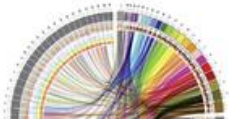
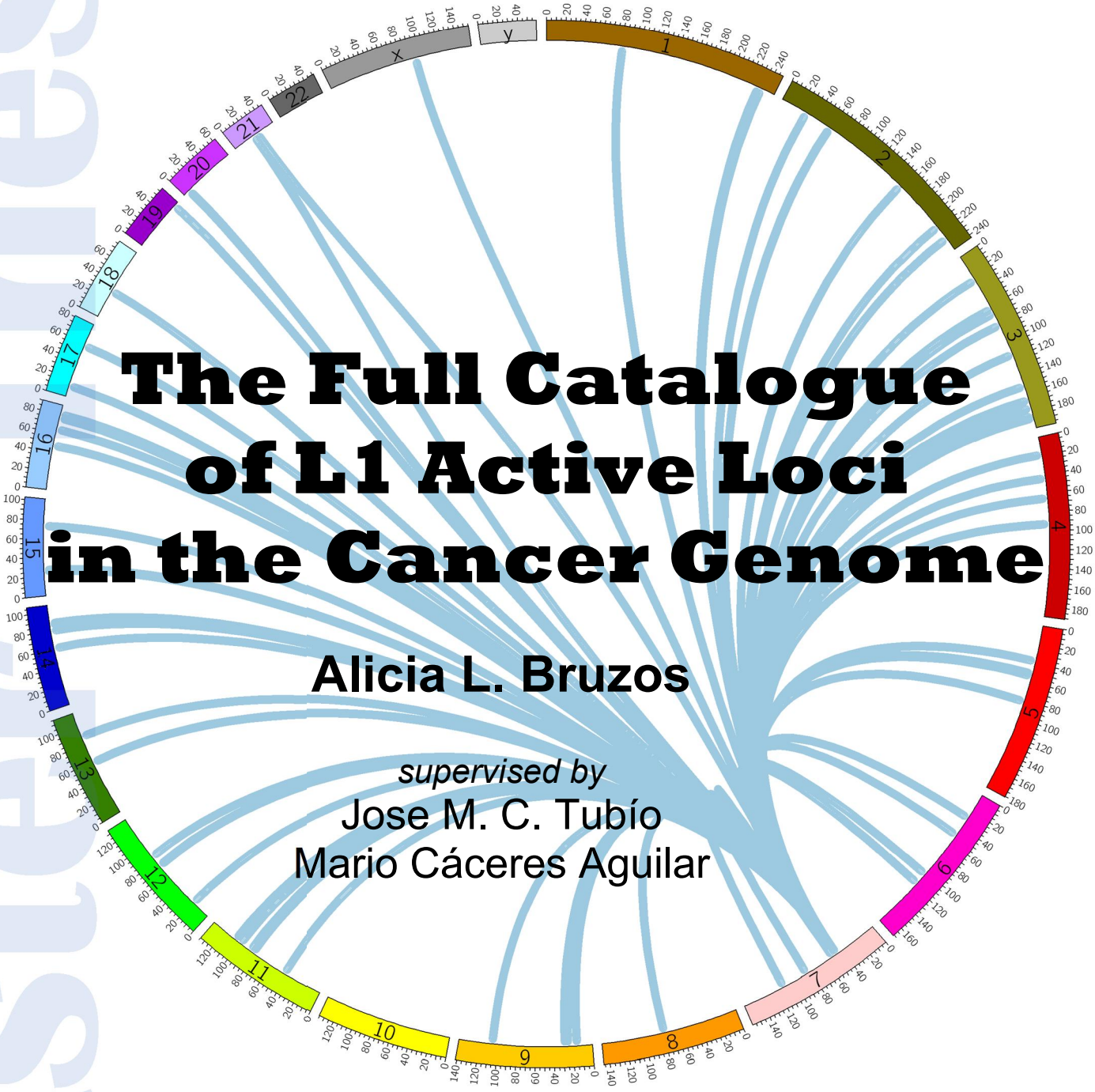
Master Thesis



MSc. in Bioinformatics



The Full Catalogue of L1 Active Loci in the Cancer Genome

Alicia L. Bruzos

supervised by

Jose M. C. Tubío

Mario Cáceres Aguilar

July 2016

International Cancer Genome Consortium

The Cancer Genome Atlas

Universidade Vigo

UAB Universitat Autònoma de Barcelona

Barcelona, July 4th, 2016

To whom it may concern:

We certify the student *Alicia L. Bruzos*, a candidate for Master's Degree in Bioinformatics at Universitat Autònoma de Barcelona, has successfully completed her Master thesis project entitled "The full catalogue of L1 active loci in the cancer genome", which has been reviewed by us, and we believe it has been completed.

Jose M. C. Tubío
Thesis director
University of Vigo

Mario Cáceres Aguilar
Academic tutor
Universitat Autònoma de Barcelona

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

— Marie Skłodowska-Curie

Acknowledgements

I would like to express my deep gratitude to my master thesis director, *Dr. Jose M. C. Tubío*, who gave me the opportunity to develop this research project and because this work could not have been done without his help and ideas. I am thankful, as well, to my academic tutor, *Dr. Mario Cáceres Aguilar*, for his helpful discussions.

Secondly, I am very grateful to the *Phylogenomics group* at University of Vigo for the opportunity to work with a very professional team. Particularly, I would like to thank the IT technician of the group, *Rubén Fernández Lago,* and three PhD students, *Merly Escalona*, *Bernardo Rodríguez Martín,* and *Adrián Báez Ortega* (University of Cambridge, UK) for sharing their bioinformatics skills with me.

My thanks are extended to all the researchers involved in the *Pan-Cancer analysis of Whole Genomes* (PCAWG) *project* of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA).

Last, but not least, I would like to thank all those people who taught me how to get here and believe in my scientific career: *family*, *friends*, *classmates* and *teacher**s***. Thank you all, because without you this Master's Thesis would have never been written.

Table of Contents

1. Abstract

Retrotransposons are repetitive elements that are constantly on the move. By poaching certain enzymes, they copy and insert themselves at new sites in the host genome generating structural variability of potential functional importance for the cancer cell. Retrotransposons can also promote genomic rearrangements by recombination, and mobilize coding and regulatory regions by transduction. Previous work showed that about half of human tumours have a variable number of somatic retrotranspositions, ranging from less than ten to several hundreds, which are caused by a relatively low number of L1 source elements. These germline L1 active elements are the source for *de novo* somatic mutations that may help cancer to survive and progress.

In this study, we aimed to catalogue the full set of L1 germline elements with somatic ativity in cancer, by analyzing somatic retrotransposition events on the largest dataset ever generated in a single project for the study of cancer retrotransposition, which consists of 2,704 cancer samples from 39 different tumour types. We run the bioinformatic pipeline TraFiC and identified 22,838 somatic retrotransposition events across the cancers analyzed. L1 elements dominate somatic retrotransposition, representing ~87% of the total retrotransposition events. Overall, ~46% of the cancer genomes analyzed have a least one L1 somatic retrotransposition event, being more frequent in lung squamous carcinoma, where 100% of the samples have a retrotransposition event, followed by esophagus cancer with 98%. Esophagus cancer has the highest retrotransposition rate, with an average of ~96 retrotranspositions per sample, followed by head-and-neck cancer, lung squamous carcinoma, and colon adenocarcinoma. L1 3'-transductions represent 20% of the total L1 somatic events. We used these transductions to identify the L1 source elements whence they derive, and found 37 novel source elements that were not reported in previous cancer retrotransposition studies, including a novel hot-L1 source element at chromosome 7. Multiple source elements can be active in single cancer genomes, representing a significant source of mutations in a tumour. Added to the 72 previously reported L1 source loci, it makes a total of 109 of germline L1 source loci with somatic retrotransposition activity in the cancer genome.

2. Introduction

The study of human genetic variation revealed great similarity between individuals; in fact, two humans are ~99.9% identical at the DNA sequence level (Lander et al. 2001; Venter et al. 2001; Hattori 2005). Thus, only a small fraction of the genome varies among individuals and it keeps the keys of phenotypic variation and susceptibility to certain diseases; hence the importance to understand the genetic variation in the human population (Reich et al. 2002; Feuk et al. 2006). A proportion of human genetic variation correspond to 'genomic structural variants' (SVs), which is the variation in the organization of the DNA molecule, typically generated during DNA break-induced repair, recombination or replication (Hastings et al. 2009; Jennes et al. 2011).

Somatic structural variation in the cancer genome

Along the development of a human being, the genome sequence of a cell acquires a set of differences relative to the fertilized egg from which is a descendant; these differences are termed somatic mutations**,** to distinguish them from the germline mutations, which are inherited from parents and transmitted to offspring (Stratton et al. 2009; Griffiths et al. 2000).

Almost all (if not all) cancers are the result of somatically acquired genetic mutations in the cells of the cancer lineage. However, that does not mean that all the somatic alterations accumulated in the genomes of a cancer lineage have contributed to the origin and/or development of the cancer. Actually, most mutations are neutral for the propagation of the cancer clone (passenger mutations), while only a very small fraction of somatic changes, called driver mutations, may result in the alteration of key genes that drive oncogenesis (Stratton et al. 2009).

Somatic structural variation (i.e., those genomic rearrangements arising in a tumour) has been found in almost all cancers studied in detail. Although most of these changes are thought to be selectively neutral during the evolution of the disease, some

play a major role as drivers of the oncogenic process (Calin et al. 2002; Fabbri et al. 2005; Ren 2005). For the last years, cancer genomics has benefited greatly from next-generation sequencing (NGS) technology. Prior to the advent of NGS, strategies for the identification and characterization of rearrangements were laborious and had low sensitivity and resolution. Current technology using paired-end NGS enable the identification of genomic structural variation at unprecedented base pair resolution (Korbel et al. 2007), revealing a complex landscape of somatic rearrangements in cancer genomes (Stephens et al. 2009; Campbell et al. 2010; Stephens et al. 2011; Campbell et al. 2008).

Thousands of cancer genomes from tens of different cancer types have been sequenced and analyzed to date, showing a variable number of genomic structural rearrangements acquired somatically, ranging from zero to thousands, and a variable relative proportion of rearrangement types (Yang et al. 2013). Most of these studies have focused on certain classes of rearrangement, namely, deletions, amplifications, inversions and translocations. Recently, the development of new algorithms allowed the identification of genomic insertions acquired somatically, a type of structural variation that includes retrotransposition (Helman et al. 2014; Lee et al. 2013; Tubio et al. 2014).

Transposable elements

Broadly, transposable elements (TEs), also called mobile genetic elements, are defined as repetitive DNA sequences that have the ability to move in the genome where they reside. First discovered in maize by Barbara McClintock in the 1940's, TEs have been detected in all eukaryotic genomes sequenced to date, differing in their abundance and diversity across species (Hattori 2005; Lander et al. 2001; Venter et al. 2001). In Humans they represent up to 45% of the total genome size (Lander et al. 2001), although some studies estimate that about two-thirds of our genome may have resulted from the activity of mobile genetic elements (Koning et al. 2011; Solyom & Kazazian 2012).

The mechanism of TE mobilization is called transposition and, depending on the mode of transposition, TEs are classified into two categories (Figure 1A): DNA transposons (or Class II), which are mobilized by a cut-and-paste mechanism by a transposase, and Retrotransposons (or Class I), which are mobilized by a copy-and-

paste mechanism that involves an intermediary mRNA that is retrotranscribed prior to the integration of the new TE copy into a new location of the genome (Hancks & Kazazian 2012; Wicker et al. 2007).

Retrotransposons are further subclassified (Figure 1B) into those flanked by long terminal repeat (LTR), and non-LTR elements. In the Human genome, LTR retrotransposons are represented by human endogenous retroviruses (HERVs). Non-LTR retrotransposons are classified into long interspersed elements (LINE), which includes LINE-1 (L1) retrotransposons, and short interspersed elements (SINEs), which include Alu and SVA retrotransposons. While LINEs are autonomous elements that encode for all the enzymes required to complete their transposition cycle, SINEs are non-autonomous, requiring the enzymes encoded by other elements (Kazazian 2004). In the human genome, LINEs and SINEs represent ~27% of the total genome size (Figure 1C), being the dominant TE type.

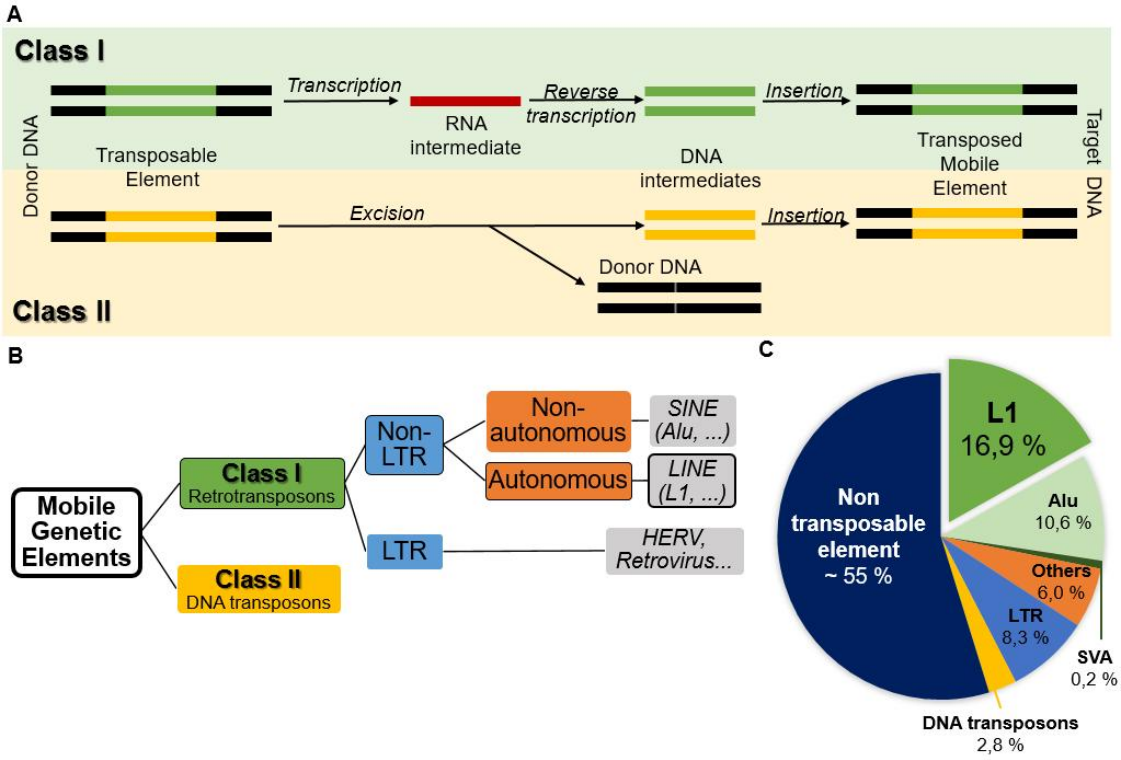


Figure 1. **(A)** TEs classification: *Class I* include retrotransposons that transpose via an RNA intermediate. The element is transcribed, and the transcript is copied into DNA by a reverse transcriptase encoded by autonomous elements. The new DNA copy is integrated elsewhere in the genome. *Class II* are transposons that excise themselves from the genome and move into a new location (Lisch 2012). **(B)** In the human genome, the majority of retrotransposons are non-LTR retrotransposons, a group that includes

SINEs and LINEs (Koito & Ikeda 2013). **(C)** In the human genome, TEs represent 45% of the total genome size (Cordaux & Batzer 2009).

L1 retrotransposons

L1 elements are the master retrotransposons in mammalian genomes (Figure 2A) (Ostertag & Kazazian 2001). Besides duplicating themselves, they also cause the genomic expansion of the non-autonomous retrotransposons Alu and SVA, and of processed pseudogenes (Ostertag & Kazazian 2001).

The human genome bears about half a million copies of L1, which represents ~17% (Figure 1C) of the total genome size (Belancio et al. 2010; Brouha et al. 2003). The vast majority of these L1 copies represent remains of ancient activity, which have been inactivated as a result of truncation, mutation, and internal rearrangement.

It has been estimated that in the human genome there are about 100 active L1 loci, which preserve their full structure (Seluanov et al. 2015; Solyom et al. 2015; Brouha et al. 2003). Only a handful of this active elements are known to be highly active — earning the moniker 'hot-L1' — (Brouha et al. 2003; Beck et al. 2010; Solyom et al. 2015).

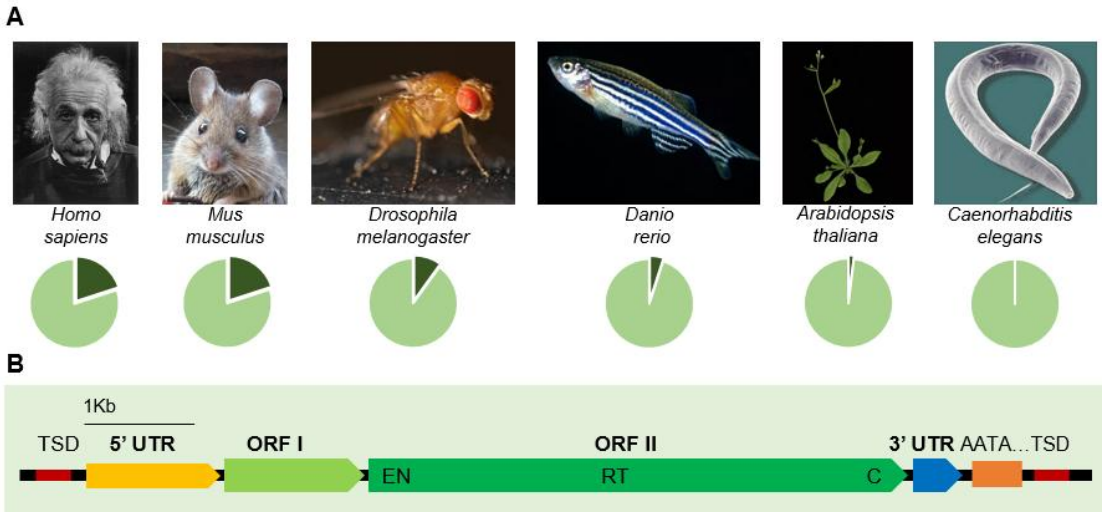


Figure 2. **(A)** LINE-1 elements are present in a wide-range of animal species (Huang et al. 2012) highlighting in dark green the proportion of L1 for that specie. **(B)** The structure of a typical full-length human L1 element consist of a 5' UTR, two ORFs separated by a short intergenic region, a 3' UTR, a polyA signal (AATAAA), and a

poly(A) tail. L1 elements are often flanked by 7–20bp target site duplication (TSD) (Boissinot et al. 2004; Ostertag & Kazazian 2001).

Active L1 retrotransposons (Figure 2B) are around 6 kb long (Brouha et al. 2003). L1 elements have a 5' untranslated region (UTR) with internal promoter activity, two open reading frames (ORFs), a 3' UTR that ends in an AATAAA polyadenylation signal, and a polyA tail (Ostertag & Kazazian 2001). They encode a RNA-binding protein and an endonuclease with reverse-transcriptase activity, which allow the element to autonomously replicate in the host genome via the retrotranscription of a RNA intermediate (Seluanov et al. 2015).

Transcription is initiated from an internal promoter located within its 5' UTR (Swergold 1990), and the RNA is transported outside of the nucleus. Once in the cytoplasm, the ORF1 and ORF2 proteins are translated. At least one L1 RNA molecule, one ORF2 molecule, and one or more ORF1 molecules may assemble into a ribonucleoprotein (RNP) complex that is an intermediate in retrotransposition (Esnault et al. 2000; Goodier 2014; Kazazian & Moran 1998). The resultant RNP complex re-enters the nucleus where L1 integration is thought to occur by target-primed reverse transcription (TPRT) (Brouha et al. 2003; Luan et al. 1993). The TPRT process creates 7–20-bp target site duplications that flank the L1 element. Many elements undergo 5' truncation, or 5' inversion and truncation, during the TPRT process, resulting in an inactive DNA copy of the original element (Ostertag & Kazazian 2001).

The impact of transposable elements on genome function and disease

Despite the majority of TE insertions in the germline are likely to be phenotypically silent, some TEs have been identified which lead to severe phenotypic consequences and disease (O'Donnell & Burns 2010). There are several mechanisms by which integration of TEs can impact gene function (Beck et al. 2011); and most of them are reviewed in Figure 3. Moreover, TEs can also promote chromosomal rearrangements by ectopic recombination (Lim & Simmons 1994). The resulting rearrangements may have important consequences for gene function: The *Antp73b* mutation in *Drosophila*, in which antennae are transformed into second legs, results from such phenomenon (Schneuwly et al. 1987). In humans, hair growth deficiency is linked to a deletion of exon 4 of gene LIPH, which is caused by the recombination between homologous regions of two Alu elements (Kazantseva et al. 2006). Finally, L1 and SVA

retrotransposons can promote the mobilization of pseudogenes in trans, and the mobilization of adjacent (usually downstream) genomic regions, a mechanism called L1-mediated transduction.

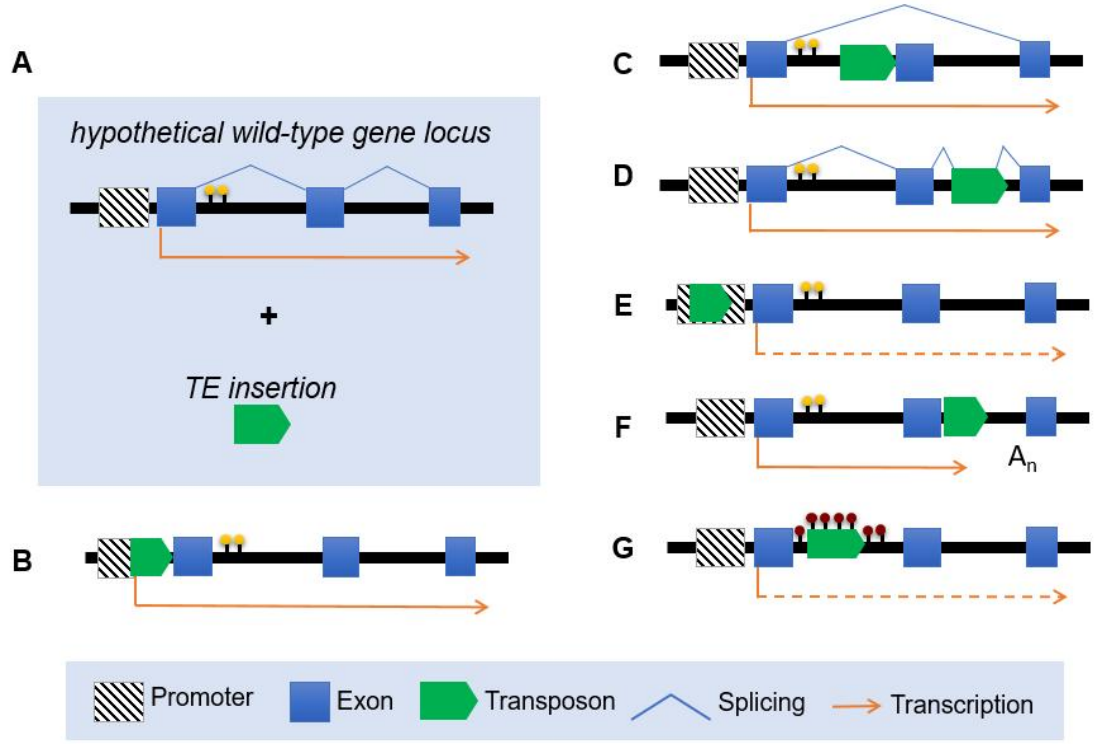


Figure 3. Diagram highlighting some mechanisms by which insertional mutagenesis of TEs can impact genome function. **(A)** A hypothetical wild-type gene locus. Grey boxes represent exons; black line represents introns; striped black rectangle the promoter region. Normal transcription is represented by an orange arrow; normal splicing by blue lines over the gene; and normal methylation profile by red lollypops. **(B)** Introduction of novel transcription start sites by the TE promoter. **(C)** Alternative/aberrant splicing via exon skipping. **(D)** Alternative/aberrant splicing by exonization (i.e., creation of a new exon as a result of mutations in intronic sequences). **(E)** Deregulation due to gene promoter/enhancer disruption; discontinued red arrow indicates aberrant transcription. **(F)** Premature end of transcription by addition of new polyadenylation 3' ends. **(G)** TEs can also be epigenetically silenced during or after their integration, being affected the epigenetic status of adjacent regions as well (hypermethylation is represented by red lollypops and consequent impact on transcript levels by discontinued red arrow).

Little is known about the extent to which TEs can generate diversity in somatic cells and contribute to the development of cancer. For the last years, several NGS studies intended the identification of somatic retotrotransposition in cancer genomes (Helman et al. 2014; Lee et al. 2013; Solyom & Kazazian 2012; Tubio et al. 2014).

These works revealed high retrotransposition rates in lung and colorectal cancers, where tens to hundreds of somatic integrations of L1 retrotransposons are frequently found in their genomes.

L1 transcription often mobilizes unique DNA located downstream of the element by a process called 3'-transduction (Solyom & Kazazian 2012) (Figure 4). This mechanism is the consequence of a weak transcription termination signal at the end of the L1, causing the transcription machinery to bypass the first L1 polyadenylation signal and use another polyadenylation site located downstream, incorporating to the mRNA unique DNA material located downstream, and causing its mobilization and integration elsewhere in the genome. In cancer, L1-mediated transductions represent 20% of all somatic retrotranspositions, and have a strong potential to impact cancer genome function, given by their ability to duplicate and spread exons, genes, and regulatory regions (Tubio et al, 2014).

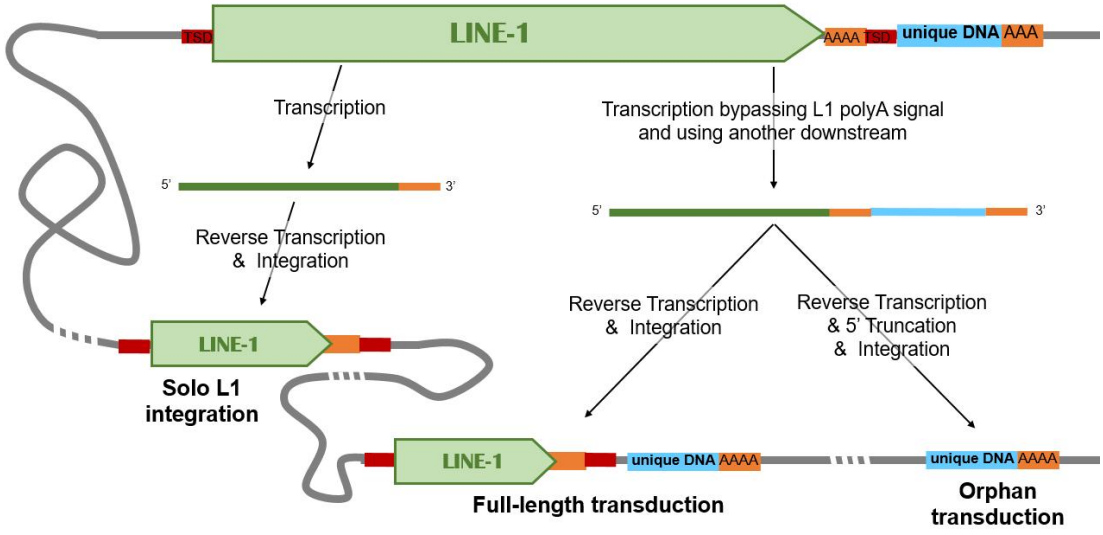


Figure 4. L1 3'-transductions are the mobilization of DNA sequences located downstream to the element. The retrotransposition of L1 involves the transcription of the element, which starts in the 5´extreme of the element and finishes in the poly tail located at the 3´extreme. Because the polyA tail is not strong enough to stop transcription, the polymerase sometimes overpasses this first polyA tail, incorporating to the transcript unique DNA sequences. After retrotranscription and integration, in the insertion point we will see the element together with a copy of the unique material; this is a process called full length 3´transduciton. Nevertheless, sometimes there is a truncation just before integration producing an orphan 3' transduction.

3. Hypothesis

From the nearly ~500,000 L1 loci that bears the human genome (Lander et al. 2001), only a handful of elements are full-length copies able to retrotranspose (Sassaman et al. 1997; Beck et al. 2010; Brouha et al. 2003). These active elements may carry along adjacent unique genomic DNA sequences, a process called 3′-transduction (Moran et al. 1999).

Transductions can be used to, unambiguously, identify the L1 source element whence they derive. So far, researchers have identified 72 transduction-competent L1-loci in the cancer genome (Tubio et al. 2014), but we hypothesize that many other L1 source elements remain undiscovered. The main purpose of this research project is to identify the full set of active L1 loci (i.e. those transduction competent L1 loci), using bioinformatic tools to analyse 2,704 cancer genomes generated within the framework of the Pan-cancer initiative.

4. Objectives

The present study aims to catalogue the full set of germline L1 loci active in the cancer genome. We use bioinformatic resources, pipelines and tools, generated within the framework of the International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA), to identify all germline L1 copies competent for 3'-transduction in 2,705 cancer genomes from 20 different cancer types.

There are four specific objectives:

1) Run the pipeline TraFiC (Transposome Finder in Cancer) for the identification of somatic retrotransposition in cancer genomes.
2) Development of the scripts necessary for the curation of the dataset generated by TraFiC, for the identification of source L1 copies.
3) Validation of source L1 loci by visual inspection of the read-mapping data.
4) Representation of somatic retrotransposition using Circos and R-scripting.

5. Materials and Methods

5.1. Workflow

A schematic diagram of the workflow is shown in Figure 5.

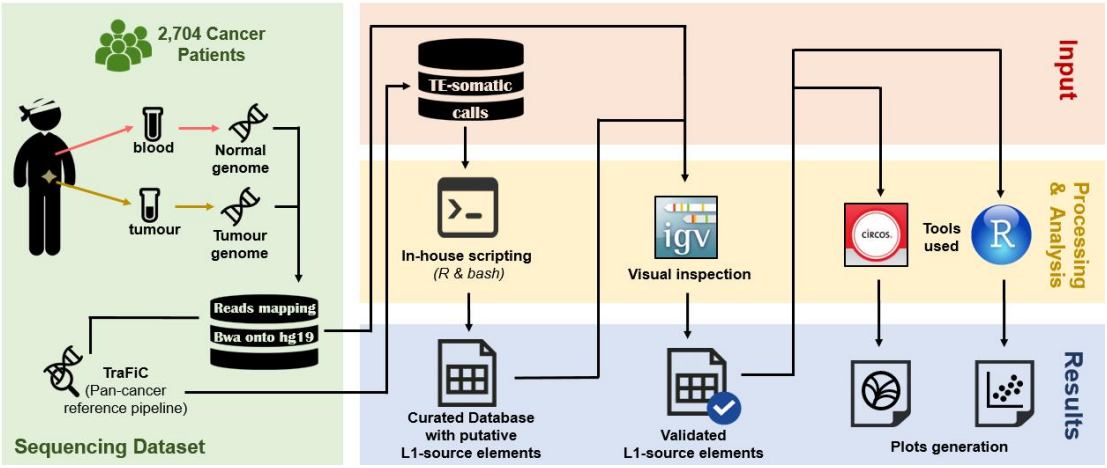


Figure 5. Tumour and normal genomes from 2,704 cancer patients were sequenced and mapped onto the human reference genome. Then, we ran TraFiC (Tubio et al. 2014) on this dataset to identify somatic retrotransposition in cancer, and we developed the scripts necessary for the curation of the somatic calls from TraFiC, in order to get a list of putative L1 source elements. Source elements were validated by visual inspection using the Integrative Genomics Viewer IGV (Thorvaldsdóttir et al. 2013; Robinson et al. 2011) and, finally, we used tools like Circos (Krzywinski et al. 2009) and R-scripting to plot the results.

5.2. Sequencing data

Pan-cancer is a joint initiative between the International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA) that aims the analysis of the full landscape of molecular aberrations that characterize 2,834 cancer genomes, together with RNA sequencing data, methylation profiles and clinical data (Weinstein et al. 2014).

Taking advantage of the enrollment of Dr Tubio's laboratory in the Pan-cancer initiative, a total of 2,750 cancer genomes and their matched normal genomes from

patients across 20 cancer types were analyzed in our study (Table S1). These genomes were sequenced up to 35x coverage by paired-end sequencing using Illumina technology (libraries of ~300 nucleotides and sequencing reads of 100 nucleotides). Sequencing reads were mapped onto the human reference genome (hg19) using Bwa (Li & Durbin 2009).

5.3. Identification of somatic retrotransposition

TraFiC (Transposome Finder in Cancer), developed by Tubio et al (2014), is the reference pipeline for the identification of somatic retrotransposition within the framework of the Pan-cancer initiative. The pipeline is capable of detecting three types of L1 retrotranspositions: solo-L1 events, partnered transductions, and orphan transductions. TraFiC relies on the identification of genomic hallmarks of retrotransposition at both the integration point and the L1 source element locus.

5.4. Identification of novel L1 source elements

The main piece of bioinformatic work carried out for this Master thesis intended the identification of novel (unreported) germline L1 source elements with somatic activity in cancer genomes, by the curation of the data provided by TraFiC. Firstly, we developed a collection of scripts, written in R and bash, to identify high-confident germline L1 source elements, which were defined as those L1 source elements giving rise to somatic transductions in two or more cancer samples. All the scripts were stored in the following GitHub repository:

https://github.com/ALBruzos/Bioinformatics_MasterThesis

Secondly, we used the tool IGV Browser (Thorvaldsdóttir et al. 2013; Robinson et al. 2011) to carry out visual inspection of the paired-end mapping data in the alignment files, and confirm these high-confident L1 master copies by the identification of the following hallmarks: (1) the element is full-length (~6 kb long), indicating that it preserves the promoter region required for its transcription; (2) the element has a poly(A) tail in the 3'-extreme, a hallmark of retrotransposition which will tell us about the orientation of the master copy; (3) the element has a target site duplication, which will be used to determine the integration breakpoint at the base pair level; (4) the source element is present in the tumour genome but also in the matched-normal genome, indicating the germline status of the element.

6. Results and discussion

The main motivation for launching this master thesis project is that germline L1 active loci are source for *de novo* somatic mutations that may help cancer to survive and progress (Tubio et al. 2014) and, consequently, in some way they can be seen as loci with potential oncogenic behaviour.

Because L1 3′-transductions are defined by the retrotransposition of unique genomic sequence located downstream (Moran et al. 1999), they can be used to unambiguously identify the L1 active element whence they derive (Tubio et al. 2014). Thus, to identify the full-catalogue of germline L1 source elements that are competent for somatic retrotransposition in cancer, we would need to survey somatic transductions on a considerably large scale, across thousands of cancer genomes. To this respect, the Pan-cancer initiative represents an unprecedented opportunity for the discovery of new L1 source elements, as it intends to catalogue the full set of somatic genomic alterations in a 2,834 cancer samples and their host genomic DNA (Weinstein et al. 2013).

Identification of somatic retrotransposition and transductions

For the last years, the advent of next-generation sequencing technologies has allowed the development of bioinformatic algorithms for the identification of somatic retrotransposition (Lee et al. 2013; Helman et al. 2014; Tubio et al. 2014). In cancer, several studies showed that about half of all human tumours have a variable number of somatic retrotranspositions, ranging from less than ten to several hundreds (Lee et al. 2013; Helman et al. 2014; Tubio et al. 2014; Solyom & Kazazian 2012).

In this study, we ran TraFiC for the identification of somatic retrotransposition on 2,704 cancer samples (and their normal host DNA) from patients across 39 different tumour types (Table S1). This dataset is the largest ever generated in a single project for the study of somatic retrotransposition. TraFiC found 19,411 L1 somatic retrotranspositions, which represents ~87% (19,441/22,383) of the total retrotransposition events mobilized somatically, and confirms L1 as the dominant retrotransposon type active in the cancer genome (Lee et al. 2013; Helman et al. 2014;

Tubio et al. 2014; Solyom & Kazazian 2012). In the head-and-neck cancer sample *9988eb07-01f6-4f83-8699-bb63e0525f08*, for example, TraFiC found 738 L1 somatic events, 3 Alu, and 1 SVA.

Overall, ~46% (1,258/2,704) of the cancer genomes analyzed have a least one L1 somatic retrotransposition event, most frequently lung squamous carcinoma and rectum adenocarcinoma, where 100% of the samples have a retrotransposition event, followed by esophagus cancer with 98% (Figure 6A). Esophagus cancer has the highest retrotransposition rate, with an average of ~96 retrotranspositions per sample, followed by head-and-neck cancer (52 retrotranspositions per sample), lung squamous carcinoma (39 retrotranspositions per sample), and colon adenocarcinoma (31 retrotranspositions per sample) (Figure 6B).

L1 3'-transductions represent ~20% (3,881/19,411) of the total L1 somatic events, of which 36% are full-length (1,402 insertions) and 64% orphan transductions (2,479 insertions). For example, in the esofagous cancer sample b8f3137e-5e92-4a56-90d4-884a4ed2ef9c, the activity of L1 produced 496 somatic insertions of which 186 (37%) are L1-mediated 3'-transductions. These results corroborate preliminary analyses of somatic L1 activity, which also showed that L1 transductions represent a substantial contribution to the mutational landscape of cancer genomes (Tubio et al. 2014).

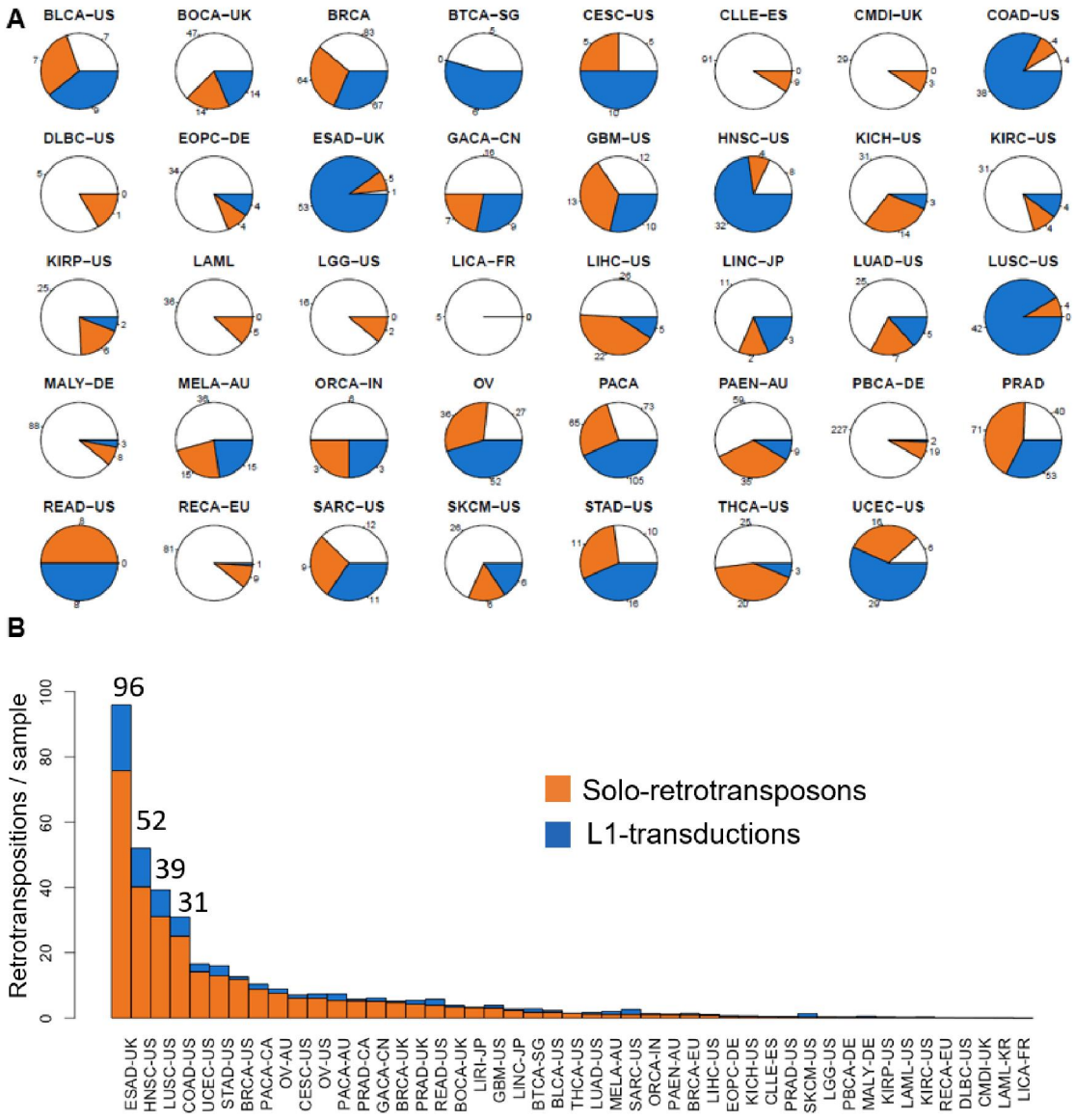


Figure 6. The somatic L1 retrotransposition activity in 2,704 cancer samples **(A)** Proportion of samples with L1 somatic retrotransposition across 39 cancer types (white, samples with no L1 retrotransposition; orange, samples with solo-L1 retrotranspositions but no transductions; blue, samples with at least one L1 transduction). **(b)** Bars show the average number of L1 somatic retrotransposition events per sample in each tumour type (orange, solo-L1 retrotranspositions; blue, L1-transductions).

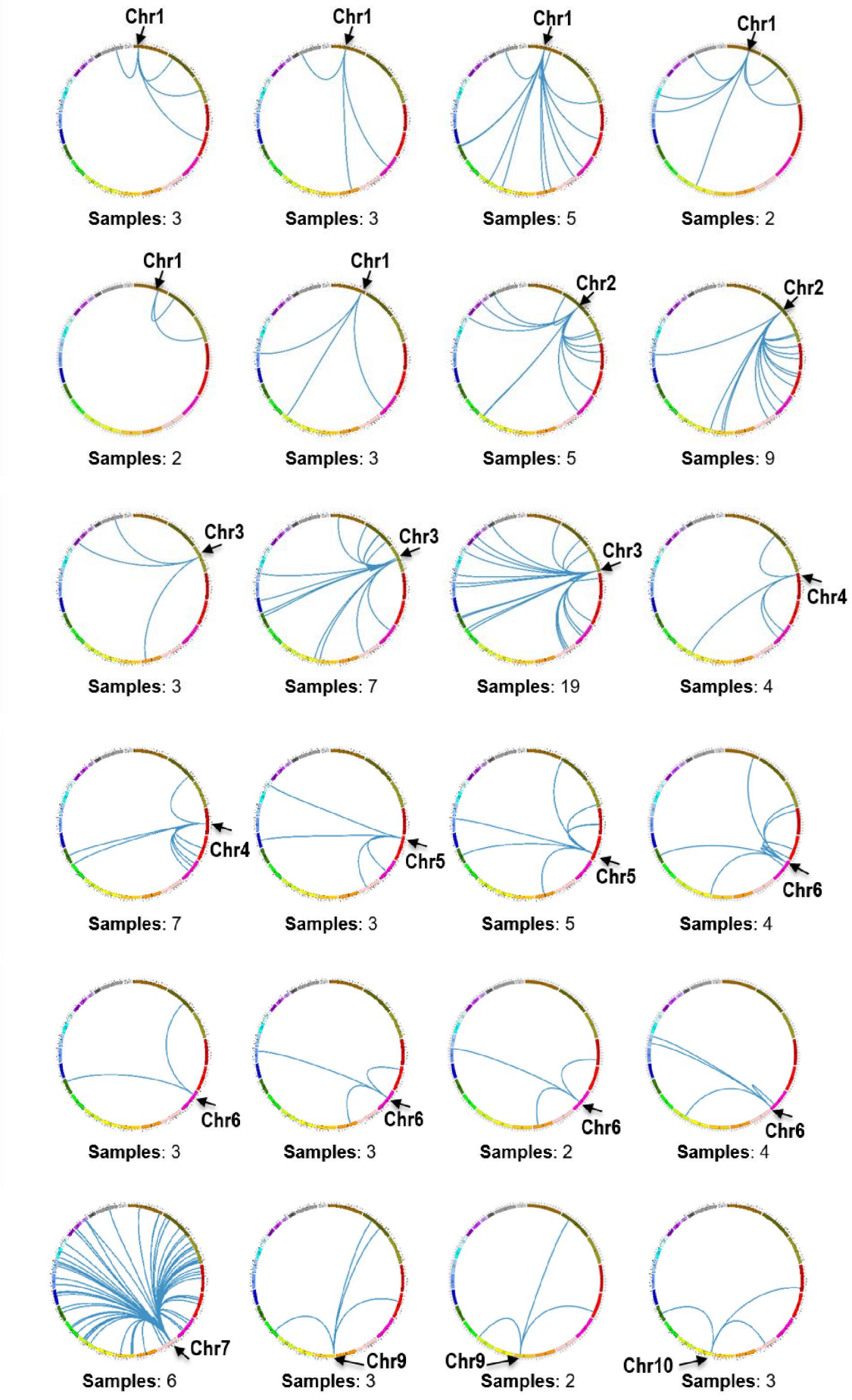
Novel (unreported) L1 source elements in the cancer genome

Several in-vitro retrotransposition assays using artificial constructions of human L1 elements estimated that, from the nearly 500,000 L1 copies present in a human genome, only 50-120 L1 elements are competent for retrotransposition (Beck et al. 2010; Brouha et al. 2003; Sassaman et al. 1997). From this list, only a small set of elements show high activity rates, earning the moniker of hot-L1 loci.

In a pilot study of the Pan-cancer initiative, the analysis of 290 cancer samples from 12 cancer types showed that somatic retrotransposition in cancer is dominated by at least 72 L1 active source elements (Tubio et al. 2014). From these retrotransposition-competent L1-copies, only four elements showed high activity rates, causing tens of transductions in individual cancer genomes and half of the total number of transductions found across the cancers analysed.

Similarly, in our dataset, although about half of human tumours have a variable number of somatic retrotranspositions, ranging from less than ten to several hundreds, these retrotransposition events are caused by a relatively low number of L1 source elements. We looked for high-confident novel source elements - i.e., those putative source elements giving rise to somatic transductions in 2 or more cancer samples -, finding 47 loci. This filter removed 116 putative L1 loci present in one-single cancer sample, which most likely represent L1 source copies of somatic origin or TraFiC miscalls.

We then carried out validation on the 47 high-confident candidate source elements by visual inspection of the bam files using IGV (Thorvaldsdóttir et al. 2013). L1 source elements (1) must be ~6, 100 nucleotides long, for preserving the promoter region that allow their transcription, and (2) may keep some diagnostic hallmarks of retotransposition, including the presence of a poly(A) tail in the 3'-extreme and a target site duplication (TSD). The visualization analysis with IGV allowed us to confirm the full-length status of most candidate L1 loci, and the identfication of poly(A) tail and target site duplication. By this method, we validated 37/47 source elements not reported in previous cancer retrotransposition studies, and identified their integration breakpoints into the cancer genome to base pair resolution (Figure 7; Table S2).


Chr1
Samples: 3
Chr1
Samples: 3
Chr1
Samples: 5
Chr1
Samples: 2
Chr1
Samples: 2
Chr1
Samples: 3
Chr2
Samples: 5
Chr2
Samples: 9
Chr3
Samples: 3
Chr3
Samples: 7
Chr3
Samples: 19
Chr4
Samples: 4
Chr4
Samples: 7
Chr5
Samples: 3
Chr5
Samples: 5
Chr6
Samples: 4
Chr6
Samples: 3
Chr6
Samples: 3
Chr6
Samples: 2
Chr6
Samples: 4
Chr7
Samples: 6
Chr9
Samples: 3
Chr9
Samples: 2
Chr10
Samples: 3

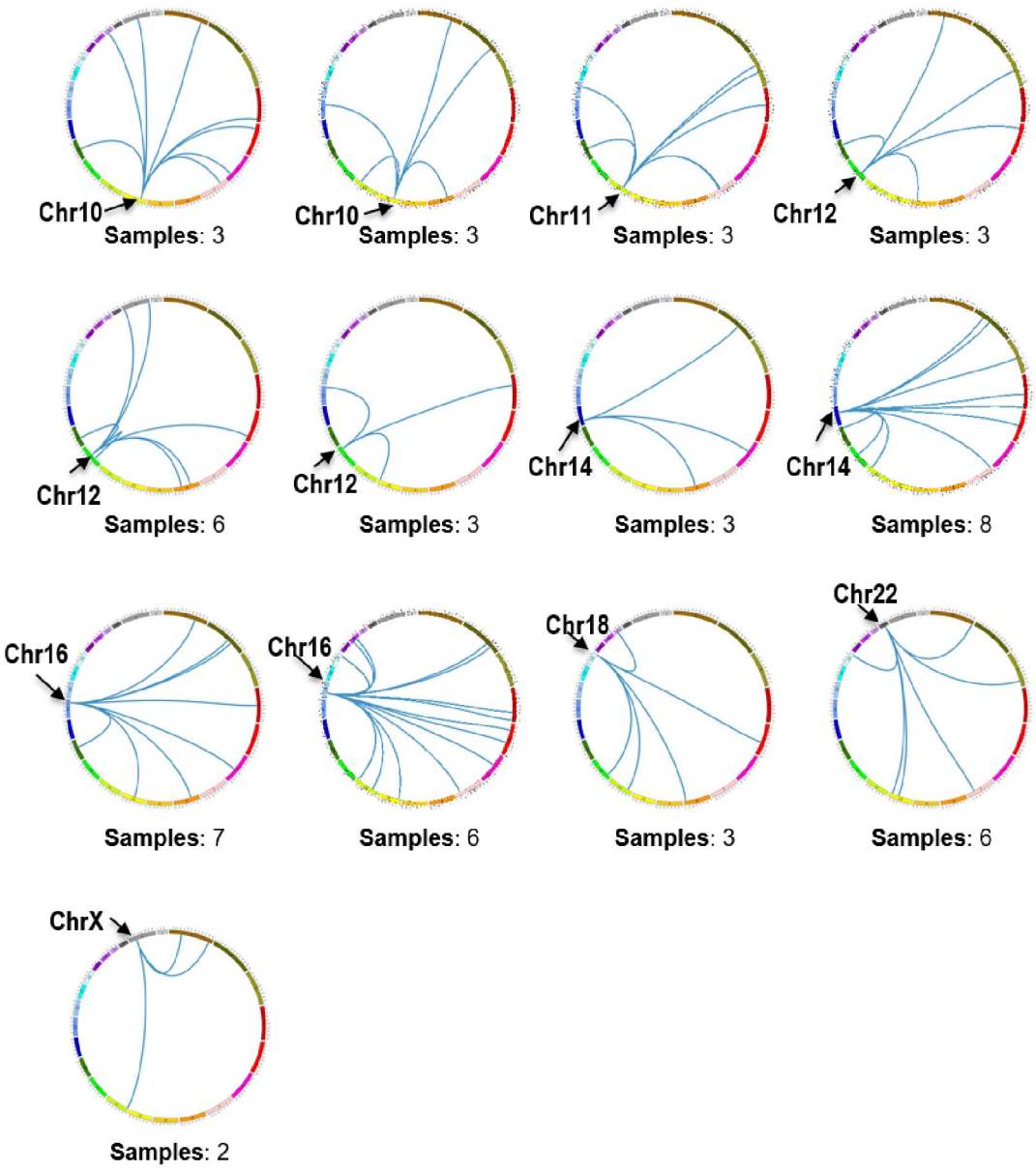


Figure 7. The catalog of 37 novel L1 source elements reported in this study. Each circos plot represents the activity of a germline source element. Arrows indicate the genomic location of each L1 element; blue lines connecting chromosomes represent somatic L1-transductions. The number of samples where the element is active is also shown below the circos plot.

Some source elements, called hot-L1, can give rise to tens of transductions in a single cancer genome. The somatic retrotransposition pilot study carried out by Tubio et al (2014) revealed that just four hot-L1 loci cause half of the total number of L1-transductions acquired somatically in a cancer. Similarly, here we report a novel hot-L1 element, located at chromosome 7p12.3 (Table S2; Figure 7), which showed somatic activity in 6 cancer samples and gave rise to 78 total transductions.

Adding the 37 novel germline L1 source elements to the 72 previously reported in the Pan-cancer retrotransposition pilot study carried out by Tubio et al (2014), make a total of 109 of germline L1 source loci that are active in the cancer genome (Figure 8). Cataloguing these loci is very valuable for the correct characterization of the landscape of structural variants in cancer genomes, because transductions are frequently misclassified as chromosomal translocations by current algorithms (Tubio et al, 2014).

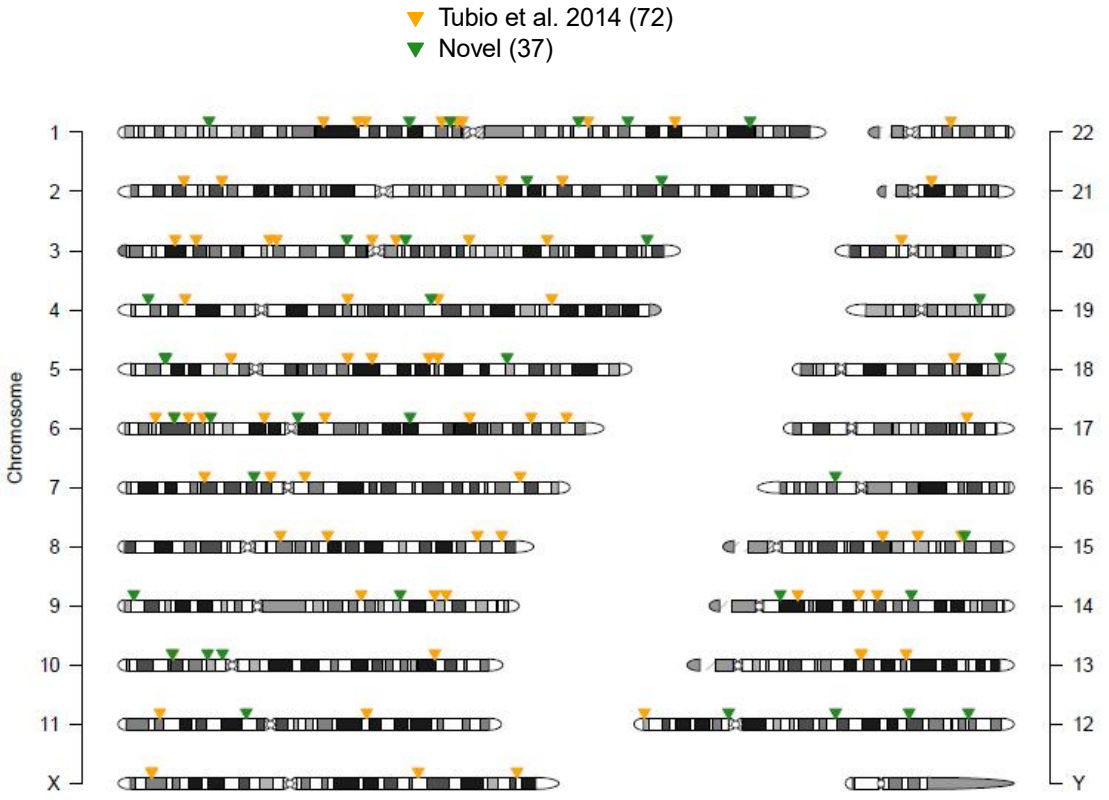


Figure 8. The genomic topography of 109 germline L1 source elements active in human cancers. Triangles represent the genomic location of each source element (orange triangles, the source elements reported by Tubio et al. 2014; green triangles, the novel source elements reported in this study).

Multiple source elements can be active in single cancer genomes, representing a significant source of mutations in a tumour. In an oesophageal carcinoma sample, for example, we found up to 10 source L1 elements active that gave rise to 83 somatic transductions (Figure 9).

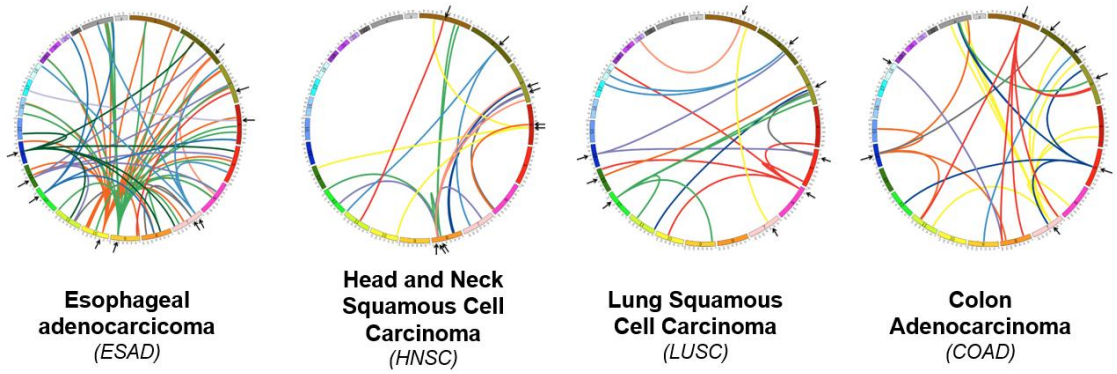


Figure 9. Multiple different germline L1 source elements can segregate somatic transductions in single tumour samples. From left to right: esophageal cancer sample where 10 source elements gave rise to 83 transductions; a head-and-neck cancer sample with 9 L1 active loci and 21 transductions; a lung squamous carcinoma sample where 9 source elements gave rise to 19 somatic transductions; and a colon adenocarcinoma sample with 31 transductions promoted by 8 L1 loci.

L1 source elements activity may impact cancer genome function

L1 source copies act as a source for new retrotransposition events in a cancer lineage. Although the majority of these retrotransposition events are expected to be passenger mutations, there are multiple ways by which somatic retrotransposition of L1 may alter the function of a cancer genome (Hancks & Kazazian 2012; Beck et al. 2011), including disruption and regulation of gene transcription, alternative splicing and exonization, promotion of chromosomal rearrangements, mobilization of pseudogenes in trans, and transduction of adjacent (usually downstream) genomic regions.

56% (12,487/22,383) of somatic retrotranspositions analysed in this study fall within gene boundaries. Nonetheless, insertional mutagenesis may have low effect on cancer gene expression, because the vast majority of these events are frequently located in late-replication regions of the genome (Helman et al. 2014), most likely heterochromatic regions (Tubio et al. 2014).

L1 transductions have a strong potential to impact cancer genome function, given by their ability to duplicate and spread exons, genes, and regulatory regions (Moran et

al. 1999; Tubio et al, 2014). For example, in our cancer dataset, a novel source L1 element reported in this study, which is located at locus 14q24.2, promotes the transduction of complete and partial coding regions of the Mitogen activated protein kinase *MAP3K9*. To understand how the duplication of this, and other, coding and regulatory regions mediated by L1 transductions can impact cancer genome function, we will need to explore the RNA sequencing data and gene methylation profile arrays available from Pan-cancer.

7. Concluding remarks

About half of human tumours have a variable number of somatic retrotranspositions, ranging from less than ten to several hundreds, which are caused by a relatively low number of L1 source elements. These germline L1 active elements are source for *de novo* somatic mutations that may help cancer to survive and progress.

In this study, we aimed to catalogue the full set of L1 germline elements with somatic activity in cancer, by analyzing somatic retrotransposition events on the largest dataset ever generated in a single project for the study of cancer retrotransposition, which consists of 2,704 cancer samples from 39 different tumour types.

We found that L1 elements dominate somatic retrotransposition across the cancers analyzed, representing ~87% of the total retrotransposition events. Overall, ~46% of the cancer genomes analyzed have a least one L1 somatic retrotransposition event with somatic retrotransposition being more frequent in lung squamous carcinoma, although esophageal carcinoma has the highest retrotransposition rate.

We confirmed that L1 3'-transductions represent a substantial contribution to the mutational landscape of cancer genomes, being 20% of the total L1 somatic events. We used these transductions to identify the L1 source elements whence they derive, and found 37 novel source elements that were not reported in previous cancer retrotransposition studies, including a novel hot-L1 source element at chromosome 7. Added to the 72 previously reported L1 source loci, it makes a total of 109 of germline L1 source loci with somatic retrotransposition activity in the cancer genome.

8. References

Beck, C.R. et al., 2011. LINE-1 Elements in Structural Variation and Disease. *Annu Rev Genomics Hum Genet*, 12(60), pp.187–215.

Beck, C.R. et al., 2010. LINE-1 retrotransposition activity in human genomes. *Cell*, 141(7), pp.1159–1170. Available at: http://dx.doi.org/10.1016/j.cell.2010.05.021.

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

Boissinot, S. et al., 2004. The insertional history of an active family of L1 retrotransposons in humans. *Genome Research*, 14(7), pp.1221–1231.

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. Available at: http://www.pnas.org/content/100/9/5280.short\nhttp://www.pnas.org/content/100/9/5280.abstract.

Calin, G.A. et al., 2002. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24), pp.15524–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=137750&tool=pmcentrez&rendertype=abstract.

Campbell, P.J. et al., 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6), pp.722–729. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18438408\nhttp://www.nature.com/ng/journal/v40/n6/pdf/ng.128.pdf.

Campbell, P.J. et al., 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319), pp.1109–13. Available at: http://dx.doi.org/10.1038/nature09460.

Cordaux, R. & Batzer, M.A., 2009. The impact of retrotransposons on human genome evolution. *Nature reviews. Genetics*, 10(10), pp.691–703. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2884099&tool=pmcentrez&rendertype=abstract.

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

Fabbri, M. et al., 2005. miR-15 and miR-16 induce apoptosis by targeting BCL2. *PNAS*, 102(39), pp.13944–13949.

Feuk, L., Carson, A. & Scherer, S., 2006. Structural variation in the human genome. *Nat Rev Genet*, 7(2), pp.85–97.

Goodier, J.L., 2014. Retrotransposition in tumors and brains. *Mobile DNA*, 5(1), p.11. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3995500&tool=pmcentrez&rendertype=abstract.

Griffiths, A.J. et al., 2000. Somatic versus germinal mutation. Available at: http://www.ncbi.nlm.nih.gov/books/NBK21894/ [Accessed June 2, 2016].

Hancks, D.C. & Kazazian, H.H., 2012. Active human retrotransposons: Variation and disease. *Current Opinion in Genetics and Development*, 22(3), pp.191–203. Available at: http://dx.doi.org/10.1016/j.gde.2012.02.006.

Hastings, P.J. et al., 2009. Mechanisms of change in gene copy number. *Nature reviews. Genetics*, 10(8), pp.551–64. Available at: http://www.nature.com/nrg/journal/v10/n8/pdf/nrg2593.pdf.

Hattori, M., 2005. Finishing the euchromatic sequence of the human genome. *Tanpakushitsu kakusan koso. Protein, nucleic acid, enzyme*, 50(2), pp.162–168.

Helman, E. et al., 2014. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Research*, 24(7), pp.1053–1063.

Huang, C.R.L., Burns, K.H. & Boeke, J.D., 2012. Active transposition in genomes. *Annual review of genetics*, 46, pp.651–75. Available at: http://www.annualreviews.org/doi/abs/10.1146/annurev-genet-110711-155616.

International Cancer Genome Consortium, 2015. ICGC Data Coordination Center. Available at: http://docs.icgc.org/submission/projects/.

Jennes, I. et al., 2011. Breakpoint characterization of large deletions in EXT1 or EXT2 in 10 multiple osteochondromas families. *BMC medical genetics*, 12, p.85. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3152881&tool=pmcentrez&rendertype=abstract.

Kazantseva, A. et al., 2006. Human hair growth deficiency is linked to a genetic defect in the phospholipase gene LIPH. *Science (New York, N.Y.)*, 314(5801), pp.982–985. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17095700.

Kazazian, H.H., 2004. Mobile elements: drivers of genome evolution. *Science (New York, N.Y.)*, 303(5664), pp.1626–32. Available at: http://www.sciencemag.org/content/303/5664/1626.full.pdf\nhttp://www.ncbi.nlm.nih.gov/pubmed/15016989.

Kazazian, H.H. & Moran, J. V., 1998. The impact of L1 retrotransposons on the human genome. *Nature Genetics*, 18(3), pp.231–236.

Koito, A. & Ikeda, T., 2013. Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. *Frontiers in Microbiology*, 4(FEB), pp.1–9.

Koning, A.P.J. et al., 2011. Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*, 7(12).

Korbel, J.O. et al., 2007. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *October*, 318(5849), pp.420–426.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11237011\nhttp://www.nature.com/nature/journal/v409/n6822/pdf/409860a0.pdf.

Lee, E. et al., 2013. Landscape of Somatic Retrotransposition in Human Cancers. *NIH Public Access*, 337(6097), pp.967–971.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

Lim, J.K. & Simmons, M.J., 1994. Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 16(4), pp.269–275.

Lisch, D., 2012. How important are transposons for plant evolution? *Nat Rev Genet*, 14(1), pp.49–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23247435\nhttp://www.nature.com/doifinder/10.1038/nrg3374.

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.

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.

O'Donnell, K.A. & Burns, K.H., 2010. Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mobile DNA*, 1(1), p.21. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2941744&tool=pmcentrez&rendertype=abstract.

Ostertag, E.M. & Kazazian, H.H., 2001. Biology of Mammalian L1 Retrotransposons. *Annual Review of Genetics*, 35, pp.501–538. Available at: http://www.annualreviews.org/doi/pdf/10.1146/annurev.genet.35.102401.091032

Reich, D.E. et al., 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature genetics*, 32(1), pp.135–142.

Ren, R., 2005. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nature reviews. Cancer*, 5(3), pp.172–183.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature Biotechnology*, 29(1), pp.24–26. Available at: http://www.nature.com/nbt/journal/v29/n1/abs/nbt.1754.html\nhttp://www.nature.com/nbt/journal/v29/n1/pdf/nbt.1754.pdf.

Sassaman, D.M. et al., 1997. Many human L1 elements are capable of retrotransposition. *Nature Genetics*, 15, pp.57–61.

Schneuwly, S., Kuroiwa, A. & Gehring, W.J., 1987. Molecular analysis of the dominant homeotic Antennapedia phenotype. *The EMBO journal*, 6(1), pp.201–206.

Seluanov, A., Van Meter, M. & Gorbunova, V., 2015. Wrangling Retrotransposons. *The Scientist Magazine*. Available at: http://www.the-scientist.com/?articles.view/articleNo/42274/title/Wrangling-Retrotransposons/.

Solyom, S. et al., 2015. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Research*, 22(12), pp.2328–2338.

Solyom, S. & Kazazian, H.H., 2012. Mobile elements in the human genome : implications for disease. *Genome Medicine*, 1, pp.1–8.

Stephens, P.J. et al., 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276), pp.1005–10. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3398135&tool=pmcentrez&rendertype=abstract.

Stephens, P.J. et al., 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1), pp.27–40. Available at: http://dx.doi.org/10.1016/j.cell.2010.11.055.

Stratton, M., Campbell, P. & Futreal, A., 2009. The cancer genome. *Nature*, 458(7239), pp.719–724. Available at: citeulike-article-id:4292667\nhttp://dx.doi.org/10.1038/nature07943.

Swergold, G.D., 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and cellular biology*, 10(12), pp.6718–29. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=362950&tool=pmcentrez&rendertype=abstract.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), pp.178–192.

Tubio, J.M.C. et al., 2014. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science (New York, N.Y.)*, 345(6196), p.1251343. Available at: http://science.sciencemag.org/content/345/6196/1251343.abstract.

Venter, J.C. et al., 2001. The sequence of the human genome. *Science*, 291(5507), pp.1304–1351.

Weinstein, J.N. et al., 2014. NIH Public Access. , 45(10), pp.1113–1120.

Wicker, T. et al., 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12), pp.973–982. Available at: http://www.nature.com/nrg/journal/v8/n12/abs/nrg2165.html.

Yang, L. et al., 2013. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153(4), pp.919–929. Available at: http://dx.doi.org/10.1016/j.cell.2013.04.010.

Appendices

APPENDIX A | ABBREVIATIONS

CNV	.....	Copy Number Variant
EN	.....	Endonuclease
HERVs	.....	Human endogenous retroviruses
ICGC	.....	International Cancer Genome Consortium
IGV	.....	Integrative Genomics Viewer
kb	.....	Kilobases
LINEs	.....	Long-interspersed elements
LIPH	.....	Lipase H
LTR	.....	Long Terminal Repeat
NGS	.....	Next-Generation Sequencing
ORF	.....	Open Reading Frame
PCAWG	.....	Pan-Cancer analysis of Whole Genomes
RNP	.....	Ribonucleoprotein
RT	.....	Reverse Transcriptase
SINEs	.....	Short- interspersed elements
SV	.....	Structural Variation
SVA	.....	SINE/VNTR/Alu
TCGA	.....	The Cancer Genome Atlas
TE	.....	Transposable Element
TPRT	.....	Target-primed reverse transcription
TraFiC	.....	Transposon Finder in Cancer
TS	.....	Target Site
TSD	.....	Target Site Duplication
UTR	.....	Untranslated Region

APPENDIX B | SUPPLEMENTARY TABLES

Table S1. Data from different projects all around the world were collected by ICGC and were used in the Pan-Cancer Project. The subset and samples of them used on this master thesis are the ones on this table (Adapted from International Cancer Genome Consortium 2015).

Code	Project Name	Number of Samples
BLCA-US	Bladder Urothelial Cancer - TGCA, US	23
BOCA-UK	Bone Cancer - UK	75
BRCA-EU	Breast ER+ and HER2- Cancer - EU/UK	215
BRCA-UK	Breast Triple Negative/Lobular Cancer - UK	
BRCA-US	Breast Cancer - TCGA, US	
BTCA-SG	Biliary tract cancer - Gall bladder cancer/Cholangiocarcinoma - SG	12
CESC-US	Cervical Squamous Cell Carcinoma - TCGA, US	20
CLLE-ES	Chronic Lymphocytic Leukemia - ES	100
CMDI-UK	Chronic Myeloid Disorders - UK	32
COAD-US	Colon Adenocarcinoma - TCGA, US	46
DLBC-US	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma - TCGA, US	7
EOPC-DE	Early Onset Prostate Cancer - DE	43
ESAD-UK	Esophageal Adenocarcinoma - UK	59
GACA-CN	Gastric Cancer - CN	32
GBM-US	Brain Glioblastoma Multiforme - TCGA, US	36
HNSC-US	Head and Neck Squamous Cell Carcinoma - TCGA, US	44
KICH-US	Kidney Chromophobe - TCGA, US	49
KIRC-US	Kidney Renal Clear Cell Carcinoma - TCGA, US	39
KIRP-US	Kidney Renal Papillary Cell Carcinoma - TCGA, US	34
LAML-KR	Acute Myeloid Leukemia - KR	43
49LAML-US	Acute Myeloid Leukemia - TCGA, US	
LGG-US	Brain Lower Grade Glioma - TCGA, US	19
LINC-JP	Liver Cancer - NCC, JP	16
LIRI-JP	Liver Cancer - RIKEN, JP	246
LUAD-US	Lung Adenocarcinoma - TCGA, US	37
LUSC-US	Lung Squamous Cell Carcinoma - TCGA, US	46
MALY-DE	Malignant Lymphoma - DE	100
MELA-AU	Skin Cancer - AU	66
ORCA-IN	Oral Cancer - IN	13
OV-AU	Ovarian Cancer - AU	116
PACA-AU	Pancreatic Cancer Endocrine Neoplasms- AU	243
PACA-CA	Pancreatic Cancer - CA	
PAEN-AU	Pancreatic Cancer Endocrine neoplasms - AU	104
PBCA-DE	Pediatric Brain Cancer - DE	249
PRAD-CA	Prostate Adenocarcinoma - CA	165
PRAD-UK	Prostate Adenocarcinoma - UK	
PRAD-US	Prostate Adenocarcinoma - TCGA, US	
READ-US	Rectum Adenocarcinoma - TCGA, US	16
RECA-EU	Renal Cell Cancer - EU/FR	92
SARC-US	Sarcoma - TCGA, US	32
SKCM-US	Skin Cutaneous melanoma - TCGA, US	38
STAD-US	Gastric Adenocarcinoma - TCGA, US	37
THCA-US	Head and Neck Thyroid Carcinoma - TCGA, US	49
UCEC-US	Uterine Corpus Endometrial Carcinoma – TCGA, US	51

Table S2. L1-source element results table showing the novel and old L1 source elements described in Tubío et al. (2014) and in this Master Thesis with their characteristics (chromosome, coordinates, strand, status).

Chromosome	Position1	Position2	Strand	Novel
X	11725369	11731399	plus	Known
X	11953208	11959433	minus	Known
X	105712516	105718549	plus	Known
X	140515278	140515267	minus	Known
1	72353887	72359918	plus	Known
1	84518060	84524089	minus	Known
1	87144764	87150794	minus	Known
1	114039843	114045879	plus	Known
1	119394975	119401003	plus	Known
1	121274033	121280059	minus	Known
1	165553149	165553134	minus	Known
1	196188501	196194532	minus	Known
2	23190961	23191000	minus	Known
2	36570274	36570269	minus	Known
2	135117676	135117685	plus	Known
2	156527841	156527820	plus	Known
3	20090525	20090512	minus	Known
3	27529295	27529293	minus	Known
3	53399324	53405352	minus	Known
3	55788573	55788577	plus	Known
3	89509975	89516006	minus	Known
3	97929285	97929325	plus	Known
3	123590727	123590709	plus	Known
3	151148535	151148548	minus	Known
4	23616395	23622409	plus	Known
4	80888061	80894087	plus	Known
4	112628956	112628982	plus	Known
4	152732645	152732711	minus	Known
5	39787755	39793773	plus	Known

Chromosome	Position1	Position2	Strand	Novel
5	80911910	80917937	plus	Known
5	89450783	89450771	plus	Known
5	109480275	109480254	minus	Known
5	112703080	112703070	plus	Known
6	13191015	13191013	plus	Known
6	19765124	19771149	plus	Known
6	24811907	24817711	minus	Known
6	29920101	29920277	plus	Known
6	51531356	51536769	minus	Known
6	72799421	72799524	plus	Known
6	123853932	123853976	minus	Known
6	145500621	145500846	minus	Known
6	157968403	157968568	minus	Known
7	30478858	30484890	plus	Known
7	53652494	53652523	minus	Known
7	65751841	65757871	minus	Known
7	141620482	141626512	minus	Known
8	57161904	57161887	plus	Known
8	73787792	73793823	minus	Known
8	126595131	126601133	minus	Known
8	135082987	135089016	minus	Known
9	85664455	85670486	minus	Known
9	111564969	111564954	plus	Known
9	115560408	115566439	minus	Known
10	19377514	19383547	minus	Known
10	111572184	111578215	minus	Known
11	14737455	14743484	plus	Known
11	87566584	87566645	minus	Known
12	3608362	3614394	minus	Known

Table S2 (cont.). L1-source element results table showing the novel and old L1 source elements described in Tubío et al. (2014) and in this Master Thesis with their characteristics (chromosome, coordinates, strand, status).

Chromosome	Position1	Position2	Strand	Novel
13	61221130	61221153	plus	Known
13	61462309	61462332	minus	Known
13	77186959	77192986	minus	Known
14	31150830	31150811	minus	Known
14	52667751	52667802	minus	Known
14	59220408	59220378	plus	Known
15	56251153	56251200	plus	Known
15	68688010	68688056	minus	Known
15	84052404	84058368	plus	Known
17	64637258	64637272	plus	Known
18	57071172	57077202	minus	Known
20	23406746	23412777	plus	Known
21	19082273	19082366	minus	Known
22	29059272	29065303	plus	Known
1	32004320	32004345	minus	Novel
1	102568874	102568951	plus	Novel
1	116980762	116980766	minus	Novel
1	162184318	162184334	minus	Novel
1	179575362	179575378	minus	Novel
1	222579104	222579158	minus	Novel
2	144010775	144010794	plus	Novel
2	191478700	191478715	minus	Novel
3	80590163	80590177	minus	Novel
3	101279644	101279658	plus	Novel
3	186372124	186372142	plus	Novel
4	10632679	10632695	minus	Novel
4	110248316	110248332	minus	Novel
5	16464296	16464310	plus	Novel
5	16883011	16883056	minus	Novel
5	137014774	137014791	minus	Novel
6	19793116	19793130	plus	Novel
6	32613066	32613367	plus	Novel
6	63368196	63368208	plus	Novel
6	102846086	102846095	minus	Novel
7	47864826	47864839	plus	Novel
9	5491407	5491419	plus	Novel
9	99364826	99364842	plus	Novel
10	19065785	19065799	minus	Novel
10	31517463	31517471	minus	Novel
10	36759271	36759287	plus	Novel
11	45157031	45157046	plus	Novel
12	33325794	33325804	minus	Novel
12	71020035	71020046	plus	Novel
12	96912555	96912570	minus	Novel
12	117814447	117814461	minus	Novel
14	24992912	24992926	minus	Novel
14	71197777	71197800	plus	Novel
15	85140894	85140910	plus	Novel
16	27425927	27425941	minus	Novel
18	73290637	73290645	minus	Novel
19	47024925	47024943	minus	Novel

APPENDIX C | SUPPLEMENTARY FIGURES

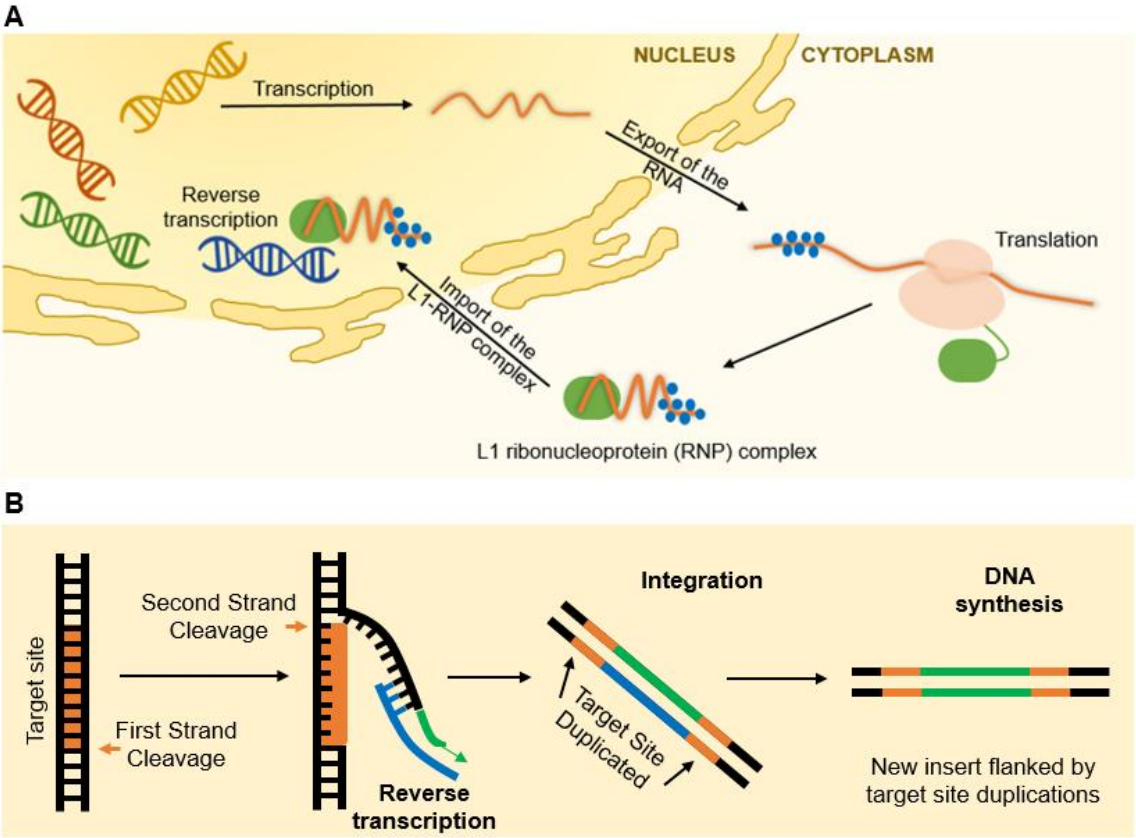


Figure. L1 retrotransposition cycle. **(A)** L1 mRNA is exported into the cytoplasm, translated, and L1-encoded proteins (L1 ORF1p, L1 ORF2p) bind to their own mRNA and form RNP complexes which are reimported into the nucleus (Adapted from Singer et al. 2010 and Ostertag & Kazazian 2001). **(B)** Subsequently, L1 RNA is reverse transcribed and the cDNA is inserted into the genome by a mechanism termed target-primed reverse transcription (TPRT) (Adapted from Ostertag & Kazazian 2001).

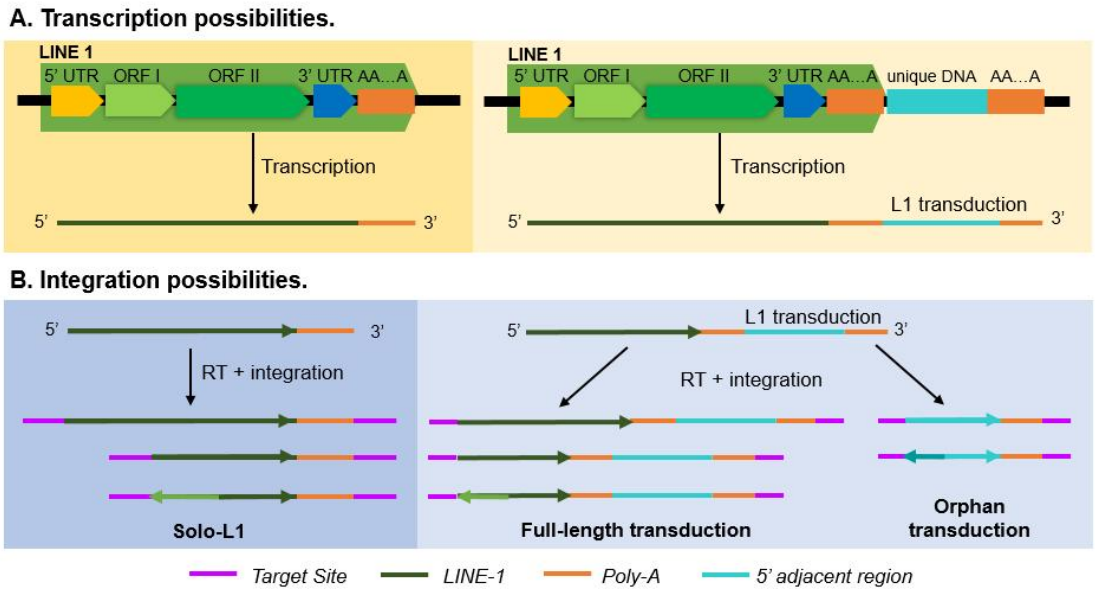


Figure. LINE-1 retrotransposition (Adapted from *Tubio et al., 2014*). **(A)** Transcription possibilities of a typical full-length human L1 element. **(B)** Integration possibilities of the two transcripts possibilities.

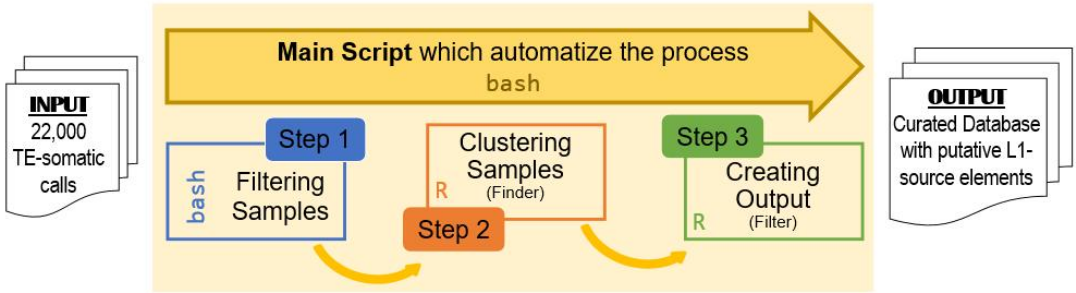


Figure. Scripting pipeline developed for the curation of data in this Master Thesis.

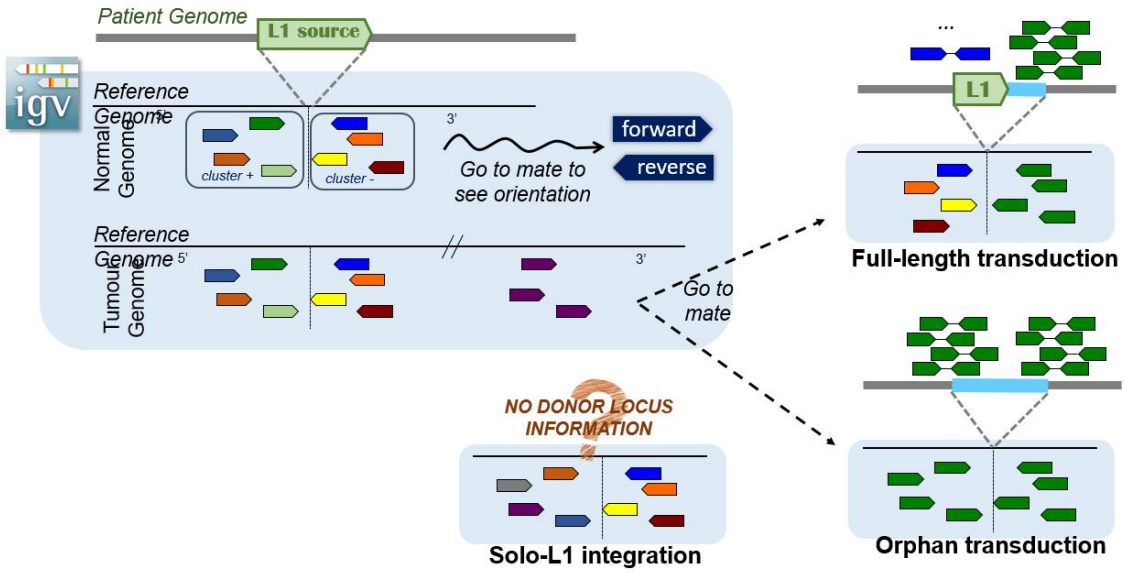
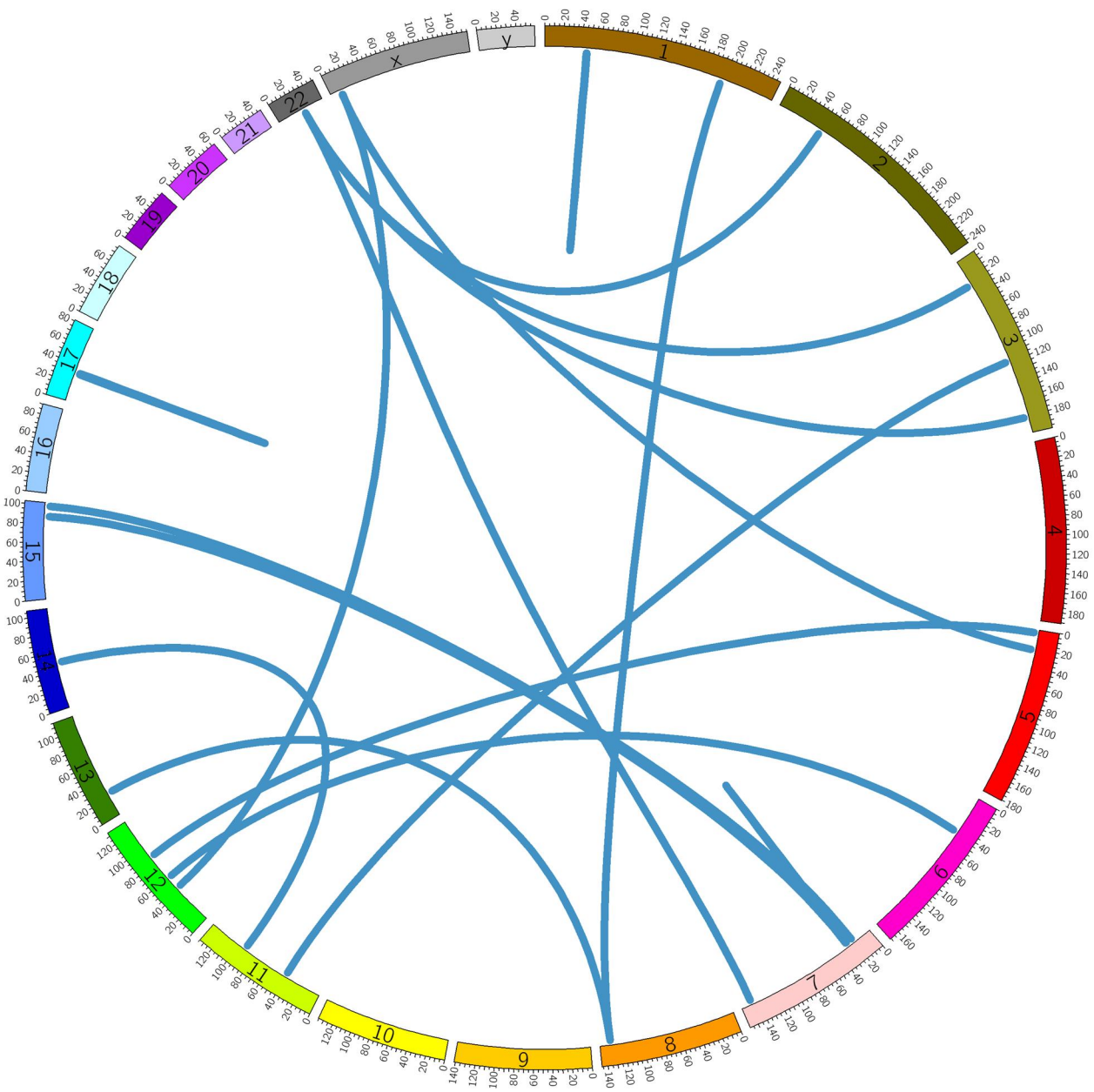


Figure. L1 identification approach using IGV. Load normal genome and note the orientation of the L1 source element regarding the mapping of the reads from each cluster. Load, as well, tumour genome and look for transduction after the insertion coordinates of the L1 source element. Taking a look to the mates of those reads, two types of transductions – full-length and orphan transductions – can be detected.

APPENDIX D | LIST OF FIGURES


y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X