, whether endonuclease driven or otherwise.

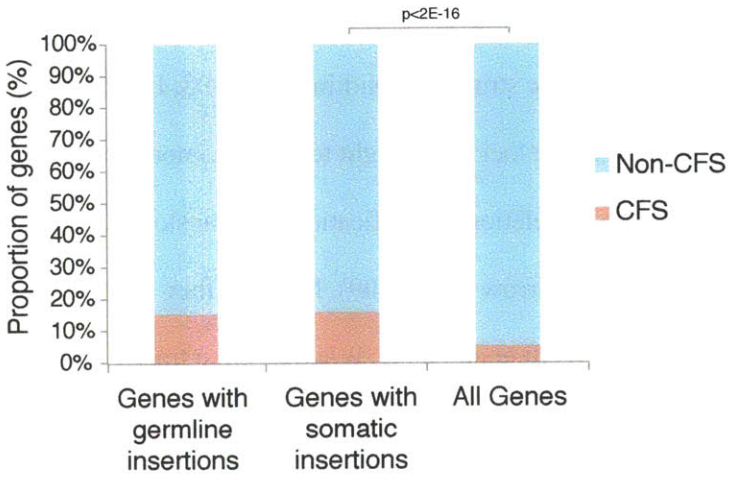


Figure 4-12 Common fragile sites and retrotransposoon integration.
Proportion of genes with germline and somatic retrotransposon insertions, and all RefSeq genes, that overlap with common fragile sites (CFS).

Homozygous deletion analysis

Homozygous deletions of recessive cancer genes and common fragile sites are common occurrences in human cancers (Cox et al. 2005). Somatic retrotransposon insertion could act as the "second hit" for a tumor suppressor gene that requires both alleles to be inactivated in order to have tumorigenic effects. Homozygous deletions have been essential to identifying recessive cancer genes; however, many genes that are homozygously deleted could occur in regions of the genome susceptible to rearrangement, such as common fragile sites, and may not confer any growth advantage to the tumor. However, a genome-wide survey of homozygous deletions in cancer cell lines has shown that most events cannot be explained by fragile sites or copy-number polymorphisms (Cox et al. 2005). These events are often found in regions that may entail fewer adverse consequences for the cell.

We gathered genic absolute copy number data across a panel of 3,450 tumor samples in TCGA Pan-Cancer study for which SNP array data was available and had been analyzed through the ABSOLUTE algorithm to quantify allelic copy numbers (Carter et al. 2012). For each gene, we recorded the fraction of samples in which it was homozygously deleted, leaving no alleles. We find that genes with somatic retrotransposon insertions are enriched for homozygously deleted genes in cancer. Genes with germline retrotransposon insertions are also enriched for these genes, as compared with the general distribution of proportions of samples with homozygous deletions (Figure 4-13).

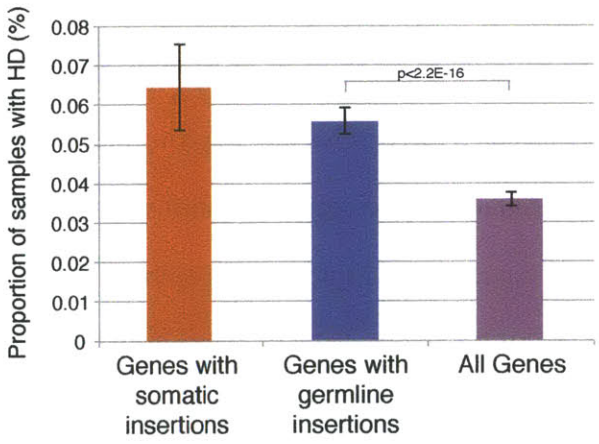


Figure 4-13 Genes with retrotransposon insertions are often homozygously deleted. For each gene, proportion of tumor samples (%) in which that gene is homozygously deleted was plotted across all genes that contain somatic and germline retrotransposon insertions, as well as all RefSeq genes.

GO analysis

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 (Table 4-1) as well as the neuronal synapse cellular component, and cAMP

and calcium ion binding cellular functions (data not shown). Genes that are frequently mutated via other, more typically studied, somatic mutations in gastric cancer are also enriched for cell adhesion pathway (Zang et al. 2012). Many neurodevelopmental disorders are linked to defects in cell adhesion (Greenberg 2003), and recent studies have asserted a high concordance between cancer and neurodevelopmental genes. Many of these adhesion and neurodevelopmental 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). In order to control for gene length, we performed a Gene Ontology analysis of the genes with somatic insertions on a background of the top 75th percentile of long genes (5,609 genes that are longer than 52,023bp). Against this background, genes with somatic retrotransposons insertions are no longer enriched for any biological processes in the gene ontology; however, the cAMP binding and transmembrane signaling receptor activity cellular functions remain, implying the association between these functions and genes with retrotranspositions cannot be attributed to the sheer length of these genes. Finally, other studies have found L1 insertions in cadherin genes such as *CDH11* and *CDH12* genes (E. Lee et al. 2012; Solyom et al. 2012) that mediate calcium-dependent cell-cell adhesion, and imply that the role of cell adhesion genes retrotransposon insertion-associated cancers may deserve further investigation.

The processes associated with genes harboring somatic retrotransposons differ significantly from genes with germline retrotransposon insertions where only the "multicellular organismal process" was slightly enriched; a vague term encapsulating any biological process occurring at the level of a multicellular organism that might be pertinent to its function.

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

Table 4-1 Gene Ontology Biological Processes enriched in genes containing somatic retrotransposon insertions.

4.4 HPV infection in Head & Neck Squamous Cell carcinoma versus retrotransposition

Next, we wondered whether any other genomic feature of the tumor is correlated with retrotransposition rate. Using TCGA and the Broad Institute's GDAC framework, we gathered somatic mutation, methylation, copy number, and miRNA data. Within each tumor type, we correlated these data where available to retrotransposon clusters (RTI-H and RTI-L). We find that in HNSC samples, both *TP53* mutation and p16/*CDKN2A* focal deletion are significantly correlated to high retrotransposition activity (Fisher's p=0.01481, Figure 4-13). Since HPV-positive HNSC tumors are less likely to have *TP53* mutation (Gillison et al. 2000), we looked at

somatic retrotransposition versus HPV status in the 28 HNSC samples and found that, accordingly, samples with high retrotransposition are disproportionately HPV negative (Fisher's exact p=0.041, Figure 4-13).

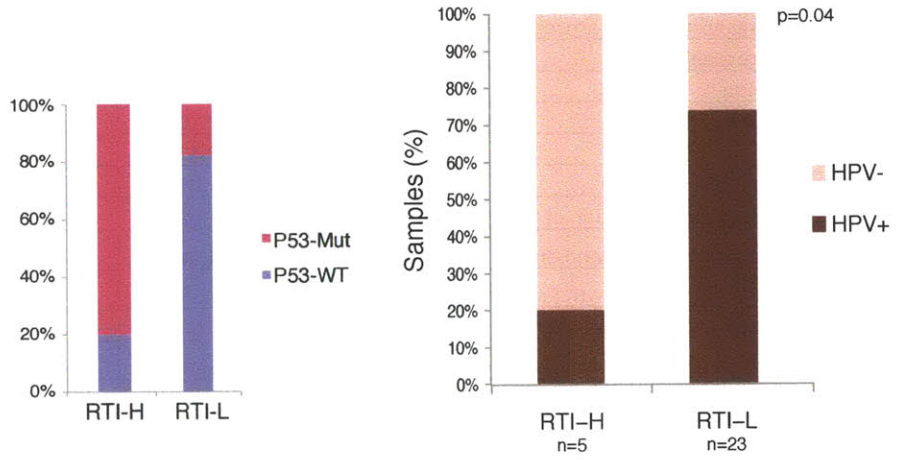


Figure 4-14 HPV status and rate of retrotransposition in HNSC.
Proportion of HNSC samples with *TP53* mutations (left) and HPV infection (right) in retrotransposon-high (RTI-H) and low (RTI-L) clusters.

Thus, we find that in HNSC, retrotransposition occurs more often in the absence of HPV infection. The murine L1Md retroelement is similar in function to L1HS, and has been shown to be downregulated by HPV E7 (Montoya-Durango & Ramos 2012). In human cells, expression of HPV E6 attenuates L1 retrotransposition (Haoudi et al. 2004; Wallace et al. 2013). Additionally, HPV-negative HNSC tumors exhibit heightened genomic instability and global LINE and Alu hypomethylation (Richards et al. 2009; Furniss et al. 2008). It remains to be seen whether HPV proteins in HNSC directly affect L1 regulation or whether samples with active viral infection simply do not have the need for genomic rearrangement and somatic retrotransposition.

4.5 Somatic 3'-sequence transductions elucidate active retrotransposon elements in cancer

In several samples, we find evidence for the retrotransposition of an L1 along with a short unique genomic sequence. These unique sequences originate from the region downstream of both reference and non-reference germline L1 elements. Known as 3'-transduction, this process is thought to result from the read-through of the weak L1 poly(A) signal and is estimated to occur in 15-23% of all genomic L1s (Goodier et al. 2000; Pickeral et al. 2000; Szak et al. 2002; J. V. Moran et al. 1999; Holmes et al. 1994). L1s carrying 3' transductions have been shown to disrupt several human genes in disease including *APC* (Miki et al. 1992), *DMD* (Holmes et al. 1994), *CYBB* (Meischl et al. 2000; Brouha et al. 2002), *RP2* (Schwahn et al. 1998) and *CHM* (van den Hurk et al. 2003). 3'-transductions exhibit known TPRT characteristics, including TSDs, the L1 endonuclease motif at the insertion point, and poly-adenylation of the 3'-transduced segment. Recently, an orphan transduction, or a non-reference L1 so severely truncated that only the uniquely transduced sequence remained, was found to be pathogenic (Solyom et al. 2011).

Short transductions were identified here when reads on one side spanned the transduction and therefore the event maintained evidence for a retrotransposon insertion on both sides. A transduction was called when the 3' end junction of the insertion spanned across the poly(A) sequence into a region 3' of an active (either reference or germline/somatic) L1 element. Element characteristics were assessed using L1Base (Penzkofer 2004).

3'-transductions enable the unique and incredibly important capability of determining which active L1 element constitutes the source of novel somatic insertions. With this information, it is

possible to discern between various scenarios of somatic L1 movement. It is possible that there exists only one element in the genome that is reactivated in cancer, due to its specific genomic or epigenetic context. This element would be important to identify to both elucidate cancer biology and for therapeutic purposes. Alternatively, it is equally possible that any or all of the ~80 known active L1 elements are reactivated in cancer, which would indicate somatic retrotransposition capability from multiple genomic locations and contexts.

One HNSC sample displayed several such 3'-transduction events from different regions of the genome, suggesting that at least three separate L1HS elements were active in the tumor sample. In another sample, we find a known non-reference polymorphic full-length germline L1HS element (chr6:29920436) to be highly active and result in at least four separate instances of somatic 5'-truncated L1HS insertions on chromosomes 3, 9, 11, and X (Table 4-2).

Sample	Source element	Somatic insertions	Length of 3' Transduction
TCGA-BA-4076	Full-length Ta1-nd germline L1HS at chr8:57161596	Chr10	480bp
TCGA-BA-4076	Full-length Ta1-nd reference L1HS at chr22:29059272	Chr2, ChrX, Chr8	604bp
TCGA-BA-4076	Full-length Ta1-d reference L1HS at chr8:135082972	Chr4	528bp
TCGA-BA-5873	Full-length Ta1-nd germline L1HS at chr3:55788568	Chr4	523bp
TCGA-CV-7180	Full-length Ta1-d germline L1HS at chr6:29920436	Chr3, Chr9, Chr11, ChrX	412bp

Table 4-2 Select instances of 3'-transduction events.
Shown are several examples of 3'-transductions found in HNSC samples. The "Source element" column refers to the identity and location of the active retrotransposon, just 5' of the putative 3'-transduction sequence. "Somatic insertions" lists the chromosomes where instances of this 3'-transduction were inserted.

Thus through our analysis, we see evidence for two models of somatic retrotransposon activity in

cancer: i. a single hyperactive source element inserts itself multiple times throughout the genome in the tumor sample, and ii. multiple elements become active in the tumor sample (Figure 4-14).

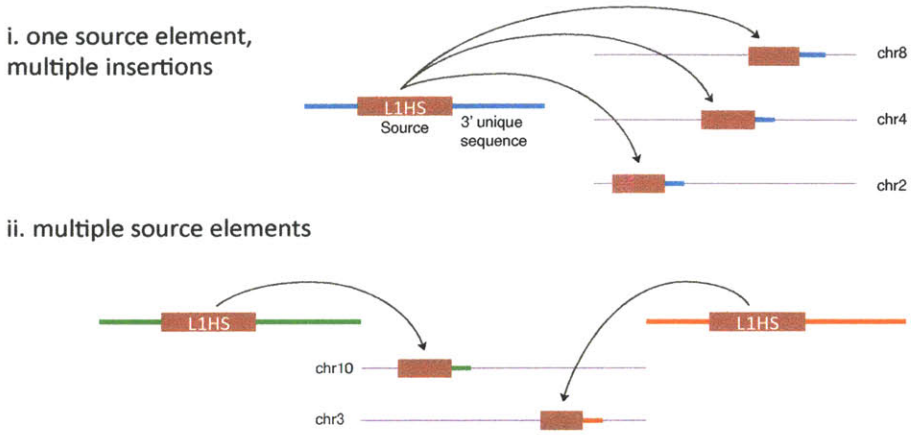


Figure 4-15 Two models of somatic retrotransposition activity in cancer.
Top panel, i., a single hyperactive source element inserts itself multiple times throughout the genome in the tumor sample. Bottom panel, ii., multiple elements become active in the tumor sample.

4.6 Expression of retrotransposable elements

The cell has a variety of defense mechanisms against retrotransposable elements, normally suppressing their expression and subsequent mobilization. It is unclear, however, whether expression of active retrotransposons is upregulated in cancer, and whether this has any impact on retrotransposition. Heightened LINE-1 expression was found in some cases of human breast cancer (Bratthauer et al. 1994), testicular cancer (Bratthauer & Fanning 1992), and pediatric germ cell tumors (Bratthauer & Fanning 1993); however, a sufficiently systematic survey of L1 expression in human cancers cancer has not been performed (Schulz 2006). More recently, several epithelial neoplasms including renal, ovarian, lung and prostate were found to express L1 RNA at detectable levels (Belancio et al. 2010) and L1 RNA levels have even been shown to

correlate with poorer clinical outcomes in pancreatic ductal adenocarcinoma (Ting et al. 2011). Moreover, it is unclear whether increased expression of retrotransposable elements leads to increased retrotransposition activity. If gene expression is, in fact, an indicator for retrotransposition, assessing whether a tumor sample has active somatic retrotransposition would require only expression data, instead of whole-genome sequencing and computationally-intensive algorithms.

To answer this question, we took RNAseq reads from the 19 LUSC samples and aligned raw fasta files to consensus L1HS and AluYa5 sequences using Bowtie2 (Langmead et al. 2009) allowing for 1 mismatch. Values were then converted to Reads Per Kilobase per Million (RPKM) by the formula:

$$number\ of\ mapped\ reads \div length\ of\ transcript\ (kb) \div total\ number\ of\ reads\ (Mb)$$

Comparing L1HS and AluYa5 expression between samples with high and low retrotransposition revealed that expression does not appear to correlate with retrotransposition activity in LUSC (Figure 4-15). Thus, measuring retrotransposon gene expression is not a suitable proxy for assessing retrotransposition activity in a tumor sample. Sequence data for genomic DNA, coupled with algorithms such as TransposSeq, must be utilized to identify novel integration events and establish the extent of retrotransposition in a sample.

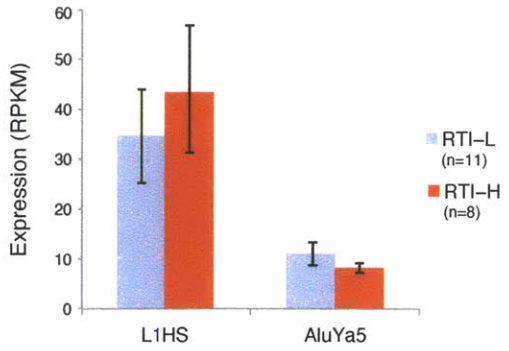


Figure 4-16 Expression of retrotransposon elements.
Gene expression (RPKM) of L1HS and AluYa5 reference elements in LUSC samples with high (red) and low (blue) rates of somatic retrotransposition.

4.7 Summary

We examine the genomic features associated with somatic retrotransposition, i.e., which samples exhibit these events, where retrotransposons land, and where they come from. First, we find that genomic rearrangement load is highly associated with the frequency of retrotransposon insertions. Genomic rearrangements and L1 insertions are both forms of double-strand break repair; however, it appears that these mechanisms are distinct in that genomic rearrangements mend nicks in the DNA created by the L1 endonuclease, and L1 does not mediate repair of double-strand breaks. Samples with high rates of retrotransposition are often microsatellite unstable and have greater frequency of somatic mutation than samples with low retrotransposition. In head and neck squamous cell carcinoma, specifically, samples with high retrotransposition rates are by in large HPV-negative. We show that somatic retrotransposon insertions tend to occur in genes that are long, late-replicating, and in closed chromatin. Importantly, using 3'-transductions, we are able to determine the active source L1 element in several samples and present two models of somatic retrotransposition that occur in cancer.

Chapter 5. Discussion

We present here a large-scale comprehensive analysis of somatic retrotransposon movement in cancer. We find that colorectal, lung squamous cell, head and neck squamous cell and endometrial carcinomas exhibit considerable L1 retrotransposition. Other tumor types, including glioblastoma multiforme, acute myeloid leukemia and kidney clear cell carcinoma, remain quiet. We demonstrate the novel insertion of L1HS into known and putative tumor suppressor genes, such as *RUNX1* and *REV3L*, and identify genes that undergo recurrent insertion across samples and tumor types, such as *CNTNAP2*. We also present the first analysis of retrotransposon insertions using exome-capture data, revealing several important exonic insertions, including one into *PTEN*. Our findings suggest that somatic retrotransposon insertions are an important class of cancer-associated structural variation with the potential to play a role in the tumorigenesis of certain cancers.

A small set of active L1s accounts for most of the L1 activity in humans (Brouha et al. 2002; Beck et al. 2010). We find that the majority of somatically inserted L1s are severely 5'-truncated, and are thus rendered inactive upon insertion. Nonetheless, we do identify several full-length L1HS somatic insertions, as well as common full-length germline polymorphisms that mobilize in the tumor sample, as evidenced by their transduction of unique 3'-sequences. This raises the possibility that polymorphic transposable elements in the germline may predispose to increased somatic retotransposon activity.

The typical mechanism of retrotransposition, TPRT, leads to double-stranded breaks, and so it is thought that L1 transposition has genome-destabilizing effects (Belgnaoui et al. 2006). Whether

it is the L1 that is causing these double-strand breaks or rather contributing to L1-mediated repair of preexisting breaks (Morrish et al. 2002) is an open question; however, it is less likely that L1 elements are being used here as retrotransposon-mediated repair of existing double-strand breaks because somatic L1 insertions exhibit the hallmark L1-endonuclease cleavage site sequence motif, implicating the L1 as the agent of double-stranded nicks.

In addition to identifying novel somatic retrotransposon insertions across multiple tumor and sequencing data types, we also sought to answer the question: what does somatic retrotransposition target? We show here that somatic retrotransposition recurrently targets large, common-fragile site genes that are late-replicating and tend to be located in regions of closed chromatin. Whether these regions are specifically targeted by L1 or whether negative selection eliminated the cells with insertions into other areas remains to be elucidated.

We present the diverse landscape of somatic retrotransposition across human cancers. Tumors of squamous origin appear to be more prone to somatic retrotransposition (lung squamous cell and head and neck squamous cell carcinomas); however, endometrial and colorectal adenocarcinomas also exhibit high rates of retrotransposon insertions. Additional analyses indicate that tumors with general genomic instability, including genomic rearrangements and microsatellite instability, will often also manifest retrotransposon movement. Clinical variates, however, do not appear to correlate with frequency of somatic retrotransposition. In HNSC, specifically, the absence of HPV deems a sample more likely to exhibit highly active retrotransposons. Finally, blood and brain cancers do not harbor a single novel retrotransposon insertion in cancer, whereas most other epithelial tumors have some instances of this genomic

aberration. Additional features distinguishing tumors of varying levels of retrotransposition should be further examined in the future.

5.1 Are retrotransposon insertions driver or passenger events in tumorigenesis?

Passenger vs. Driver

Somatic retrotransposon insertions have the inherent potential to drive tumorigenesis by disrupting tumor suppressor genes or activating oncogenes. Few definitively causal events of this nature, however, have been reported to date. This study represents the largest and most comprehensive investigation of somatic retrotransposition in cancer, which gives us some power to revisit the question of whether retrotransposon insertions act as driver or passenger events in cancer.

The arguments for retrotransposons' role in driving cancer cite that reverse transcriptase inhibition in cell-culture-based cancer models has been shown to promote senescence and differentiation, and reduce invasive growth (Sciamanna et al. 2005; Carlini et al. 2010), however there is debate over whether the effects of the drugs administered are mediated through pathways other than inhibition of the L1 reverse transcriptase specifically. The argument follows that L1 has been shown to be expressed in epithelial neoplasms including renal, ovarian, lung and prostate carcinomas, and this expression is linked to poor prognosis (Belancio et al. 2010). We find here that L1 expression is not a predictor of somatic retrotransposition activity. Finally, there is evidence that L1 can be mutagenic in cell-culture by inducing DNA breaks via its endonuclease, and this could lead to the genomic instability necessary for tumors to

initiate/progress. We find that the L1 endonuclease most likely does not contribute to the double-strand breaks responsible for other somatic genomic rearrangements.

Driver genes in cancer are typically tumor suppressors, which can be mutated in a variety of locations so long as they are rendered dysfunctional, and oncogenes, which are usually mutated at a specific site or domain in order to achieve activation. We find whole genes with recurrent somatic retrotransposon insertions, some of which have been implicated as putative tumor suppressor genes. However, the downstream effects of these insertions are unknown. Lee et al. (2012) found that genes with L1 insertions are commonly mutated in tumors, suggesting that these insertions may contribute to cancer formation; however, we find that genes with L1 insertions display the typical characteristics of passenger events in cancer including long and late-replicating. We do not find any specific loci with recurrent somatic insertions to indicate potential proto-oncogene activation.

The allelic fraction of somatic retrotransposon insertions may give some indication of whether these events play a role in tumor initiation or progression. We find that somatic L1 insertions are typically clonally heterogeneous, present at an allelic fraction centered around 0.5. These analyses are skewed by detection sensitivity, however, especially because TranspoSeq requires a conservative threshold of supporting evidence to call an insertion a true event. Thus, future studies must sequence to a greater depth and sample from multiple regions of the tumor in order to distinguish subclonal populations of specific, presumably later events in tumor evolution.

A recent review cites L1 insertions as frequent genomic passenger events in cancer, while their ability to act as drivers has yet to be demonstrated (Rodić & Burns 2013). While other studies are either too small to assess this question or make grand claims of potential correlation and causality, we find, through the most comprehensive analysis of somatic retrotransposon insertions to date, that indeed most insertions are likely passenger events. Ongoing work to catalog and profile L1 position in additional tumor types – especially the ones found here to have high retrotransposition rates, metastatic samples and disease recurrence after therapy may help elucidate this question further.

5.2 Future studies

Using the framework and analyses presented here as a foundation, we can further pursue questions pertaining to repeat element mobilization in the cancer genome. Studies investigating the correlation between DNA methylation and retrotransposon insertion events, the novel insertion of endogenous retrotroviruses, as well as the rearrangement of repetitive DNA such as telomeres and centromeres, should be performed in order to elucidate this potentially important aspect of tumor biology.

Methylation

DNA methylation is vital for normal cellular function, including gene regulation, cellular differentiation, embryogenesis, X-inactivation and genomic imprinting (Lister et al. 2009). Cytosines within scattered CpG dinucleotides are normally methylated across repetitive elements in the genome, while CpG-dense regions, termed CpG islands, often present in gene promoters

are generally unmethylated. In contrast, the genome of cancer cells is often globally hypomethylated, mostly in scattered CpG dinucleotides in repetitive DNA sequences, which is believed to cause general genomic instability (Feinberg & Vogelstein 1983; Fraga et al. 2005). At the same time, certain CpG islands become hypermethylated in tumors, including those in the promoter regions of known tumor suppressor genes (Greger et al. 1989; Sakai et al. 1991; Gonzalez-Zulueta et al. 1995; Herman et al. 1995; Merlo et al. 1995; Hayslip & Montero 2006). Methylation-induced transcriptional silencing offers a mechanism to inactivate genes, in addition to loss-of-function point mutations and genomic deletions.

DNA methylation acts as the cell's defense mechanism against retrotransposable elements, normally suppressing their expression and subsequent mobilization (Yoder et al. 1997; Bourc'his & Bestor 2004). L1 promoter hypomethylation has been reported in several hematologic malignancies, such as multiple myeloma (Bollati et al. 2009), chronic myeloid leukemia (Roman-Gomez et al. 2005), and chronic lymphocytic leukemia (Fabris et al. 2011). A global hypomethylation signature has been associated with higher levels of somatic retrotransposition in lung cancer (Iskow et al. 2010). In colorectal cancers, however, Solyom et al. (2012) found L1 promoter hypomethylation at 4 CpG sites in tumor samples compared to paired normal tissue, but no correlation was observed between L1 methylation status and the number of L1 insertions (Solyom et al. 2012). Conversely, Lee et al. (2012) found that hypomethylated regions were more likely to harbor somatic retrotransposon insertions.

To assess methylation status across genomes, we will gather Reduced Representation Bisulfite Sequencing (RRBS) data for tumor and matched normal samples. RRBS consists of bisulfite

conversion – the process by which genomic DNA is treated with sodium bisulphite, which converts specifically unmethylated cytosines to uracil leaving methylated cytosines unconverted – on a portion of the genome that is enriched for CpGs. This increases coverage of CpGs and decreases cost. Namely, purified genomic DNA is first digested with a methylation-insensitive restriction enzyme, for example, MspI, which generates fragments with CpG dinucleotides at both ends and is not influenced by their DNA methylation status. Adapters are then added to the ends of the CpG-rich DNA fragments, followed by size selection, bisulfite conversion, PCR amplification and end-sequencing (Gu et al. 2011). Methylation ratios for each CpG dinucleotide capture in the sequencing, i.e., (number of methylated reads)/(total number of reads) covering the CpG will be obtained. These ratios constitute a beta distribution, with values ranging from [0, 1].

The objective in this future study will be to determine the overall methylation level of repeat elements in cancer versus matched normal, as well as specific families of L1 and Alu retrotransposons. Additionally, since we have some indication of potential source L1HS elements through episodic 3'-transduction events (see Section 4.5), we can test whether these elements are active due to demethylation. Finally, we can assess whether somatic retrotransposons tend to land in regions of hypomethylation, as hypothesized by Lee et al. (2012).

We will be limited here by the variability in methylation detection through RRBS as well as the coverage of repeat element CpG dinucleotides. Firstly, bisulfite converts single-stranded but not double-stranded DNA, so incomplete denaturation or reannealing leads to incomplete conversion of unmethylated cytosine to uracil. Thus, it is not always possible to determine whether an

unconverted cytosine is a true methylation or an experimental artifact. Genome-wide coverage is also skewed because RRBS preferentially targets CpG-rich islands, which are often depleted in repeat elements. Finally, because RRBS reads are only 29bp and single-end, there are some inherent alignment ambiguities, especially in the context of repeat elements.

Endogenous retroviruses

Other repeat elements, such as human endogenous retroviruses (HERVs), resemble retroviruses in both their structure and mobility mechanism. Most HERVs contain a dysfunctional *ENV* gene, which prevents them from traveling out of the cell (Bannert & Kurth 2006). HERVs comprise ~8% of the human genome (Lander et al. 2001). Like L1s, the majority of genomic HERVs are incapable of retrotransposition; however, a small number of elements within the HERV-K subfamily are polymorphic in human populations (Belshaw et al. 2005). The HERV-K subfamily differs by the host lysine transfer RNA (tRNA) that initiates HERV-K negative (-) strand cDNA synthesis (Beck et al. 2011). Some HERV-K elements retain intact ORFs (Mayer et al. 1997). Recent studies indicate that HERV-K retrotransposons are expressed in breast and ovarian cancers (Wang-Johanning et al. 2003; Moyes et al. 2007).

Using the TranspoSeq framework, we will search for potential movement of HERV-K elements in tumor genomes. HERV element consensus sequences will replace those of L1, Alu and SVA as an input parameter. Although preliminary data does not indicate evidence for HERV-K mobilization in cancer (data not shown), these may be rare events that will only be discovered through systematic investigation of massive sequencing studies.

Telomere and centromere movement

Repetitive sequences in the human genome apart from transposable elements and endogenous retroviruses include short tandem repeats clustered in the telomeres and centromeres of chromosomes as well as interspersed throughout as microsatellites, and low-copy repetitive sequences present in the regions adjacent to telomeres and centromeres. Due to the inherent difficulty in aligning these elements to the reference genome, they are largely ignored in cancer genomic analyses, despite the fact that they are often involved in rearrangement events due to their high homology. Diseases associated with rearrangement of these elements include idiopathic mental retardation - associated with subtelomeric rearrangements (Knight & Flint 2000), several leukemias – associated with terminal non-reciprocal translocations (Temperani et al. 1995), and colorectal cancer – associated with microsatellite instability (Thibodeau et al. 1993). Using the TranspoSeq framework, it is possible to search for structural rearrangement of other repetitive regions in the cancer cell, including telomeres and centromeres. Existing tools for detecting genomic rearrangement would miss such events because they typically require that both sides of the event contain unique sequence.

Telomeres

The tips of all human chromosomes are composed of the repeated sequence (TTAGGG)n, ranging from 2 to 15 kb in length (Moyzis et al. 1988). These play a crucial role in chromosome stability and linear organization, ensuring complete replication and preventing degradation of chromosomal DNA. These sequences are maintained by telomerase, a ribonucleoprotein that uses an RNA template to direct telomere synthesis (Blackburn 1991). Telomerase has also been shown to form telomeres *de novo* to heal and stabilize broken chromosomes (Flint et al. 1994).

Telomere-repeat-like sequences are found at intrachromosomal sites and may form fragile sites for chromosomal rearrangements(Day et al. 1998). The region directly adjacent to the telomeric repeats is termed the subtelomere, comprised of dynamic and variable repeat sequences that form the transition between unique chromosome-specific sequence and telomeric repeat caps at the end of each chromosome. They are composed of a mosaic of multichromosomal blocks of sequence (Riethman 2003) and have been shown to be hotspots of recombination (Linardopoulou et al. 2005). Many genes are found in subtelomeric regions of human chromosomes and most are members of larger gene families (Mefford & Trask 2002), such as the olfactory receptor (OR) gene family. Individuals have been found to carry up to 56 copies of subtelomeric OR genes, athough there is considerable inter-individual variation in the copy number (Trask et al. 1998).

Rearrangements involving telomeric sequence have been associated with several diseases. A portion of idiopathic mental retardation cases is associated with subtelomeric rearrangements (Knight & Flint 2000), as well as facioscapulohumeral muscular dystrophy (Fisher & Upadhyaya 1997; Clapp et al. 2003). Terminal non-reciprocal translocations, when a telomere rearranges with an intra-arm chromosome segment have been found in certain leukemias (Temperani et al. 1995). Additionally, telomeres can inhibit the expression of nearby genes, called telomere position effect (TPE) (Kulkarni et al. 2010). Rearrangement of these genes away from the telomere may prevent TPE and elicit aberrant gene expression. Conversely, the relocation of a telomere may result in aberrant loss of expression of genes that now find themselves near a telomere.

Centromeres

Centromeres are special structures of eukaryotic chromosomes that hold sister chromatid together and ensure proper chromosome segregation during cell division. Similar to telomeres, centromeres also consist of a region of tandemly repeated DNA sequences, the alpha-satellite repeat, and an adjacent region of a mosaic of large low-copy repeat blocks, called the pericentromeric region. The centromeric repeat region can contain 1-4Mb of 171bp alpha-satellite (alphoid) repeats. There are several alphoid subfamilies seen across human chromosomes (Choo et al. 1991). Pericentromeric regions are prone to genomic instability (Eichler 1998). In fact, it was shown in mice that centromere mitotic recombination occurs in normal cells at a higher frequency than telomere recombination and chromosome arm recombination (Jaco et al. 2008).

TranpoSeq outputs somatic and germline events separately depending on whether the candidate event has reads from both directions supporting it or if it has just one-sided support. Here, we will focus on the one-sided events to identify fusion events with telomeric or centromeric sequence. The reference database inputted in TranspoSeq will be important for locating these events. For telomeres, we will create reference databases that include terminal repeat sequences, (TTAGGG)n, subtelomeric blocks from Linardopoulou et al. (2005) (Linardopoulou et al. 2005), and potentially genes that are located within telomeric regions. Subtelomeric blocks contain retrotransposon elements, so to ensure the identity of the block that is putatively inserted in a region, we will use RepeatMasker to mask all repeat elements in the subtelomeric blocks. Analogously, for the centromeric database, we will use the phylogeny of alpha-satellite repeat

sequences from Choo et al. (1991) (Choo et al. 1991), repeat-masked pericentromeric sequence blocks from Horvath et al. (2000) (Horvath et al. 2000) and pericentromeric genes.

Similar limitations apply when searching for any repeat element mobilization across the genome; however, the limitations for large low-copy repeat elements are not as severe as for interspersed elements because these repeats are generally localized in a few regions in the reference genome. To avoid confounding our results with reference elements and misalignments, we will filter out events that fall in the tails or centromere of a chromosome. We will not be able to identify intra-telomeric or intra-centomeric rearrangements. However, because of the low stringency of alignment to the reference sequence, our method is expected to identify interstitial degenerate telomeric and alpha-satellite repeats.

5.3 Closing remarks

The development of high-throughput technologies, along with new bioinformatics tools, has transformed the study of mobile elements in the human genome (Xing et al. 2013). Studies of the "mobilome" have lagged far behind other "-omics" analyses. This thesis, with several recent studies, marks the start of comprehensively characterizing retrotransposon movement in cancer. Although the incidence of retrotransposons acting as drivers in cancer is yet to be determined, we show that the somatic movement of these elements is a prevalent event in some tumors and is associated with genomic instability and certain biological processes. Different cancer types and samples vary greatly in their receptiveness to retrotransposon insertions. With larger sample sizes and further investigation into those tumor types we show to have high rates of

retrotransposition, it will be possible to subdivide patient populations and uncover novel insights into tumor-specific disease etiologies.

Personalized medicine, or the customization of medical decisions, practices and therapies tailored to the individual patient, is rapidly expanding and revolutionizing healthcare through more efficient and specific treatments, with the end goal of "n of 1" genomics. In the clinic, tumor samples are increasingly being sequenced to identify sample-specific mutations in known cancer genes. As discussed throughout this thesis, it is very possible that a retrotransposon element, reactivated in the tumor cell, lands near or inside one of these genes, even into an exon. Currently, this type of event will be missed because investigators are looking for more typical forms of genomic aberration, such as point mutation and gene fusion. We propose that somatic retrotransposon insertions be examined together with other mutation types, using methods such as TranspoSeq to mine sequencing data.

Thus, somatic retrotransposition should continue to be investigated in both large sequencing studies such as TCGA and alongside clinic decisions in hopes that they provide further insight into tumor biology, clinically viable targets, and potential biomarkers for patient stratification.

References

Alves, G., Tatro, A. & Fanning, T., 1996. Differential methylation of human LINE-1 retrotransposons in malignant cells. *Gene*, 176(1-2), pp.39–44.

Arlt, M.F. et al., 2006. Common fragile sites as targets for chromosome rearrangements. *DNA Repair*, 5(9-10), pp.1126–1135.

Arribas, R. et al., 1999. Prospective assessment of allelic losses at 4p14-16 in colorectal cancer: two mutational patterns and a locus associated with poorer survival. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 5(11), pp.3454–3459.

Awano, H. et al., Contemporary retrotransposition of a novel non-coding gene induces exon-skipping in dystrophin mRNA. 55(12), pp.785–790.

Babushok, D.V. & Kazazian, H.H., 2007. Progress in understanding the biology of the human mutagen LINE-1. *Human mutation*, 28(6), pp.527–539.

Bailey, J.A. et al., 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12), pp.6634–6639.

Bailey, T.L. et al., 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server), pp.W202–W208.

Baillie, J.K. et al., 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, pp.1–4.

Banerji, S. et al., 2012. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403), pp.405–409.

Bannert, N. & Kurth, R., 2006. The Evolutionary Dynamics of Human Endogenous Retroviral Families. *Annual Review of Genomics and Human Genetics*, 7(1), pp.149–173.

Bass, A.J. et al., 2011. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nature genetics*, 43(10), pp.964–968.

Battafarano, R.J. et al., 2005. Large cell neuroendocrine carcinoma: An aggressive form of non-small cell lung cancer. *The Journal of Thoracic and Cardiovascular Surgery*, 130(1), pp.166–172.

Beck, C.R. et al., 2011. LINE-1 Elements in Structural Variation and Disease. *Annual Review of Genomics and Human Genetics*, 12(1), pp.187–215.

Beck, C.R. et al., 2010. LINE-1 Retrotransposition Activity in Human Genomes. *Cell*, 141(7), pp.1159–1170.

Beerenwinkel, N. et al., 2007. Genetic progression and the waiting time to cancer. *PLoS computational biology*, 3(11), p.e225.

Belancio, V.P. et al., 2010. Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Research*, 38(12), pp.3909–3922.

Belancio, V.P., Hedges, D.J. & Deininger, P., 2008. Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Research*, 18(3), pp.343–358.

Belgnaoui, S.M. et al., 2006. Human LINE-1 retrotransposon induces DNA damage and apoptosis in cancer cells. *Cancer cell international*, 6, p.13.

Belshaw, R. et al., 2005. Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Implications for Present-Day Activity. *Journal of Virology*, 79(19), pp.12507–12514.

Beroukhim, R. et al., 2007. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50), pp.20007–20012.

Bester, A.C. et al., 2006. Fragile sites are preferential targets for integrations of MLV vectors in gene therapy. *Gene therapy*, 13(13), pp.1057–1059.

Blackburn, E.H., 1991. Structure and function of telomeres. *Nature*, 350(6319), pp.569–573.

Bochukova, E.G. et al., 2009. Rare mutations of FGFR2causing apert syndrome: identification of the first partial gene deletion, and an Aluelement insertion from a new subfamily. *Human mutation*, 30(2), pp.204–211.

Boeke, J.D., 1997. LINEs and Alus--the polyA connection. *Nature genetics*, 16(1), pp.6–7.

Boissinot, S. & Furano, A.V., 2001. Adaptive evolution in LINE-1 retrotransposons. *Molecular biology and evolution*, 18(12), pp.2186–2194.

Boissinot, S. et al., 2006. Fitness cost of LINE-1 (L1) activity in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 103(25), pp.9590–9594.

Boissinot, S., Chevret, P. & Furano, A.V., 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Molecular biology and evolution*, 17(6), pp.915–928.

Boissinot, S., Entezam, A. & Furano, A.V., 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Molecular biology and evolution*, 18(6), pp.926–935.

Bollati, V. et al., 2009. Differential repetitive DNA methylation in multiple myeloma molecular subgroups. *Carcinogenesis*, 30(8), pp.1330–1335.

Bourc'his, D. & Bestor, T.H., 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, 431(7004), pp.96–99.

Bratthauer, G.L. & Fanning, T.G., 1992. Active LINE-1 retrotransposons in human testicular cancer. *Oncogene*, 7(3), pp.507–510.

Bratthauer, G.L. & Fanning, T.G., 1993. LINE-1 retrotransposon expression in pediatric germ cell tumors. *Cancer*, 71(7), pp.2383–2386.

Bratthauer, G.L., Cardiff, R.D. & Fanning, T.G., 1994. Expression of LINE-1 retrotransposons in human breast cancer. *Cancer*, 73(9), pp.2333–2336.

Brondello, J.-M. et al., 2008. Novel evidences for a tumor suppressor role of Rev3, the catalytic subunit of Pol ζ. *Oncogene*, 27(47), pp.6093–6101.

Brouha, B. et al., 2002. Evidence consistent with human L1 retrotransposition in maternal meiosis I. *American journal of human genetics*, 71(2), pp.327–336.

Brouha, B. et al., 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9), pp.5280–5285.

Bruder, C.E.G. et al., 2008. Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. *The American Journal of Human Genetics*, 82(3), pp.763–771.

Burns, K.H. & Boeke, J.D., 2012. Human Transposon Tectonics. *Cell*, 149(4), pp.740–752.

Burrow, A.A. et al., 2009. Over half of breakpoints in gene pairs involved in cancer-specific recurrent translocations are mapped to human chromosomal fragile sites. *BMC Genomics*, 10(1), p.59.

Burwinkel, B. & Kilimann, M.W., 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *Journal of molecular biology*, 277(3), pp.513–517.

C, M. et al., 1991. Reverse transcriptase encoded by a human transposable element. *Science (New York, N.Y.)*, 254(5039), pp.1808–1810.

Carlini, F. et al., 2010. The Reverse Transcription Inhibitor Abacavir Shows Anticancer Activity in Prostate Cancer Cell Lines C. Creighton, ed. *PLoS ONE*, 5(12), p.e14221.

Carroll, M.L. et al., 2001. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *Journal of molecular biology*, 311(1), pp.17–40.

Carter, S.L. et al., 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology*, 30(5), pp.413–421.

Chai, H. & Brown, R.E., 2009. Field effect in cancer-an update. *Annals of clinical and laboratory science*, 39(4), pp.331–337.

Chapman, M.A. et al., 2011. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339), pp.467–472.

Chen, C.L. et al., 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Research*, 20(4), pp.447–457.

Chen, J.-M. et al., 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Human Genetics*, 117(5), pp.411–427.

Cho, N.-Y. et al., 2007. Hypermethylation of CpG island loci and hypomethylation of LINE-1 and Alu repeats in prostate adenocarcinoma and their relationship to clinicopathological features. *The Journal of Pathology*, 211(3), pp.269–277.

Choi, I.-S. et al., 2007. Hypomethylation of LINE-1 and Alu in well-differentiated neuroendocrine tumors (pancreatic endocrine tumors and carcinoid tumors). *Modern Pathology*, 20(7), pp.802–810.

Choo, K.H. et al., 1991. A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Research*, 19(6), pp.1179–1182.

Chow, J.C. et al., 2010. LINE-1 Activity in Facultative Heterochromatin Formation during X Chromosome Inactivation. *Cell*, 141(6), pp.956–969.

Clapp, J., Bolland, D.J. & Hewitt, J.E., 2003. Genomic analysis of facioscapulohumeral muscular dystrophy. *Briefings in functional genomics & proteomics*, 2(3), pp.213–223.

Cordaux, R. & Batzer, M.A., 2009. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10), pp.691–703.

Cost, G.J. et al., 2002. Human L1 element target-primed reverse transcription in vitro. *The EMBO journal*, 21(21), pp.5899–5910.

Cost, G.J. et al., 2001. Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucleic Acids Research*, 29(2), pp.573–577.

Coufal, N.G. et al., 2009. L1 retrotransposition in human neural progenitor cells. *Nature*, 460(7259), pp.1127–1131.

Cowin, P.A. et al., 2012. LRP1B Deletion in High-Grade Serous Ovarian Cancers Is Associated with Acquired Chemotherapy Resistance to Liposomal Doxorubicin. *Cancer research*, 72(16), pp.4060–4073.

Cox, C. et al., 2005. A survey of homozygous deletions in human cancer genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12), pp.4542–4547.

D, F. et al., 2010. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research*, 39(Database), pp.D945–D950.

Dall, K.L. et al., 2008. Characterization of Naturally Occurring HPV16 Integration Sites Isolated from Cervical Keratinocytes under Noncompetitive Conditions. *Cancer research*, 68(20), pp.8249–8259.

Day, J.P., Limoli, C.L. & Morgan, W.F., 1998. Recombination involving interstitial telomere repeat-like sequences promotes chromosomal instability in Chinese hamster cells. *Carcinogenesis*, 19(2), pp.259–265.

Dedes, K.J. et al., 2010. PTEN Deficiency in Endometrioid Endometrial Adenocarcinomas Predicts Sensitivity to PARP Inhibitors. *Science Translational Medicine*, 2(53), pp.53ra75–53ra75.

Deininger, P.L. & Batzer, M.A., 1999. Alu repeats and human disease. *Molecular genetics and metabolism*, 67(3), pp.183–193.

Deininger, P.L. et al., 1992. Master genes in mammalian repetitive DNA amplification. *Trends in genetics : TIG*, 8(9), pp.307–311.

Djordjevic, B. et al., 2012. Clinical assessment of PTEN loss in endometrial carcinoma: immunohistochemistry outperforms gene sequencing. 25(5), pp.699–708.

Dulak, A.M. et al., 2012. Gastrointestinal Adenocarcinomas of the Esophagus, Stomach, and Colon Exhibit Distinct Patterns of Genome Instability and Oncogenesis. *Cancer research*, 72(17), pp.4383–4393.

Durkin, S.G. et al., 2008. Replication stress induces tumor-like microdeletions in FHIT/FRA3B. *Proceedings of the National Academy of Sciences of the United States of America*, 105(1), pp.246–251.

Dutt, A. et al., 2008. Drug-sensitive FGFR2 mutations in endometrial carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, 105(25), pp.8713–8717.

Eichler, E.E., 1998. Masquerading repeats: paralogous pitfalls of the human genome. *Genome Research*, 8(8), pp.758–762.

Esnault, C., Maestre, J. & HEIDMANN, T., 2000. Human LINE retrotransposons generate processed pseudogenes. *Nature genetics*, 24(4), pp.363–367.

Evrony, G.D. et al., 2012. Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. *Cell*, 151(3), pp.483–496.

Ewing, A.D. & Kazazian, H.H., 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Research*, 20(9), pp.1262–1270.

Ewing, A.D. & Kazazian, H.H., 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Research*, 21(6), pp.985–990.

Fabris, S. et al., 2011. Biological and clinical relevance of quantitative global methylation of repetitive DNA sequences in chronic lymphocytic leukemia. *Epigenetics*, 6(2), pp.188–194.

Faulkner, G.J., 2011. Retrotransposons: Mobile and mutagenic from conception to death. *FEBS Letters*, 585(11), pp.1589–1594.

Feinberg, A.P. & Vogelstein, B., 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301(5895), pp.89–92.

Feng, Q. et al., 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, 87(5), pp.905–916.

Ferlay J, S.H.B.F.F.D.M.C., GLOBOCAN 2008 v2.0. *GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 [Internet]*. Available at: http://globocan.iarc.fr/ [Accessed September 4, 2013].

Fisher, J. & Upadhyaya, M., 1997. Molecular genetics of facioscapulohumeral muscular dystrophy (FSHD). *Neuromuscular disorders : NMD*, 7(1), pp.55–62.

Flint, J.J. et al., 1994. Healing of broken human chromosomes by the addition of telomeric repeats. *American journal of human genetics*, 55(3), pp.505–512.

Fraga, M.F. et al., 2005. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30), pp.10604–10609.

Freeman, J.D., Goodchild, N.L. & Mager, D.L., 1994. A modified indicator gene for selection of retrotransposition events in mammalian cells. *BioTechniques*, 17(1), pp.46–48–9– 52.

Freudenreich, C.H., 2007. Chromosome fragility: molecular mechanisms and cellular consequences. *Frontiers in bioscience : a journal and virtual library*, 12, pp.4911–4924.

Fujita, P.A. et al., 2010. The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*, 39(Database), pp.D876–D882.

Fungtammasan, A. et al., 2012. A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome? *Genome Research*, 22(6), pp.993–1005.

Furniss, C.S. et al., 2008. Line Region Hypomethylation Is Associated with Lifestyle and Differs by Human Papillomavirus Status in Head and Neck Squamous Cell Carcinomas. *Cancer Epidemiology Biomarkers & Prevention*, 17(4), pp.966–971.

Gasior, S.L. et al., 2006. The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. *Journal of molecular biology*, 357(5), pp.1383–1393.

Gasior, S.L., Roy-Engel, A.M. & Deininger, P.L., 2008. ERCC1/XPF limits L1 retrotransposition. *DNA Repair*, 7(6), pp.983–989.

Gerlinger, M. et al., 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England journal of medicine*, 366(10), pp.883–892.

Gibbs, R.A. et al., 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science (New York, N.Y.)*, 316(5822), pp.222–234.

Gilbert, N. et al., 2005. Multiple Fates of L1 Retrotransposition Intermediates in Cultured Human Cells. *Molecular and cellular biology*, 25(17), pp.7780–7795.

Gilbert, N., Lutz-Prigge, S. & Moran, J.V., 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell*, 110(3), pp.315–325.

Gillison, M.L. et al., 2000. Evidence for a causal association between human papillomavirus and a subset of head and neck cancers. *Journal of the National Cancer Institute*, 92(9), pp.709–720.

Gonzalez-Zulueta, M. et al., 1995. Methylation of the 5' CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing. *Cancer research*, 55(20), pp.4531–4535.

Goodier, J.L., Ostertag, E.M. & Kazazian, H.H., 2000. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Human Molecular Genetics*, 9(4), pp.653–657.

Grabherr, M.G. et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Publishing Group*, 29(7), pp.644–652.

Greenberg, D.A., 2003. Linking acquired neurodevelopmental disorders to defects in cell adhesion. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14), pp.8043–8044.

Greenman, C. et al., 2007. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132), pp.153–158.

Greger, V. et al., 1989. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Human Genetics*, 83(2), pp.155–158.

Gu, H. et al., 2011. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols*, 6(4), pp.468–481.

Han, K., 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Research*, 33(13), pp.4040–4052.

Hancks, D.C. & Kazazian, H.H., Jr, 2012. Active human retrotransposons: variation and disease. *Current Opinion in Genetics & Development*, 22(3), pp.191–203.

Haoudi, A. et al., 2004. Retrotransposition-Competent Human LINE-1 Induces Apoptosis in Cancer Cells With Intact p53. *Journal of Biomedicine and Biotechnology*, 2004(4), pp.185–194.

Hayslip, J. & Montero, A., 2006. Tumor suppressor gene methylation in follicular lymphoma: a comprehensive review. *Molecular cancer*, 5, p.44.

Helman, E. & Meyerson, M., 2012. Genomic Impact Of Eukaryotic Transposable Elements. pp.1–145.

Helman, E. & Meyerson, M., RetroSeq: A Tool To Discover Somatic Insertion of Retrotransposons - Elena Helman - TCGA. *cancergenome.nih.gov*. Available at: http://cancergenome.nih.gov/newsevents/multimedialibrary/videos/retroseqhelman [Accessed June 4, 2013].

Helman, E. & Meyerson, M., Translation of the Cancer Genome Program. *aacr.org*. Available at: http://www.aacr.org/home/scientists/meetings--workshops/special-conferences/previous-special-conferences/2011---2012-special-conferences/translation-of-the-cancer-genome/program.aspx [Accessed June 7, 2013].

Herman, J.G. et al., 1995. Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer research*, 55(20), pp.4525–4530.

Holmes, S.E. et al., 1994. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nature genetics*, 7(2), pp.143–148.

Hormozdiari, F. et al., 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics (Oxford, England)*, 26(12), pp.i350–i357.

Horvath, J.E., Schwartz, S. & Eichler, E.E., 2000. The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. *Genome Research*, 10(6), pp.839–852.

Huang, C.R.L. et al., 2010. Mobile Interspersed Repeats Are Major Structural Variants in the Human Genome. *Cell*, 141(7), pp.1171–1182.

Iskow, R.C. et al., 2010. Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons. *Cell*, 141(7), pp.1253–1261.

Iwakawa, R. et al., 2012. Contribution of germline mutations to PARK2 gene inactivation in lung adenocarcinoma. *Genes, Chromosomes and Cancer*, 51(5), pp.462–472.

Jaco, I. et al., 2008. Centromere mitotic recombination in mammalian cells. *The Journal of Cell Biology*, 181(6), pp.885–892.

Kashiwagi, E. et al., 2012. Downregulation of phosphodiesterase 4B (PDE4B) activates protein kinase A and contributes to the progression of prostate cancer. *The Prostate*, 72(7), pp.741–

751.

Kazazian, H.H., 1999. An estimated frequency of endogenous insertional mutations in humans. *Nature genetics*, 22(2), p.130.

Kazazian, H.H., 2004. Mobile Elements: Drivers of Genome Evolution. *Science (New York, N.Y.)*, 303(5664), pp.1626–1632.

Kazazian, H.H. et al., 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, 332(6160), pp.164–166.

Keane, T.M., Wong, K. & Adams, D.J., 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics (Oxford, England)*, 29(3), pp.389–390.

Khan, H., Smit, A. & Boissinot, S., 2005. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Research*, 16(1), pp.78–87.

Knight, S.J. & Flint, J., 2000. Perfect endings: a review of subtelomeric probes and their use in clinical diagnosis. *Journal of medical genetics*, 37(6), pp.401–409.

Koboldt, D.C. et al., 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), pp.61–70.

Koren, A. et al., 2012. AR TICLEDifferential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. *American journal of human genetics*, 91(6), pp.1033–1040.

Kulkarni, A. et al., 2010. Effect of Telomere Proximity on Telomere Position Effect, Chromosome Healing, and Sensitivity to DNA Double-Strand Breaks in a Human Tumor Cell Line. *Molecular and cellular biology*, 30(3), pp.578–589.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Langbein, S. et al., 2002. Alteration of the LRP1B gene region is associated with high grade of urothelial cancer. *Laboratory investigation; a journal of technical methods and pathology*, 82(5), pp.639–643.

Langmead, B. et al., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), p.R25.

Lawrence, M.S. et al., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, pp.1–5.

Lebeau, A. et al., 2008. Oestrogen receptor gene (ESR1) amplification is frequent in endometrial carcinoma and its precursor lesions. *The Journal of Pathology*, 216(2), pp.151–157.

Lee, E. et al., 2012. Landscape of Somatic Retrotransposition in Human Cancers. *Science (New York, N.Y.)*.

Lee, J. et al., 2007. Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene*, 390(1-2), pp.18–27.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–2079.

Liao, R.G. et al., 2013. Inhibitor-Sensitive FGFR2 and FGFR3 Mutations in Lung Squamous Cell Carcinoma. *Cancer research*, 73(16), pp.5195–5205.

Lieberman-Aiden, E. et al., 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (New York, N.Y.)*, 326(5950), pp.289–293.

Lin, C. et al., 2009. Nuclear Receptor-Induced Chromosomal Proximity and DNA Breaks Underlie Specific Translocations in Cancer. *Cell*, 139(6), pp.1069–1083.

Linardopoulou, E.V. et al., 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, 437(7055), pp.94–100.

Lister, R. et al., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), pp.315–322.

Littink, K.W. et al., 2010. Mutations in the EYS gene account for approximately 5% of autosomal recessive retinitis pigmentosa and cause a fairly homogeneous phenotype. *Ophthalmology*, 117(10), pp.2026–33– 2033.e1–7.

Liu, C.X. et al., 2000. LRP-DIT, a putative endocytic receptor gene, is frequently inactivated in non-small cell lung cancer cell lines. *Cancer research*, 60(7), pp.1961–1967.

Liu, W.-B. et al., 2012. ANKRD18A as a novel epigenetic regulation gene in lung cancer. *Biochemical and Biophysical Research Communications*, 429(3-4), pp.180–185.

Loman, N.J. et al., 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5), pp.434–439.

Lotery, A.J. et al., 2001. Mutations in the CRB1 gene cause Leber congenital amaurosis. *Archives of ophthalmology*, 119(3), pp.415–420.

Luan, D.D. & Eickbush, T.H., 1995. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Molecular and cellular biology*, 15(7), pp.3882–3891.

Luan, D.D. et al., 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, 72(4), pp.595–605.

Mardis, E.R., 2012. PERSPECTIVE. *Nature*, 470(7333), pp.198–203.

Martin, S.L. & Bushman, F.D., 2001. Nucleic Acid Chaperone Activity of the ORF1 Protein from the Mouse LINE-1 Retrotransposon. *Molecular and cellular biology*, 21(2), pp.467–475.

Mayer, J., Meese, E. & Mueller-Lantzsch, N., 1997. Multiple human endogenous retrovirus (HERV-K) loci with gag open reading frames in the human genome. *Cytogenetics and cell genetics*, 78(1), pp.1–5.

Meehan, M. et al., 2007. Alpha-T-catenin (CTNNA3) displays tumour specific monoallelic expression in urothelial carcinoma of the bladder. *Genes, Chromosomes and Cancer*, 46(6), pp.587–593.

Mefford, H.C. & Trask, B.J., 2002. THE COMPLEX STRUCTURE AND DYNAMIC EVOLUTION OF HUMAN SUBTELOMERES. *Nature Reviews Genetics*, 3(2), pp.91–102.

Meischl, C. et al., 2000. A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. *European journal of human genetics : EJHG*, 8(9), pp.697–703.

Merlo, A. et al., 1995. 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nature medicine*, 1(7), pp.686–692.

Meyer, L.R. et al., 2012. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Research*, 41(D1), pp.D64–D69.

Meyerson, M., Gabriel, S. & Getz, G., 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, 11(10), pp.685–696.

Miki, Y. et al., 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer research*, 52(3), pp.643–645.

Miller, D.G., 1980. On the nature of susceptibility to cancer. The presidential address. *Cancer*, 46(6), pp.1307–1318.

Mills, R.E. et al., 2006. Recently mobilized transposons in the human and chimpanzee genomes. *American journal of human genetics*, 78(4), pp.671–679.

Montoya-Durango, D.E. & Ramos, K.S., 2012. HPV E7 viral oncoprotein disrupts transcriptional regulation of L1Md retrotransposon. *FEBS Letters*, 586(1), pp.102–106.

Moran, C.A. et al., 2009. Neuroendocrine Carcinomas of the Lung: A Critical Analysis. *American Journal of Clinical Pathology*, 131(2), pp.206–221.

Moran, J.V. et al., 1996. High frequency retrotransposition in cultured mammalian cells. *Cell*, 87(5), pp.917–927.

Moran, J.V., DeBerardinis, R.J. & Kazazian, H.H., 1999. Exon shuffling by L1 retrotransposition. *Science (New York, N.Y.)*, 283(5407), pp.1530–1534.

Morrish, T.A. et al., 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nature genetics*, 31(2), pp.159–165.

Moyes, D., Griffiths, D.J. & Venables, P.J., 2007. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. *Trends in Genetics*, 23(7), pp.326–333.

Moyzis, R.K. et al., 1988. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 85(18), pp.6622–6626.

Muotri, A.R. et al., 2010. L1 retrotransposition in neurons is modulated by MeCP2. *Nature*, 468(7322), pp.443–446.

Muotri, A.R. et al., 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature*, 435(7044), pp.903–910.

Musova, Z. et al., 2006. A novel insertion of a rearranged L1 element in exon 44 of the dystrophin gene: Further evidence for possible bias in retroposon integration. *Biochemical and Biophysical Research Communications*, 347(1), pp.145–149.

Mutter, G.L. et al., 2000. Altered PTEN expression as a diagnostic marker for the earliest endometrial precancers. *Journal of the National Cancer Institute*, 92(11), pp.924–930.

Mülhardt, C. et al., 1994. The spastic mouse: aberrant splicing of glycine receptor beta subunit mRNA caused by intronic insertion of L1 element. *Neuron*, 13(4), pp.1003–1015.

Narita, N. et al., 1993. Insertion of a 5" truncated L1 element into the 3" end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *The Journal of clinical investigation*, 91(5), pp.1862–1867.

NHGRI, N., 2009. Executive SummaryToward a Comprehensive Genomic Analysis of Cancer. pp.1–8.

Nowell, P.C. & Hungerford, D.A., 1961. Chromosome studies in human leukemia. II. Chronic granulocytic leukemia. *Journal of the National Cancer Institute*, 27, pp.1013–1035.

Oldridge, M. et al., 1999. De novo alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. *American journal of human genetics*, 64(2), pp.446–461.

Ostertag, E.M. & Kazazian, H.H., 2001. Biology of mammalian L1 retrotransposons. *Annual review of genetics*, 35, pp.501–538.

Ostertag, E.M. et al., 2002. A mouse model of human L1 retrotransposition. *Nature genetics*, 32(4), pp.655–660.

Ovchinnikov, I., 2001. Genomic Characterization of Recent Human LINE-1 Insertions: Evidence Supporting Random Insertion. *Genome Research*, 11(12), pp.2050–2058.

Penzkofer, T., 2004. L1Base: from functional annotation to prediction of active LINE-1 elements. *Nucleic Acids Research*, 33(Database issue), pp.D498–D500.

Pickeral, O.K. et al., 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Research*, 10(4), pp.411–415.

Plun-Favreau, H. et al., 2010. Cancer and Neurodegeneration: Between the Devil and the Deep Blue Sea M. S. Horwitz, ed. *PLoS Genetics*, 6(12), p.e1001257.

Polascik, T.J. et al., 1995. Distinct regions of allelic loss on chromosome 4 in human primary bladder carcinoma. *Cancer research*, 55(22), pp.5396–5399.

Prendergast, J.G. et al., 2007. Chromatin structure and evolution in the human genome. *BMC Evolutionary Biology*, 7(1), p.72.

Rangwala, S.H. & Kazazian, H.H., 2009. The L1 retrotransposition assay: a retrospective and toolkit. *Methods (San Diego, Calif.)*, 49(3), pp.219–226.

Reed, K.E., 1997. Early hominid evolution and ecological change through the African Plio-Pleistocene. *Journal of human evolution*, 32(2-3), pp.289–322.

Rehen, S.K. et al., 2001. Chromosomal variation in neurons of the developing and adult mammalian nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23), pp.13361–13366.

Reintjes, N. et al., 2013. Activating Somatic FGFR2 Mutations in Breast Cancer S. Deb, ed. *PLoS ONE*, 8(3), p.e60264.

Richards, K.L. et al., 2009. Genome-Wide Hypomethylation in Head and Neck Cancer Is More Pronounced in HPV-Negative Tumors and Is Associated with Genomic Instability S. D. Fugmann, ed. *PLoS ONE*, 4(3), p.e4941.

Riethman, H., 2003. Mapping and Initial Analysis of Human Subtelomeric Sequence Assemblies. *Genome Research*, 14(1), pp.18–28.

Robberecht, C. et al., 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Research*, 23(3), pp.411–418.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature Publishing Group*, 29(1), pp.24–26.

Rodenas-Cuadrado, P., Ho, J. & Vernes, S.C., 2013. Shining a light on CNTNAP2: complex functions to complex disorders. *European Journal of Human Genetics*, pp.1–8.

Rodić, N. & Burns, K.H., 2013. Long Interspersed Element–1 (LINE-1): Passenger or Driver in Human Neoplasms? S. M. Rosenberg, ed. *PLoS Genetics*, 9(3), p.e1003402.

Rodriguez, J. et al., 2007. Genome-wide tracking of unmethylated DNA Alu repeats in normal and cancer cells. *Nucleic Acids Research*, 36(3), pp.770–784.

Roman-Gomez, J. et al., 2005. Promoter hypomethylation of the LINE-1 retrotransposable elements activates sense/antisense transcription and marks the progression of chronic myeloid leukemia. *Oncogene*, 24(48), pp.7213–7223.

Rozen, S. & Skaletsky, H., 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.)*, 132, pp.365–386.

Sakai, T. et al., 1991. Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *American journal of human genetics*, 48(5), pp.880–888.

Salem, A.-H. et al., 2003. LINE-1 preTa Elements in the Human Genome. *Journal of molecular biology*, 326(4), pp.1127–1146.

Sassaman, D.M. et al., 1997. Many human L1 elements are capable of retrotransposition. *Nature genetics*, 16(1), pp.37–43.

Schulz, W.A., 2006. L1 Retrotransposons in Human Cancers. *Journal of Biomedicine and Biotechnology*, 2006, pp.1–13.

Schuster-Böckler, B. & Ben Lehner, 2013. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412), pp.504–507.

Schwahn, U. et al., 1998. Positional cloning of the gene for X-linked retinitis pigmentosa 2. *Nature genetics*, 19(4), pp.327–332.

Sciamanna, I. et al., 2005. Inhibition of endogenous reverse transcriptase antagonizes human tumor growth. *Oncogene*, 24(24), pp.3923–3931.

Sen, S.K. et al., 2007. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Research*, 35(11), pp.3741–3751.

Shivapurkar, N. et al., 1999. Deletions of chromosome 4 at multiple sites are frequent in malignant mesothelioma and small cell lung carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 5(1), pp.17–23.

Shukla, R. et al., 2013. Endogenous Retrotransposition Activates Oncogenic Pathways in Hepatocellular Carcinoma. *Cell*, 153(1), pp.101–111.

Silva, F.P.G. et al., 2003. Identification of RUNX1/AML1 as a classical tumor suppressor gene. *Oncogene*, 22(4), pp.538–547.

Singer, T. et al., 2010. LINE-1 retrotransposons: mediators of somatic variation in neuronal

genomes? *Trends in Neurosciences*, 33(8), pp.345–354.

Smit, A.F. et al., 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *Journal of molecular biology*, 246(3), pp.401–417.

Solyom, S. et al., 2012. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Research.*

Solyom, S. et al., 2011. Pathogenic orphan transduction created by a nonreference LINE-1 retrotransposon. *Human mutation*, 33(2), pp.369–371.

Speek, M., 2001. Antisense Promoter of Human L1 Retrotransposon Drives Transcription of Adjacent Cellular Genes. *Molecular and cellular biology*, 21(6), pp.1973–1985.

Stamatoyannopoulos, J.A. et al., 2009. Human mutation rate associated with DNA replication timing. *Nature genetics*, 41(4), pp.393–395.

Stewart, C. et al., 2011. A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans H. S. Malik, ed. *PLoS Genetics*, 7(8), p.e1002236.

Stout, D. et al., 2008. Neural correlates of Early Stone Age toolmaking: technology, language and cognition in human evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1499), pp.1939–1949.

Stratton, M.R., Campbell, P.J. & Futreal, P.A., 2009. The cancer genome. *Nature*, 458(7239), pp.719–724.

Sundvall, M. et al., 2008. Role of ErbB4 in Breast Cancer. *Journal of Mammary Gland Biology and Neoplasia*, 13(2), pp.259–268.

Swergold, G.D., 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and cellular biology*, 10(12), pp.6718–6729.

Symer, D.E. et al., 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell*, 110(3), pp.327–338.

Szak, S.T. et al., 2002. Molecular archeology of L1 insertions in the human genome. *Genome biology*, 3(10), p.research0052.

Takahara, T. et al., 1996. Dysfunction of the Orleans reeler gene arising from exon skipping due to transposition of a full-length copy of an active L1 sequence into the skipped exon. *Human Molecular Genetics*, 5(7), pp.989–993.

Temperani, P. et al., 1995. Chromosome rearrangements at telomeric level in hematologic disorders. *Cancer genetics and cytogenetics*, 83(2), pp.121–126.

Thibodeau, S.N., Bren, G. & Schaid, D., 1993. Microsatellite instability in cancer of the proximal colon. *Science (New York, N.Y.)*, 260(5109), pp.816–819.

Thomas, C.A., Paquola, A.C.M. & Muotri, A.R., 2012. LINE-1 Retrotransposition in the Nervous System. *Annual Review of Cell and Developmental Biology*, 28(1), pp.555–573.

Thomas, R.K. et al., 2006. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature medicine*, 12(7), pp.852–855.

Ting, D.T. et al., 2011. Aberrant Overexpression of Satellite Repeats in Pancreatic and Other Epithelial Cancers. *Science (New York, N.Y.)*, 331(6017), pp.593–596.

Trask, B.J. et al., 1998. Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Human Molecular Genetics*, 7(1), pp.13–26.

van den Hurk, J.A.J.M. et al., 2003. Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon. *Human Genetics*, 113(3), pp.268–275.

Van Loo, P. et al., 2010. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(39), pp.16910–16915.

Veeriah, S. et al., 2009. Somatic mutations of the Parkinson's disease–associated gene PARK2 in glioblastoma and other human malignancies. *Nature Publishing Group*, 42(1), pp.77–82.

Venter, J.C., 2001. The Sequence of the Human Genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–1351.

Vidaud, D. et al., 1993. Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. *European journal of human genetics : EJHG*, 1(1), pp.30–36.

Wallace, N.A. et al., 2010. Feedback inhibition of L1 and alu retrotransposition through altered double strand break repair kinetics. *Mobile DNA*, 1(1), p.22.

Wallace, N.A. et al., 2013. HPV 5 and 8 E6 expression reduces ATM protein levels and attenuates LINE-1 retrotransposition. *Virology*, pp.1–11.

Wang, H. et al., 2005. SVA elements: a hominid-specific retroposon family. *Journal of molecular biology*, 354(4), pp.994–1007.

Wang, J. et al., 2006. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Human mutation*, 27(4), pp.323–329.

Wang-Johanning, F. et al., 2003. Quantitation of HERV-K env gene expression and splicing in human breast cancer. *Oncogene*, 22(10), pp.1528–1535.

Wei, W. et al., 2000. A Transient Assay Reveals That Cultured Human Cells Can Accommodate Multiple LINE-1 Retrotransposition Events. *Analytical Biochemistry*, 284(2), pp.435–438.

Wei, W. et al., 2001. Human L1 Retrotransposition: cis Preference versus trans Complementation. *Molecular and cellular biology*, 21(4), pp.1429–1439.

Wetterstrand, K.A., DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *genome.gov*. Available at: http://www.genome.gov/sequencingcosts/ [Accessed October 10, 2013].

Wimmer, K. et al., 2011. The NF1 Gene Contains Hotspots for L1 Endonuclease-Dependent De Novo Insertion N. B. Spinner, ed. *PLoS Genetics*, 7(11), p.e1002371.

Witherspoon, D.J. et al., 2010. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics*, 11(1), p.410.

Wolff, E.M. et al., 2010. Hypomethylation of a LINE-1 Promoter Activates an Alternate Transcript of the MET Oncogene in Bladders with Cancer B. Ren, ed. *PLoS Genetics*, 6(4), p.e1000917.

Xie, L. et al., 2013. FGFR2 Gene Amplification in Gastric Cancer Predicts Sensitivity to the Selective FGFR Inhibitor AZD4547. *Clinical Cancer Research*, 19(9), pp.2572–2583.

Xing, J. et al., 2009. Mobile elements create structural variation: Analysis of a complete human genome. *Genome Research*, 19(9), pp.1516–1526.

Xing, J., Witherspoon, D.J. & Jorde, L.B., 2013. Mobile element biology: new possibilities with high-throughput sequencing. *Trends in Genetics*.

Yoder, J.A., Walsh, C.P. & Bestor, T.H., 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends in genetics : TIG*, 13(8), pp.335–340.

Zang, Z.J. et al., 2012. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nature Publishing Group*, 44(5), pp.570–574.

Zhang, S. et al., 2012. REV3L 3′UTR 460 T>C polymorphism in microRNA target sites contributes to lung cancer susceptibility. pp.1–9.

Zingler, N., 2005. Analysis of 5" junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5-"end attachment requiring microhomology-mediated end-joining. *Genome Research*, 15(6), pp.780–789.