INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Sonia
Zumalave Duro

PhD Thesis

THE DYNAMICS OF SOMATIC RETROTRANSPOSITION IN HUMAN CANCER

Santiago de Compostela, 2024

Doctoral Programme in Molecular Medicine

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

DOCTORAL THESIS

THE DYNAMICS OF SOMATIC RETROTRANSPOSITION IN HUMAN CANCER

Author

Sonia Zumalave Duro

Director: José Manuel Castro Tubío

Tutor: José Manuel Castro Tubío

PHD PROGRAM IN MOLECULAR MEDICINE

SANTIAGO DE COMPOSTELA

Á miña familia,

1 TABLE OF CONTENTS

2 AUTHOR'S DECLARATION

A autora e o director deste traballo acordaron libremente presentar os resultados desta tese de doutoramento, e decláranse sen ningún conflito de intereses en relación á mesma.

A autora declara que todas as figuras usadas no manuscrito foron creadas por ela mesma.

A autora declara que todas as mostras biolóxicas deste proxecto foron conseguidas a través da Rede de Biobancos Nacionais e contan co permiso do Comité Ético de Investigación Clínica (CEIC) pertinente (Anexo "13.3 Ethics Committee Approval"). Outros datos xenómicos utilizados neste estudo son anónimos e públicos, polo que non se identifica ningún tipo de implicación ética ou legal.

3 FUNDING

O contrato da autora desta tese foi financiado pola Xunta de Galicia, no marco do seu programa de contratos predoutorais, na convocatoria de 2018 (Referencia: ED481A 2018/199). Ademais, a presente tese foi sufragada a través dos seguintes proxectos de investigación:

- PGC2018-102245-B-I00 "The impact of L1-mediated somatic rearrangements in the origin and development of human cancer (L1-ARCHITECT)". State Research Agency. Proyectos de I+D de Generación de Conocimiento 2018. 01/01/2019-31/12/2021 (University of Santiago de Compostela). 350,900.00€. PI: JMC. Tubío.

- LABAE20053TUBI "The functional impact of retrotransposon- mediated chromatin interactions in the 3D cancer genome". Fundación Científica de la Asociación Española Contra el Cáncer (AECC). Lab AECC 2020. 01/12/2020-30/11/2023 (University of Santiago de Compostela). 300,000.00€. PI: JMC. Tubío.

4 ACKNOWLEDGEMENTS

To everyone that has somehow contributed to this PhD project, whether through insightful discussions, invaluable support or collaborative efforts, I express my deepest gratitude. Your contributions have played a vital role in shaping the success and progress of this research. I am truly grateful for your dedication and support. In particular, I would like to express a special thank you to my PhD supervisor.

5 SUMMARY

5.1 ENGLISH

Title: The dynamics of somatic retrotransposition in human cancer

Somatic retrotransposition, a mutational process where retrotransposable elements are copied and inserted into new genomic sites within somatic cells, is implicated in the initiation and progression of certain human tumours. In a previous analysis that included 2,954 cancer genomes, we detected high rates of somatic retrotransposition in various cancer types, including oesophageal adenocarcinoma (ESAD), head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC) and colorectal adenocarcinomas (COAD). Notably, these four cancer types collectively represented 70% of all somatic retrotranspositions, despite accounting for merely 9% of the samples. However, some relevant aspects of this mutational process remained uncharacterized in previous studies due to technological limitations of Illumina short-read sequencing. This doctoral project was conceived to explore the potential of long-read sequencing technologies and develop specific computational tools to overcome these limitations. First, we developed MEIGA, a novel computational method specifically designed to detect somatic retrotransposition events in long-read sequencing data in the context of cancer. MEIGA has proven to outperform current bioinformatics tools, being the most robust and accurate tool in detecting and characterizing somatic retrotransposition events.

To comprehensively characterize the dynamics of somatic retrotransposition in human cancer, we conducted a prospective screening using shallow sequencing in 150 primary tumours to identify samples with high retrotransposition rates, selecting 10 tumours each with over 100 somatic retrotransposition events. This includes five HNSC, four LUSC and a COAD, which were subsequently sequenced using Oxford Nanopore long-reads technologies. Then, we ran our computational approach, MEIGA, on these samples and their adjacent matched-normal tissues, finding 6,266 somatic insertions of retrotransposons and 152 genomic rearrangements mediated by retrotransposition. Notably, these retrotransposition-mediated rearrangements encompassed multiple large-scale reciprocal rearrangements, such as translocations and inversions, that remained largely undetected by previous cancer retrotransposition studies. We developed computational approaches to time these retrotransposon events to unprecedented

resolution, which showed that retrotransposition is an early mutational process active throughout tumorigenesis, even years before tumour diagnosis. In addition, we developed a bioinformatics pipeline to study somatic retrotransposition in healthy tissues isolated by laser-capture microdissection. This analysis confirmed not only that somatic retrotransposition is also active in multiple healthy tissues but also showed variable activity levels across different tissue types. Overall, our findings indicate that somatic retrotransposition is an early mutational process that, after malignant transformation, remains active along primary tumour evolution, providing cancer cells with a wide spectrum of mutations required for their survival and proliferation. Our study adds substantial knowledge for the understanding of the dynamics of somatic retrotransposition in cancer, providing valuable insights for future research.

Keywords: cancer, tumour evolution, somatic mutation, structural variants, retrotransposition.

5.2 GALEGO

Título: A dinámica da retrotransposición somática no cancro humano

A retrotransposición somática, un proceso mutacional no que os elementos retrotranspoñibles son copiados e inseridos en novos sitios do xenoma dentro de células somáticas, está implicado na iniciación e progresión de certos tumores humanos. Nunha análise previa que incluía 2,954 xenomas de cancro, detectamos altas taxas de retrotransposición somática en varios tipos de cancros. Estes incluíron o adenocarcinoma de esófago (ESAD), o carcinoma escamoso de cabeza e pescozo (HNSC), o carcinoma escamoso de pulmón (LUSC) e o adenocarcinoma colorectal (COAD). De maneira notábel, estes catro tipos de cancros representaron o 70% de todas as retrotransposicións somáticas, aínda que só constituían o 9% das mostras. Con todo, algúns aspectos relevantes deste mecanismo mutacional quedaron sen caracterizar debido ás limitacións tecnolóxicas da secuenciación de lectura curta de Illumina. Esta tese doutoral foi concibida para explorar o potencial das tecnoloxías de secuenciación de lectura longa e desenvolver ferramentas computacionais específicas para superar estas limitacións. Primeiro, desenvolvemos MEIGA, un novo método computacional especificamente deseñado para detectar eventos de retrotransposición somática en datos de secuenciación de lectura longa no contexto do cancro. MEIGA demostrou superar ás ferramentas bioinformáticas actuais, sendo a ferramenta máis robusta e precisa na detección e caracterización de eventos de retrotransposición.

Para unha caracterización exhaustiva da dinámica da retrotransposición somática no cancro humano, efectuamos un cribado prospectivo con secuenciación a baixa cobertura en 150 tumores primarios para identificar mostras con altas taxas de retrotransposición. Identificamos un subconxunto de 10 tumores con máis de 100 retrotransposicións somáticas, incluíndo cinco HNSC, catro LUSC e un COAD. Estes 10 tumores e os seus tecidos non tumorais adxacentes foron secuenciados empregando a tecnoloxía de lecturas longas de Oxford Nanopore. Analizamos con MEIGA os datos resultantes, identificando un total de 6,266 insercións e 152 reordenamentos cromosómicos orixinados pola actividade somática dos retrotransposóns. Especificamente, entre estes reordenamentos mediados por retrotransposóns, descubrimos unha paisaxe oculta de reordenamentos recíprocos de gran escala, incluíndo translocacións e inversións, non descritos en estudos previos. Ademais, desenvolvemos enfoques computacionais para datar estes eventos de retrotransposición dentro da historia evolutiva do

tumor cunha resolución sen precedentes. Descubrimos que a retrotransposición é un proceso mutacional que se inicia nunha etapa temperá do desenvolvemento tumoral, incluso anos antes do diagnóstico do tumor, e continúa activo durante as etapas subsecuentes da tumorixénese. Por outra banda, desenvolvemos unha ferramenta bioinformática para estudar a retrotransposición somática en tecidos sans obtidos mediante microdisección láser. Dita análise non só confirmou a presenza de retrotransposición en tecidos non tumorais, senón que tamén puxo de manifesto unha actividade diferencial entre os distintos tipos de tecidos sans examinados. Globalmente, os resultados do noso estudo sinalan que a retrotransposición somática é un proceso mutacional temperán que, tras a transformación maligna, permanece activa ao longo da evolución do tumor primario, proporcionando ás células cancerosas cunha ampla gama de mutacións necesarias para a súa supervivencia e proliferación. Esta investigación aporta unha contribución significativa ao entendemento da dinámica da retrotransposición somática no cancro, proporcionando perspectivas valiosas para futuras investigacións.

Palabras chave: cancro, evolución tumoral, mutación somática, variantes estruturais, retrotransposición.

5.3 ESPAÑOL

Título: La dinámica de la retrotransposición somática en el cáncer humano

La retrotransposición somática, un proceso mutacional en el que los elementos retrotransponibles son copiados e insertados en nuevos sitios del genoma dentro de células somáticas, está implicada en la iniciación y progresión de ciertos tumores humanos. En un análisis previo que incluía 2,954 genomas de cáncer, detectamos altas tasas de retrotransposición somática en varios tipos de cánceres. Estos incluyeron el adenocarcinoma de esófago (ESAD), el carcinoma escamoso de cabeza y cuello (HNSC), el carcinoma escamoso de pulmón (LUSC) y el adenocarcinoma colorrectal (COAD). De manera notable, estos cuatro tipos de cánceres representaron el 70% de todas las retrotransposiciones somáticas, aunque solo constituían el 9% de las muestras. Sin embargo, algunos aspectos relevantes de este mecanismo mutacional quedaron sin caracterizar debido a las limitaciones de las tecnologías de secuenciación de lectura corta de Illumina. Esta tesis doctoral fue concebida para explorar el potencial de las tecnologías de secuenciación de lectura larga y desarrollar herramientas computacionales específicas para superar estas limitaciones. Primero, desarrollamos MEIGA, un nuevo método computacional específicamente diseñado para detectar eventos de retrotransposición somática en datos de secuenciación de lectura larga en el contexto del cáncer. MEIGA demostró superar a las herramientas bioinformáticas actuales, siendo la herramienta más robusta y precisa en la detección y caracterización de eventos de retrotransposición.

Para una caracterización exhaustiva de la dinámica de la retrotransposición somática en el cáncer humano, realizamos un cribado prospectivo con secuenciación de baja cobertura en 150 tumores primarios para identificar muestras con altas tasas de retrotransposición. Identificamos un subconjunto de 10 tumores con más de 100 retrotransposiciones somáticas, incluyendo cinco HNSC, cuatro LUSC y un COAD. Estos 10 tumores y sus tejidos no tumorales adyacentes fueron secuenciados empleando la tecnología de lecturas largas de Oxford Nanopore. Analizamos con MEIGA los datos resultantes, identificando un total de 6,266 inserciones y 152 reordenamientos cromosómicos originados por la actividad somática de los retrotransposones.

Entre estos reordenamientos mediados por retrotransposones, descubrimos un paisaje oculto de reordenamientos recíprocos de gran escala, incluyendo translocaciones e inversiones, que no habían sido descritos en estudios previos. Además, desarrollamos enfoques computacionales para datar estos eventos de retrotransposición dentro de la historia evolutiva del tumor con una resolución sin precedentes. Descubrimos que la retrotransposición es un proceso mutacional que se inicia en una etapa temprana del desarrollo tumoral, incluso años antes del diagnóstico del tumor, y continúa activo durante las etapas subsecuentes de la tumorigénesis. Por otra parte, desarrollamos una herramienta bioinformática para estudiar la retrotransposición somática en tejidos sanos obtenidos mediante microdisección láser. Dicho análisis no solo confirmó la presencia de retrotransposición en tejidos no tumorales, sino que también puso de manifiesto una actividad diferencial entre los distintos tipos de tejidos sanos examinados. Globalmente, los resultados de nuestro estudio señalan que la retrotransposición somática es un proceso mutacional temprano que, tras la transformación maligna, permanece activa a lo largo de la evolución del tumor primario, proporcionando a las células cancerosas una amplia gama de mutaciones necesarias para su supervivencia y proliferación. Esta investigación aporta una contribución significativa al entendimiento de la dinámica de la retrotransposición somática en el cáncer, proporcionando perspectivas valiosas para futuras investigaciones.

Palabras clave: cáncer, evolución tumoral, mutación somática, variantes estructurales, retrotransposici

6 LIST OF ABBREVIATIONS

BFB	Breakage-Fusion-Bridge
BND	Break-End
bp	Base Pairs
cDNA	Complementary DNA
CI	Confidence Interval
CNV	Copy Number Variant
COAD	Colorectal Adenocarcinoma
DEL	Deletion
DSB	Double-Strand Break
DUP	Duplication
EI	Endonuclease-Independent
ESAD	Oesophageal Adenocarcinoma
FDR	False Discovery Rate
FISH	Fluorescence In Situ Hybridization
FN	False Negative
FoSTeS	Fork Stalling and Template Switching
FP	False Positive
FPKM	Fragments Per Kilobase of sequence per Million mapped reads
HERV	Human Endogenous Retrovirus
HNSC	Head and Neck Squamous Cell Carcinoma
ICGC	International Cancer Genome Consortium
IGV	Integrative Genomics Viewer
INS	Insertion
INV	Inversion
kp	Kilobase Pairs
L1	LINE-1

LCM	Laser Capture Microdissection
LINE	Long Interspersed Nucleotide Element
LTR	Long Terminal Repeat
LUSC	Lung Squamous Cell Carcinoma
MAPQ	Mapping Quality
MMBIR	Microhomology-Mediated Break-Induced Replication
mRNA	Messenger RNA
NA	Not Assigned
NGS	Next Generation Sequencing
nRef-L1	Source L1 elements absent in the reference genome
O/E	Observed-To-Expected Length Ratio
ONT	Oxford Nanopore Technologies
ORF	Open Reading Frame
PA	Polyadenylation Signals
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCR	Polymerase Chain Reaction
PSD	Processed Pseudogene
RC-seq	Retrotransposon Capture Sequencing
Ref-L1	Source L1 elements present in the reference genome
RT	Retrotransposition
SINE	Short Interspersed Nucleotide Element
SKY	Spectral Karyotyping
SNV	Single Nucleotide Variant
SV	Structural Variant
SVA	SINE-VNTR-*Alu*
TCGA	The Cancer Genome Atlas
TE	Transposable Element
TP	True Positive
TPRT	Target-Primed Reverse Transcription
TRA	Translocation
TSD	Target Site Duplication
UNK	Unknown
UTR	Untranslated Region
VAF	Variant Allele Frequency
WCP	Whole Chromosome Painting
WGD	Whole-Genome Doubling
WGS	Whole-Genome Sequencing
xTea-LR	xTea Module for Long-Reads
xTea-SR	xTea Module for Short-Reads

INTRODUCTION

7 INTRODUCTION

7.1 TRANSPOSABLE ELEMENTS IN THE HUMAN GENOME

Transposable elements (TEs), often referred to as transposons, are mobile DNA sequences that possess the ability to relocate within their host genome. This process, known as transposition, enables them to shift from one genomic location to another[1]. However, some TEs have lost this mobility and now exist as mere remnants of their ancient activity within their host genomes. Furthermore, not all TEs possess the capability to operate autonomously in the process of transposition[2]. Certain TEs rely on the protein machinery encoded by other TEs to facilitate their movement. TEs are present in the genomes of most, if not all, the eukaryotic organisms studied to date, where they significantly influence host biology and drive genome evolution[3]. Notably, their frequency and impact vary significantly across different species[3].

For instance, in maize (*Zea mays*), where Barbara McClintock first discovered TEs in 1940[4], they constitute up to 85% of the total genome size[5]. This high prevalence is attributed to the presence of many different TE families and the dynamic nature of maize genomes. Conversely, in Arabidopsis (*Arabidopsis thaliana*), TEs constitute only about 10-20% of the genome[6,7], a relatively modest proportion when compared to many other plant species. Additionally, certain species, including specific bacteria and archaea, exhibit extremely low levels or an apparent absence of TEs[8–10]. These cases are often found in organisms with highly compact genomes, where the presence of TEs might be detrimental to their survival or function.

In humans, the initial sequencing of the human genome in 2001[11] revealed that TEs constitute about 45% of our genome (**Fig. 1**), and that, despite their relative abundance, the activity of TEs has notably declined in the hominid lineage since the mammalian radiation[11]. With recent advancements, including the first gapless assembly of the human genome in 2022[12], it has been recognized that up to 54% of our genome is composed of these elements. However, this estimate might still be conservative, as it seems that a significant portion of the remaining DNA consists of ancient TE copies that have diverged to such an extent that they are no longer identifiable as TEs.

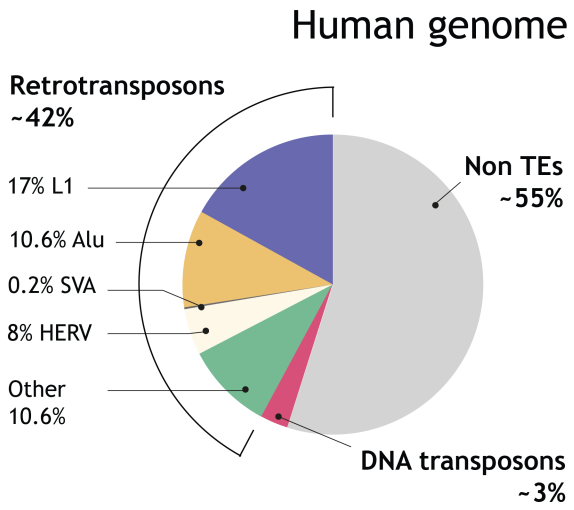


Figure 1. The transposable element content of the human genome. Over 45% of the human genome is identifiable as originating from TEs, with the predominant majority being non-LTR retrotransposons, particularly L1, *Alu* and SVA elements. HERV: Human Endogenous Retrovirus; L1: Long Interspersed Nucleotide Element 1; LTR: Long Terminal Repeat; SVA: SINE-VNTR-*Alu*; TE: Transposable Element.

TEs can be categorized into two major classes according to their mode of transposition: DNA transposons and Retrotransposons[1,13]. DNA transposons, or class II elements, move by a conservative cut-and-paste mechanism by which the element is simply excised from its original location and reinserted elsewhere into the genome. Retrotransposons, or class I elements, are TEs that transpose by a copy-and-paste mechanism, which involves the replication of the element into an RNA intermediate that is later reverse transcribed prior to insertion into the genome. Retrotransposons can be further subclassified into LTR retrotransposons, which are characterized by the presence of direct long terminal repeats (LTRs) at both extremes, and non-LTR retrotransposons, which do not have this feature.

DNA transposons are active in a diverse range of organisms[14]. However, in humans, these transposons have lost their ability to transpose within the genome[11,15–17]. They were intensively active during the early stages of primate evolution, but this activity ceased in an ancestor of anthropoid primates approximately 37 million years ago[16]. In modern humans, DNA transposons make up about 3% of the genome[11], representing fossilized remnants of once actively transposed elements. Notably, some DNA transposons have been repurposed over evolutionary time significantly influencing human biology[18,19]. A notable example is the origin of the recombination-activating genes, *RAG1* and *RAG2*[20,21]. These genes are thought to have evolved from an ancient DNA transposon that inhabited the genome of an ancestral eukaryote millions of years ago. Today, these genes play a critical role in the adaptive immunity of jawed vertebrates, enabling these organisms to develop highly sophisticated responses to a myriad of pathogens.

Retrotransposons exhibit extensive activity across eukaryotic lineages, with their activity levels and prevalence showing significant variation among different species[22]. For instance, LTR retrotransposons are particularly prevalent in plant species[23,24], whereas non-LTR retrotransposons are more commonly found in mammals compared to their LTR counterparts[11,17,25]. The transposition processes of LTR and non-LTR retrotransposons are remarkably different. LTR retrotransposons engage in a multistep process, wherein the reverse transcription is completed in the cytoplasm within virus-like particles, followed by the integration of retrotranscribed sequences into the host genome[26]. In contrast, for non-LTR retrotransposons, the transcripts of these elements are transported back into the nucleus, where reverse transcription and integration occur concurrently in a one-step process known as target-primed reverse transcription[27].

In humans, endogenous retroviruses (HERVs: Human Endogenous Retroviruses) represent a predominant type of LTR retrotransposons accounting for approximately 8% of our genome[11].

The vast majority of HERVs were inserted into the human genome over 25 million years ago, and are currently considered inactive[15,17] (**Fig. 2**). In contrast, the two major categories of mammalian non-LTR retrotransposons, long interspersed nucleotide elements (LINEs) and short interspersed nucleotide elements (SINEs), continue to be active in humans. Notably, LINE-1 (L1) represents the only autonomous transposon currently active in the human genome. SINEs, including *Alu* and SINE-VNTR-*Alu* (SVA) elements, while active, are nonautonomous and depend on L1 enzymatic activity for their propagation[28–30]. In the human population, new *Alu* insertions are estimated to occur at a rate of 2-5 new copies per 100 live births[31–33], while the insertion rates for L1 and SVA are 0.5-1 and ~0.1 new copies, respectively, for every 100 births[32–35]. At present, non-LTR retrotransposons are the predominant class of TEs in humans, constituting approximately one-third of our genome[11].

L1 elements have undergone continuous mobilization over the past 150 million years, resulting in over 850,000 copies within the human genome[11,13]. These L1 elements account for approximately 17% of our genome, making them the most abundant TE in terms of nucleotide content. Human L1 elements are characterized by a ~6 kb DNA sequence that includes a 5' untranslated region (UTR), followed by two open reading frames (ORFs), ORF1 and ORF2, and concluding with a 3' UTR holding a polyadenylation signal. While ORF1 encodes an RNA-binding protein, ORF2 encodes a protein with both endonuclease and reverse-transcriptase activities[36]. This molecular machinery enables L1 to not only to spread efficiently throughout the genome, but also to mobilize other TEs and cellular mRNAs, playing a pivotal role in the evolution of our genome.

Alu elements, which have been actively mobilized for approximately 65 million years, make up about 10% of the human genome[11,37]. They are the most abundant TEs in terms of copy number, with over 1.5 million copies of *Alu* elements present[37]. Typically spanning around 300 bp, these elements possess a dimeric structure, formed by the fusion of two monomers derived from the 7SL RNA gene, linked by an A-rich region[38]. In contrast, SVA elements, which have been active for approximately 25 million years throughout hominoid evolution, have generated around 3,000 copies, contributing up to 0.2% of the human genome[29,30]. SVA elements, though varying in length, typically average around 2 kb and consist of several distinct regions: a hexamer repeat region, an *Alu*-like region, a variable number of tandem repeats (VNTR) and a region of retroviral origin (SINE-R) [30].

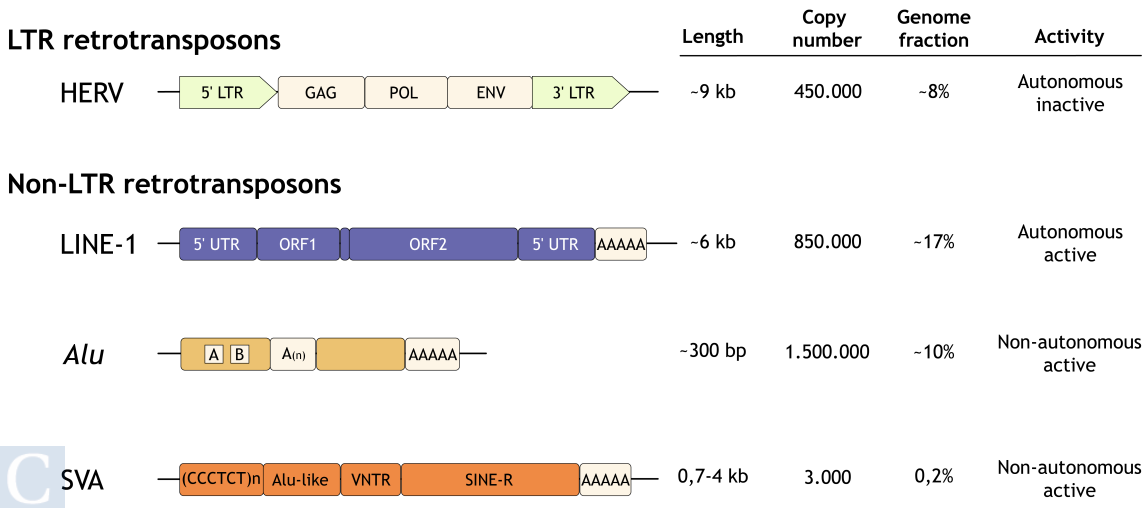


Figure 2. Description of the main human retrotransposons. (a) The canonical HERVs are

composed of viral genes, including *GAG*, *POL* and *ENV* genes, flanked by two LTRs. Notably, the LTRs encompass numerous regulatory elements, including promoter, enhancer, a primer-binding site for reverse transcription and a polyadenylation signal among others. **(b)** The canonical L1 element comprises two open reading frames, *ORF1* and *ORF2*, and is flanked by UTRs at both the 5' and 3' ends. This element concludes with an oligo dA-rich tail, following a polyadenylation signal. **(c)** The canonical *Alu* element is composed of two closely related monomers, separated by an adenine-rich linker, typically denoted as 'An'. The left monomer is distinguished by the presence of A and B boxes, which serve as promoters for transcription driven by RNA polymerase III. The structure of this element concludes with an oligo dA-rich tail. **(d)** The canonical SVA element exhibits a composite structure, characterized by several distinct regions. It begins with a CCCTCT hexamer repeat region, followed by an *Alu*-like region, a variable number of tandem repeats (VNTR) and a HERVK-like region, derived from the *ENV* gene and the 3' LTR of an HERV element. Similar to the others, this element ends with an oligo dA-rich tail and a polyadenylation signal. The lengths of the elements, their copy numbers in the genome, the proportion of the genome they occupy and their retrotransposition modes are all detailed. HERV: Human Endogenous Retrovirus; L1: Long Interspersed Nucleotide Element 1; LTR: Long Terminal Repeat; SVA: SINE-VNTR-*Alu*; UTR: Untranslated Region; VNTR: Variable Number of Tandem Repeats.

7.2 EXPLORING THE L1 RETROTRANSPOSITION CYCLE

The L1 retrotransposition cycle encompasses a series of intricate steps that rely on the coordinated interplay between L1 ORF proteins and various cellular enzymes[36]. The initial expression of L1 involves the activity of RNA polymerase II (Pol II)[39], which transcribes the L1 into a precursor RNA molecule. Following transcription, the L1 transcript is mobilized to the cytoplasm, where it undergoes translation by host cell machinery. Subsequently, L1 RNA along with the self-encoded ORF1 and ORF2 proteins assemble into ribonucleoprotein particles (RNPs) and are transported back to the nucleus. Once inside the nucleus, the process of reverse transcription takes place.

Several models have been proposed to explain the mechanism by which L1 inserts itself back into the host genome[40]. The one considered as the canonical model is referred to as Target-Primed Reverse Transcription (TPRT)[27]. The initiation of reverse transcription in the TPRT model is catalysed by the endonuclease activity of ORF2. This activity creates a single-strand break in the genomic DNA, specifically at thymidine-rich motifs like 5'-TTTT/A-3'. The cleavage exposes a poly(T) region at the intended insertion site, which anneals with the poly(A) tail of an L1 mRNA molecule. Following this, the reverse transcriptase activity of ORF2 initiates reverse transcription using the exposed 3'-OH group at the nick site as a primer. This enables the synthesis of the complementary DNA (cDNA) strand from the L1 RNA transcript. During or after this reverse transcription, a second nick is created on the opposite DNA strand, leading to the integration of the newly synthesized cDNA strand into the genome. Finally, the process concludes with the synthesis of the second DNA strand and the ligation of the 3′ ends of the inserted L1 sequence.

The retrotransposition process can result in either full-length L1 insertions or 5' truncated insertions. When the reverse transcription process successfully transcribes the entire L1 RNA sequence, the result is a full-length insertion. Conversely, if the reverse transcription complex detaches prematurely, this leads to an insertion that is truncated at the 5' end. Further research is required to develop a more detailed understanding of the specific mechanisms involved in

different stages of the retrotransposition process. This includes gaining insights into the nuclear import of RNP particles, the mechanisms underlying the creation of the second nick in the genomic DNA, the synthesis of the second DNA strand, and the complex interplay with cells DNA repair machinery.

An alternative form of TPRT, known as twin-priming[41], involves the simultaneous action of two reverse transcriptases at the integration site (**Fig. 3**). In this variant, the first nicked strand undergoes reverse transcription similarly to TPRT, while the free 3'-OH end, resulting from a second break on the opposite strand, primes reverse transcription in the opposite direction. Eventually, the retrotranscribed cDNAs from both strands engage in complementary pairing, displacing the L1 RNA and facilitating the synthesis of the second DNA strand from both ends. This twin-priming model typically leads to inversions within the newly formed L1 insertion. Notably, insertions mediated by TPRT, encompassing both the canonical and twin-priming models, are distinguished by specific structural hallmarks. These typically include a characteristic 3′ poly(A) tail and target site duplications (TSDs) of variable length. TSDs are duplications of short genomic DNA sequences flanking both sides of the new insertion, arising as a by-product of the integration process. Occasionally, TPRT can result not in TSDs but in small deletions of the target site DNA.

Besides TPRT and its variants, there is an alternative mechanism known as endonuclease-independent (EI) retrotransposition[42]. In EI retrotransposition, the process of reverse transcription is initiated at pre-existing DNA lesions, eliminating the requirement for the ORF2 endonuclease cleavage. As a result, in contrast to TPRT-mediated insertions, those driven by EI retrotransposition usually lack structural features such as 3′ poly(A) tails and TSDs. The existence of additional mechanisms like EI expands the range of pathways through which L1

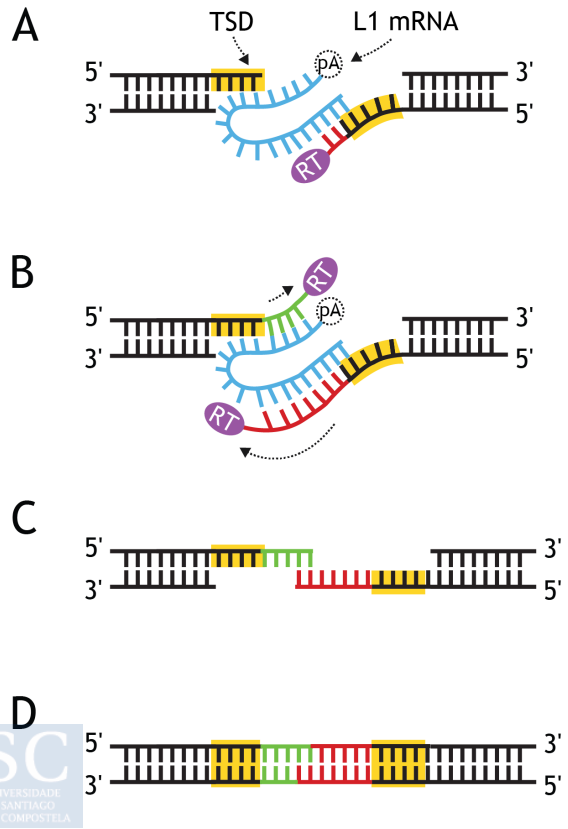


Figure 3. Twin-priming retrotransposition mechanism. The endonuclease activity of *ORF2* creates a single-strand break in the genomic DNA at thymidine-rich motifs. The cleavage exposes a poly(T) region which anneals with the poly(A) tail of an L1 mRNA molecule. **(a)** Following this, similarly to TPRT, the RT activity of *ORF2* initiates reverse transcription using the exposed 3'-OH group at the nick site as a primer. **(b)** The free 3'-OH end, resulting from a second break on the opposite strand, invades the L1 mRNA and primes reverse transcription in the opposite direction. **(c)** Eventually, the retrotranscribed cDNAs from both strands engage in complementary pairing, displacing the L1 mRNA and **(d)** facilitating the synthesis of the second DNA strand from both ends. This mechanism typically leads to inversions within the newly formed L1 insertion. cDNA: Complementary DNA; L1: Long Interspersed Nucleotide Element 1; mRNA: Messenger DNA; ORF: Open Reading Frame; pA: poly(A) tail; RT: Reverse Transcriptase; TPRT: Target-Primed Reverse Transcription; TSD: Target Site Duplication.

elements can integrate into the host genome. This diversity of mechanisms not only adds layers of complexity to the retrotransposition process but also highlights the versatility and evolutionary importance of L1 elements. At present, L1 activity drives genomic diversity mediated by TEs in humans and plays an important role in shaping the evolutionary course of our species.

7.3 THE IMPACT OF RETROTRANSPOSITION ON GENOME FUNCTION

The retrotransposition of L1, *Alu* and SVA elements can significantly impact the structure and function of the human genome[13,43–45]. While most retrotransposon insertions are likely phenotypically silent, some can cause severe phenotypic consequences and diseases[43,44,46,47]. The initial discovery of an L1 insertion dates back to 1988[46]. At that time, Haig Kazazian and his team were examining a haemophilia A patient who had no known family history of the condition. Their research led to the identification of a novel exonic L1 insertion in the X-linked gene factor VIII[46]. Since this seminal discovery, numerous instances of human genetic disorders caused by de novo retrotransposon insertions have been documented[13,47–49]. To date, over 100 cases have been linked to heritable diseases, including haemophilia, beta-thalassemia, Duchenne muscular dystrophy, cystic fibrosis, Apert syndrome, neurofibromatosis and various cancer cases[13,47–49].

There are multiple ways by which retrotransposition events can alter genome function (**Fig. 4**). First, retrotransposons can integrate within exons, disrupting gene function[44,46,47,50]. Insertions occurring into introns can alter mRNA splicing in different ways[40,51,52], including the promotion of exon skipping, intron retention and exonization of the retrotransposon sequence. Additionally, intronic insertions can attenuate gene expression or cause premature termination of transcription[53,54]. The impact of such insertions on transcriptional activity varies, often depending on whether the affected gene is in a sense or antisense orientation relative to the retrotransposon insertion. Moreover, retrotransposon insertions can disturb gene regulation in adjacent areas by interfering with promoters, enhancers and silencers[55]. They can also suppress the expression of nearby genes by inducing DNA methylation in the surrounding region[56–58].

In addition to canonical insertions, retrotransposons can also facilitate the movement of adjacent genomic sequences, a type of insertion known as transductions[59]. This is particularly the case with L1 and SVA elements, which commonly engage in 3' transductions affecting their downstream flanking regions[30,60–62]. Through this mechanism, they can mobilize neighbouring regulatory regions, a process that can result in gene deregulation. In 3′ transductions, the RNA transcription machinery bypasses the weak polyadenylation signal of the retrotransposon, instead utilising an alternative polyadenylation signal found elsewhere in the downstream 3′ flanking region of the host genome. The resulting transcript, comprising both the retrotransposon and the additional genomic sequence, is then integrated into the genome via retrotransposition. There are two types of insertions based on the final composition of the inserted sequence. An insertion that includes both the retrotransposon and the unique flanking sequence is termed a 'partnered transduction'. Conversely, if 5' truncation leads to the complete removal of the retrotransposon sequence, leaving only the flanking sequence, it is classified as an 'orphan transduction'.

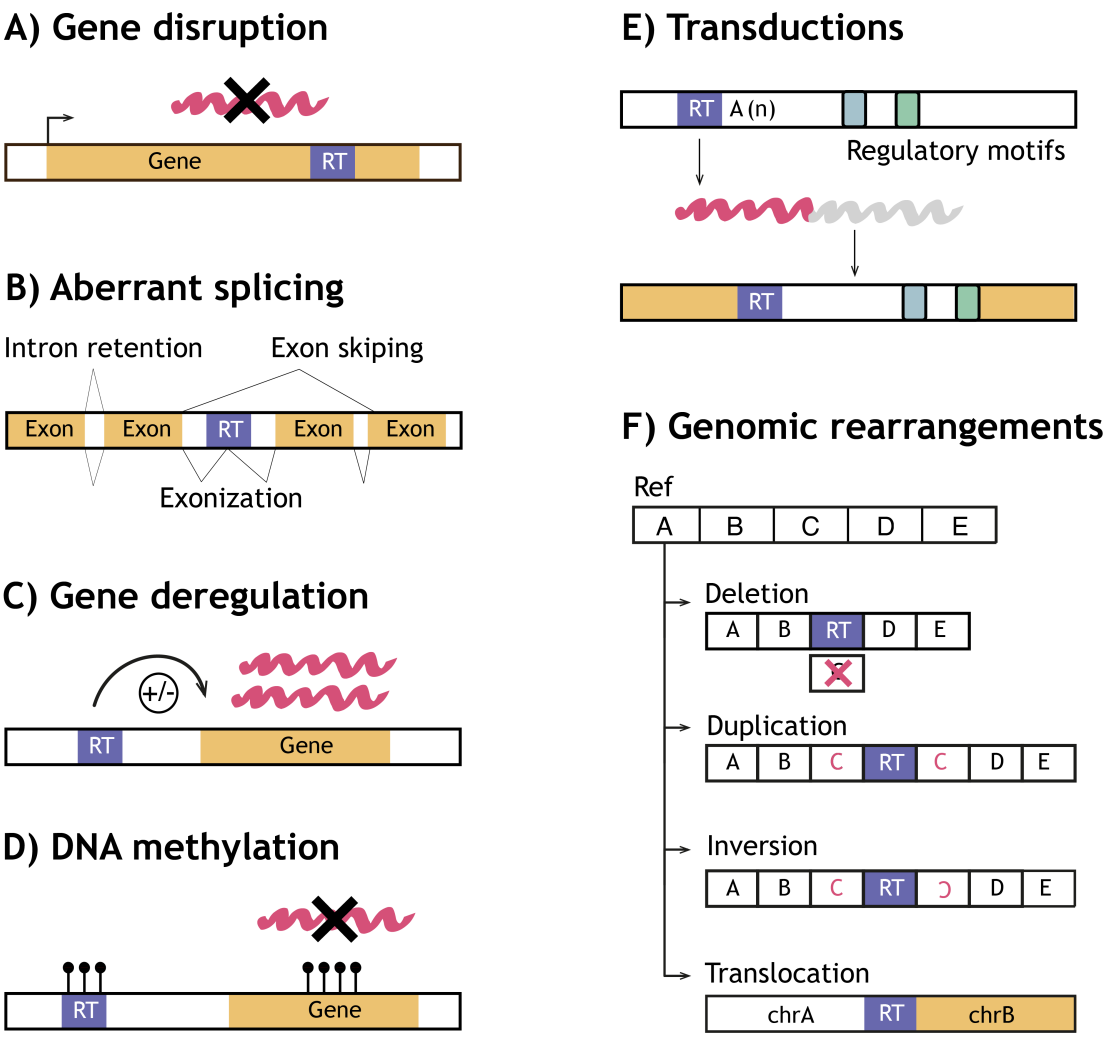


Figure 4: Mechanisms by which retrotransposition can impact human genome function. **(a)** Integration into exons disrupts gene function. **(b)** Integration into introns can lead to exon skipping, intron retention, and even exonization of the retrotransposon sequences themselves, resulting in aberrant splicing variants. **(c)** Integration near genes can deregulate them by attenuating gene expression, causing premature cessation of transcription, or interfering with promoters, enhancers and silencers. **(d)** Retrotransposon insertions can suppress the expression of adjacent genes by inducing DNA methylation in the surrounding regions. **(e)** Transductions can mobilize neighbouring regulatory regions, potentially leading to gene deregulation. **(f)** Retrotransposon insertions can facilitate genomic rearrangements, including deletions, duplications, inversions and translocations. These rearrangements can occur during the retrotransposition process or through recombination mechanisms post-integration. RT: Retrotranposition event.

Retrotransposition can also lead to a variety of genomic rearrangements during the insertion process[63–66]. These rearrangements include deletions, duplications, inversions and interchromosomal translocations[65]. Deletions and duplications arising from the retrotransposition process, while not frequent, have been extensively documented in the human genome[64–67], and can occasionally result in disease[67]. On the other hand, inversions and translocations caused by retrotransposition remain largely unexplored. So far, only a few instances of such events have been observed in cultured cells[65] and primary tumours[68]. Beyond

rearrangements directly caused by the retrotransposition process, retrotransposon insertions can also facilitate the formation of genomic rearrangements after integration[69–75]. This may occur through various recombination mechanisms, including single-strand annealing, synthesis-dependent strand annealing and possibly nonhomologous end joining. Collectively, these mechanisms highlight the profound and extensive impact that retrotransposition can have on the structure and function of the human genome.

7.4 SOMATIC RETROTRANSPOSITION AND ITS ROLE IN CANCER

Somatic retrotransposon insertions have been extensively identified in several cancers. Early cancer retrotransposition studies highlighted multiple instances of these insertions within cancer driver genes[76,77]. Notably, somatic insertions of L1 elements were found in an exon of the *APC* tumour-suppressor gene in a case of colorectal cancer[76] and within the *MYC* gene in a breast carcinoma[77]. Further experimental research expanded upon these initial findings, uncovering nine L1 insertions in six primary non-small cell lung tumours[78], 107 L1 insertions in 16 colorectal tumours[79] and a disruptive L1 insertion in the *ST18* gene in a hepatocellular carcinoma[80], all acquired somatically.

The introduction of next-generation sequencing in cancer research[81,82] enabled a more comprehensive investigation of somatic retrotransposition patterns. A remarkable study by Lee *et al.* in 2012 identified 194 somatic retrotransposon insertions across 43 tumour genomes spanning five distinct types of cancer[83]. Subsequent research expanded this understanding by examining 200 tumours across 11 different cancer types, revealing 810 somatic retrotransposon insertions[84]. This study highlighted significant variations in retrotransposition activity across different cancers. For instance, while tumour types such as lung squamous cell carcinomas and head and neck squamous cell carcinomas were characterized by a high number of somatic insertions, other tumour types displayed a notably lower frequency of these insertions.

To better understand to what extent somatic retrotransposition was relevant in cancer, we surveyed the patterns of these retrotranspositions on a considerably larger scale, across thousands of cancer genomes, and integrated them with other mutational processes and transcriptional data within the framework of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project[68], a joint initiative between the International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA).

In the PCAWG project, we developed novel strategies to analyse the patterns and mechanisms of somatic retrotransposition in 2,954 cancer genomes from 37 histological cancer subtypes[68]. This analysis revealed 19,166 somatic retrotransposon insertions which affected 35% of tumours. Additionally, we detected high rates of somatic retrotransposition in various cancer types, including oesophageal adenocarcinoma (ESAD), head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC) and colorectal adenocarcinomas (COAD). Notably, these four cancer types collectively represented 70% of all somatic retrotranspositions, despite accounting for merely 9% of the samples.

Beyond insertions, we identified 96 chromosomal rearrangements attributed to L1 retrotransposition within the PCAWG cohort[68]. These rearrangements comprised 90 deletions, along with one duplication, one translocation and one inversion. Additionally, there were three

cases where the precise nature of the rearrangement remained unresolved, a limitation stemming from the inherent short-read length of the Illumina sequencing libraries employed in this study.

Significantly, our analysis revealed that these rearrangements caused by the aberrant integration of L1 elements can result in the deletion of large, megabase-scale chromosomal regions, sometimes resulting in the loss of tumour suppressor genes such as *CDKN2A*[68]. Our findings further revealed that somatic retrotranspositions can trigger breakage-fusion-bridge cycles, potentially leading to the high-level amplification of oncogenes like *CCDN1*[68]. These observations shed light on the relevant role of L1 retrotransposition in remodelling the cancer genome, underscoring its potential impact on the origin and/or development of human tumours.

7.4.1 L1-mediated rearrangements in cancer

In our analysis of the PCAWG dataset, we identified a distinctive pattern in tumours with high rates of somatic retrotransposition[68]. We noticed that specific candidates for L1 retrotransposition were associated with a loss in copy number. This loss was uniquely defined by boundaries delineated by sequencing reads supporting the somatic integration of an L1 element. Such patterns suggested that the somatic integration of an L1 retrotransposon occurred simultaneously with the deletion event, which defines the rearrangement type called L1-mediated deletion. These deletions have previously been observed to occur somatically with engineered L1s in cultured human cells[64] and naturally in the brain[85], and are the consequence of an aberrant mechanism of L1 integration.

The importance of these findings[68] lay in demonstrating that L1-mediated deletions can naturally occur in cancer, and that they may promote extensive losses of DNA. These deletions are, on occasion, driver events, causing the loss of tumour suppressor genes. For example, in PCAWG we reported two different oesophageal adenocarcinoma tumours with large, clonal L1-mediated deletions —5.3 Mb and 8.6 Mb long— that removed a copy of the key tumour suppressor *CDKN2A*. The recurrence of this event in two independent tumours suggested that L1-mediated deletions are a mechanism that can target relevant genes in tumour evolution. Moreover, we also observed that L1-mediated deletions can affect areas essential for genome stability, such as centromeres and telomeres. For example, in a case of head-and-neck cancer, two separate L1 retrotransposon insertions led to the removal of both telomeres from a specific chromosome[68].

Notably, telomere loss promoted by L1 retrotransposition can trigger genomic instability through breakage-fusion-bridge (BFB) repair cycles and, on occasion, lead to the amplification of oncogenes[68]. This form of genetic instability typically begins with the end-to-end fusion of broken sister chromatids, forming a dicentric chromosome that results in an anaphase bridge during mitosis. While end-to-end chromosome fusions are classically attributed to telomere attrition[86], our findings suggested that somatic retrotransposition can also initiate this process. For instance, in our analysis of the PCAWG dataset, we identified three independent tumours —oesophageal, lung and breast— where large L1-mediated deletions on chromosome 11 removed the telomere and triggered BFB repair cycles that led to the amplification of the *CCND1* oncogene[68]. The independent recurrence of these patterns, all involving the

amplification of *CCND1* in different samples, underscores a L1 retrotransposition-mediated mutational mechanism that likely plays a significant role in the development of human cancer.

The analysis of the PCAWG dataset revealed that, apart from deletions, aberrant L1 retrotransposition can also promote other types of rearrangements in human tumours[68]. Though the specific patterns defining these rearrangements, as detected by Illumina paired-end short-read sequencing, differ from one to another, there is a common hallmark among them: a copy number change is always present, demarcated by sequencing reads that support the integration of an L1 element. By analysing the type of the copy number change (i.e., gain or loss) and the orientation of L1-specific reads in relation to the boundaries of the copy number variant, we were able to discern various chromosomal rearrangements mediated by L1, such as translocations, segmental duplications and fold-back inversions[68]. However, they were notably less prevalent than L1-mediated deletions and their relevance remains uncertain.

7.4.2 Hidden L1-mediated rearrangements in cancer

During the analysis of the PCAWG dataset[68], we found evidence suggesting that rearrangements mediated by somatic retrotransposition in human tumours are more frequent than we could unambiguously characterize in our work. Notably, the sequencing technology used in this study imposed considerable limitations to our analysis. The Illumina paired-end short-read sequencing libraries had short insert sizes of approximately 350 bp and read lengths of around 150 bp. This presented a significant technical challenge: the read pairs were too short to fully span and connect the sequences on both sides of rearrangements mediated by long L1 bridges. Although somatic L1 insertions are typically truncated at the 5' end, their median length for solitary events is around 1.2 kb and can extend up to around 6 kb. Moreover, L1 transductions can surpass this, exceeding 6 kb in length.

Due to the previously described technical constraints, the analysis of the sequencing data in certain PCAWG tumours revealed intriguing patterns that could not be resolved. These patterns were characterized by unresolved breakpoints that suggested the presence of hidden rearrangements mediated by L1 elements[68]. For instance, we came across patterns where paired-end reads provided evidence of an L1 integration associated with an extensive copy number change on a specific chromosome. However, in these instances, we were unable to identify the second breakpoint of the rearrangement on the same chromosome. This observation suggests that the second breakpoint of these rearrangements would be located on different chromosomes, and that this pattern could correspond to cryptic unbalanced translocations mediated by long L1 events.

The analysis of the same PCAWG samples mentioned above, allowed us to identify another non-canonical pattern where retrotransposons were involved[68]. Here, the independent clusters supporting the L1 events again have no reciprocal cluster supporting the second breakpoint but, contrary to the patterns described above, there were no associated copy number changes. We hypothesized that these patterns could represent genomic inversions or reciprocal translocations mediated by long L1 events, where there were no gains or losses of DNA.

Indeed, these specific rearrangements fall outside the detection limits of cancer retrotransposition studies conducted using Illumina short-read sequencing technologies[68,83,84]. This underscores the possibility that an unknown portion of retrotransposon-mediated

mutations remains undiscovered in cancer patients. Fortunately, long-read sequencing technologies, such as single-molecule sequencing with Oxford Nanopore Technologies (ONT) or PacBio, and Bionano Genome Mapping[87], offer the potential for a more comprehensive understanding of the extent and impact of L1-mediated rearrangements in human cancer. However, while these technologies have the capability to overcome the challenges posed by previous sequencing approaches, they have not yet been employed to investigate these types of rearrangements within a cohort of tumours.

7.5 NAVIGATING THE OBSTACLES IN DETECTING RETROTRANSPOSITION

The detection of retrotransposons has long posed significant challenges because of their repetitive nature, characterized by the presence of thousands of nearly identical copies dispersed throughout the genome. Before the availability of affordable whole-genome sequencing (WGS) technologies, the discovery and genotyping of retrotransposons primarily relied on targeted methods[35,78,88,89]. These methods typically involved enriching DNA fragments associated with retrotransposon insertions through PCR amplification, followed by gel-based techniques or targeted sequencing. These approaches, which were thoroughly reviewed by Xing *et al.*[90], were instrumental in providing initial insights into the presence and characteristics of retrotransposons within genomic sequences. However, their application was limited when it came to large-scale detection.

The turn of the millennium marked a significant milestone in DNA sequencing with the introduction of high-throughput methods, commonly referred to as next-generation sequencing (NGS). High-throughput methods paved the way for the first-ever genome-wide surveys in humans, ushering in a revolutionary era in the field of genomics. Over the following decade, DNA sequencing platforms experienced continuous improvements, facilitating their systematic and cost-effective application across various scientific fields. Consequently, the production of an enormous volume of genomic data demanded considerable research efforts within the scientific community to develop sensitive algorithms for detecting genetic variants and enabling comprehensive genomic analysis.

In current genomics, Illumina stands as the dominant NGS platform, generating the majority of available WGS data. Illumina datasets often comprise billions of short-read pairs, with each pair representing the ends of longer DNA fragments, and individual reads typically extending across 75, 100 or 150 bp. In this sequencing setup, detecting small mutations such as single-base substitutions as well as insertions and deletions shorter than the read length, can be achieved by accurately aligning the reads to a reference genome and examining variations in the aligned columns of bases. However, detecting longer structural variants (SVs) poses a greater challenge. Illumina short-read sequences typically do not cover the entire affected interval, making it difficult to directly identify SVs, including retrotransposon insertions. Moreover, the widespread distribution of retrotransposons across the genome adds complexity and poses challenges for their accurate identification. Furthermore, detecting retrotransposons necessitates an additional level of scrutiny beyond what is typically required for SV detection in order to understand their distinctive insertion mechanisms.

The computational methods devised for detecting retrotransposon insertions in short paired-end read data often depend on identifying clusters of read-pairs exhibiting discordant patterns.

Clusters that support potential insertion events are formed by read-pairs where one end aligns unambiguously at the insertion site, while the other end maps to multiple distant locations in the genome. The former, known as anchored reads, determine the genomic position of the insertion, while the latter, referred to as non-anchor reads, provide information about the identity of the inserted sequence.

When two distinct clusters of discordant read-pairs, supporting a retrotransposon insertion from the same family, are identified in close proximity with anchors aligning in forward and reverse orientations, computational methods classify it as a candidate retrotransposon insertion. Following this, methods typically proceed to search for split-reads, which are reads spanning the junction between the retrotransposon and the genomic sequence at the integration point. Split-read assemblies are often employed to accurately define several key characteristics of genuine retrotransposition insertions, including the presence of poly(A) tails, TSDs, and endonuclease cleavage sites.

Eventually, candidate retrotransposition events undergo a filtering process based on multiple criteria, including mapping quality, read support, consistency in read orientations and breakpoint positions, underlying genomic features and other relevant factors. Among the most popular retrotransposon insertion callers for short-read sequencing data are Tea[83], RetroSeq[84], Mobster[91], TraFiC[92] and MELT[93]. This topic has been extensively reviewed by Adam D. Ewing in [94]. Notably, callers like Tea, RetroSeq and TraFiC distinguish between somatic and germline insertions by comparing tumour and matched-normal samples.

Despite significant advancements, Illumina short-read sequencing studies have been limited in their ability to comprehensively capture the complete repertoire of retrotransposons and their activity within our genome due to constraints imposed by its read size. Current long-read sequencing technologies, such as ONT and PacBio, provide a more comprehensive and detailed view of the genome, allowing for the detection of structural variations, accurate assembly of complex genomes, characterization of transposable elements, haplotype phasing and investigation of epigenetic modifications.

Over the past four years, at least four tools have been developed to detect retrotransposition events in long-read sequencing data, including rMETL[95], PALMER[96], TLDR[97] and xTea[98]. The detection of retrotransposon insertions in long-read sequencing often relies on the clustering of two types of insertion-supporting reads: spanning reads, which contain the entire inserted sequence embedded within them, and clipping reads, where one segment aligns specifically to the adjacent regions of the insertion site, while the clipped segment corresponds to the inserted sequence. To determine the identity of the inserted sequences, retrotransposition callers typically employ a process of realignment, comparing the inserts to a comprehensive database of retrotransposon sequences. Additionally, this step can be utilized to gain insights into the retrotransposition hallmarks that characterize bona fide retrotransposition insertion events.

The tool rMETL[95] was published in 2019, but the study did not explicitly compare the sensitivity of long-read callers to short-read callers. Instead, the researchers used short-read callers as a ground truth for their benchmarking analysis. Additionally, the authors did not apply their method to address any particular biological inquiry or investigation. In 2020, Weichen Zhou *et al.* conducted a study introducing the PALMER[96] tool and employing it to analyse PacBio long-read sequencing data obtained from the NA12878 benchmark genome. Through this analysis, the researchers successfully identified 203 non-reference L1Hs insertions, with

90 of them being undetected by previous short-read sequencing methods. Notably, it was observed that approximately 81% (73 out of 90) of these L1Hs insertions were found within endogenous L1 sequences present in the reference genome, indicating nested insertions. These findings highlighted the limitations of standard short-read sequencing approaches in effectively detecting L1 insertions within individual genomes.

In the same year, Adam Ewing and colleagues introduced the Transposons from Long DNA Reads (TLDR) method[97] to identify retrotransposon insertions. While TLDR is a valuable approach, it has limitations in terms of its detection capacity. It relies on identifying insertions fully embedded within long reads as the initial step for generating clusters. The likelihood of finding these specific reads is influenced by two critical factors: the distribution of the library read length and the length of the insertion itself. Consequently, TLDR results can be significantly influenced by the intricate interplay between these factors. However, their primary focus was on exploring the locus-specific methylation landscape of retrotransposons. Indeed, they demonstrated how ONT long-read sequencing can simultaneously survey the epigenome patterns and detect somatic mobilization of retrotransposons in a hepatocellular carcinoma patient.

In 2021, the lab of Peter Park introduced a novel method for long reads called xTea[98], which they assert surpasses PALMER in terms of specificity and demonstrates slightly improved sensitivity. Their analysis revolved around long-read WGS, utilizing both PacBio and ONT technologies, from 20 individuals representing diverse human populations. These datasets were sourced from two distinct studies, enabling comprehensive exploration and evaluation of the xTea method. On average, they identified 217 polymorphic L1 insertions per individual. The utilization of long reads also enabled them to resolve the structural features of these events. Among these polymorphic insertions, 59% were 5' truncated L1s, 21% full-length L1s, 17% L1s with internal inversions, 3.7% L1s with both internal deletions and inversions and 0.46% were L1s with only an internal deletion. Notably, the study yielded intriguing insights into the integration of TEs within centromeric regions. Specifically, they identified up to 114 potential full-length L1s within the centromeres of the gapless assembly CHM13.

While current methods have emphasized the significant advantages of long-read sequencing in detecting retrotransposon polymorphisms, our understanding of somatic retrotransposon activity in the context of long reads remains largely unknown. The existing methods were not specifically designed for paired analysis and have not been extensively evaluated in cohorts of tumours with high retrotransposition rates. Furthermore, most of these methods are not tailored to identify non-canonical integration patterns, such as rearrangements caused by the aberrant integration of retrotransposons. Consequently, our current understanding of the true impact of somatic retrotransposition on the cancer genome remains limited.

7.6 SOMATIC RETROTRANSPOSITION IN HEALTHY TISSUES

Somatic retrotransposition, long regarded as a mutational process exclusive to germ cells, was later observed in early embryonic development[99–101]. In 2007, a genetic study unveiled that a mutagenic L1 insertion, causing X-linked choroideremia in a male patient, occurred during the early embryonic development of his mother[99]. Simultaneously, another study demonstrated the ability of engineered L1s to retrotranspose in human embryonic stem cells[100]. Furthermore,

research involving transgenic mice and rats showed that retrotransposon expression often leads to retrotransposon insertions during early embryogenesis, resulting in somatic mosaicism[101]. Collectively, these findings indicated that the transcriptional activation of retrotransposons during mammalian embryogenesis[102,103,104,105] not only significantly impacts cellular physiology but can also contribute to genetic variation within individuals.

Beyond early embryogenesis, compelling evidence indicates that somatic retrotransposition continues during later developmental stages in the brain[56,85,106–110]. In 2005, a ground-breaking study by Muotri and colleagues demonstrated that engineered L1 reporter systems in transgenic mice could undergo retrotransposition, resulting in neuronal somatic mosaicism[110]. A subsequent study using PCR-based methods found higher L1 content in specific brain regions compared to heart and liver tissues of the same individuals, indicating somatic activity of L1 retrotransposition in the brain[56]. Furthering this line of research, a 2011 study using retrotransposon capture sequencing (RC-seq) demonstrated significant somatic retrotransposon activity in the hippocampus and caudate nucleus of three individuals[108]. Employing single-cell genomic sequencing, Evrony *et al.* initially estimated that there are, on average, 0.04 unique somatic insertions per neuron[106]. Following this, Upton and colleagues expanded on these findings with single-cell RC-seq analyses, estimating about 13.7 insertions per hippocampal neuron and 6.5 insertions per glial cell[107].

Despite these significant findings, there remains a notable gap in our understanding of the relevance and extent of somatic retrotransposition outside the brain, particularly in committed cell lineages other than neurons and glial cells. To bridge this research gap, several studies have employed targeted approaches to detect somatic retrotransposition events, initially identified in tumour tissues, within their adjacent healthy counterparts[80,111,112]. These studies indicated that L1 retrotransposition is active in the histologically healthy tissues, including stomach[111], oesophagus[112], liver[80] and colon[111], which then expanded clonally in the subsequent tumour development. Notably, these studies started to reveal evidence of somatic retrotransposition activity in healthy tissues outside the brain.

Recently, a ground-breaking study employing expanded colonies has thoroughly demonstrated that somatic retrotransposition is extensively active within the healthy colorectal epithelium[113]. Notably, this study not only confirmed the presence of somatic retrotransposition driving intraindividual genetic variation, but also revealed a remarkable correlation between somatic retrotransposition burden and age. Moreover, this study has also unveiled that the epigenetic activation of a specific source element is preferentially determined during the early stages of organogenesis. Additionally, they showed that source L1 elements with lower population allele frequencies, presumably the youngest ones, exhibit higher activity.

While there is a significant interest in systematically characterizing the somatic activity of retrotransposons in healthy tissues, identifying somatic mutations in these tissues presents a challenge due to their limited accessibility with standard sequencing protocols. Healthy tissues are comprised of a varied mix of cell types, each possibly harbouring unique mutations. Unlike tumours, healthy tissues do not experience recurrent clonal expansions that significantly amplify the presence of mutations. As a result, single mutations in healthy tissues are present in a much smaller proportion of cells, making them difficult to detect with standard sequencing methods that often lack the necessary sensitivity to identify such low-frequency mutations.

To overcome this challenge, several sequencing strategies have been explored in recent years. These included single-cell sequencing[106,107,114,115], sequencing of expanded colonies from single cells[113] and sequencing of small cell populations isolated from histological sections by laser capture microdissection (LCM)[116–119]. These advancements in technology and methodology have opened new and promising avenues in the study of somatic retrotransposition. Ongoing research in this field is expected to illuminate the intricate interplay between somatic retrotransposition and the maintenance of genomic integrity in healthy tissues. Notably, fully unravelling the extent and impact of somatic retrotransposition in these tissues will not only deepen our insight into the mutational dynamics of healthy cells but also enrich our understanding of physiological cellular processes.

OBJECTIVES

8 OBJECTIVES

This doctoral thesis delves into the complex dynamics of somatic retrotransposition in primary human cancers utilizing innovative long-read sequencing technologies. By exploring these dynamics, this thesis intends to gain a deeper understanding of the extent and impact of retrotransposition on cancer development and progression. This research is poised to contribute significantly to the field of cancer genomics and could potentially influence the prevention, diagnosis and treatment of tumours characterized by high rates of somatic retrotransposition. To achieve this, the specific objectives of the present thesis project are as follows:

i. **To develop a robust pipeline for the analysis of somatic retrotransposition using long-read sequencing data.** The aim of this pipeline is to thoroughly characterize somatic retrotransposition patterns, particularly those that are cancer-specific, using long-read sequencing datasets derived from tumour-normal pairs.

ii. **To conduct an exhaustive analysis of the patterns and frequencies of somatic retrotransposition in primary human cancers through long-read sequencing.** This study focuses on a diverse cohort of tumour samples, aiming to comprehensively characterize the retrotransposition landscape within these tumours.

iii. **To discover new mutational mechanisms driven by retrotransposons within cancer genomes.** Leveraging long-read sequencing data, this objective focuses on uncovering previously unrecognized ways in which retrotransposons contribute to genomic instability and cancer progression.

iv. **To elucidate the tempo of somatic retrotransposition throughout cancer progression.** This objective involves adapting existing timing approaches for the study of somatic retrotransposition. The aim is to provide a thorough outline of retrotransposition activity across various stages of cancer development and progression.

v. **To assess somatic retrotransposition in healthy tissues.** This analysis aims to establish a baseline understanding of the occurrence of somatic retrotransposition in non-cancerous cells. This understanding is essential for distinguishing mutational patterns specific to cancer.

METHODOLOGY

9 METHODOLOGY

9.1 BIOLOGICAL SAMPLES

9.1.1 Primary tumours cohort

We focused our research on three types of cancer that have previously been shown to exhibit high rates of retrotransposition[68]. Our screening cohort consisted of 150 primary tumours, comprising 50 HNSC, 50 LUSC and 50 COAD. The samples included primary tumours, normal tissues adjacent to the tumours and blood, all of which were fresh frozen samples taken from treatment-free donors. These samples were collected and kindly provided by the Basque Biobank, part of the Spanish Biobank Network and the Galician Foundation of Genomic Medicine (Spain). For comprehensive metadata and detailed technical specifications of our sample cohort, please refer to **Supplementary Table 1**.

Approval for this study was granted by the appropriate Clinical Research Ethics Committee, as detailed in the Annex 'Ethics Committee Approval'. Throughout this study, we adhered to the latest guidelines of the International Conference on Harmonization, the Standards of Good Clinical Practice, the Declaration of Helsinki and the Oviedo Convention. The handling of personal data from subjects participating in the study was conducted in strict accordance with the European Regulation (EU) 2016/679 and the Organic Law 3/2018, of December 5, on the protection of personal data and guarantee of digital rights. To ensure the confidentiality of patients' personal data, all sample specimens were coded using a secure, anonymized system.

9.1.2 Healthy tissues dataset

In collaboration with Iñigo Marticorena and other members of the Welcome Sanger Institute (UK), a primary aim of this PhD thesis was to comprehensively characterize patterns of somatic retrotransposition across various healthy tissues. Our study involved analysing whole-genome short-read sequencing data derived from 504 LCM biopsies collected in previous studies[117–119]. These biopsies, obtained from 28 healthy donors, spanned eight different tissue types, including placenta (n=117), stomach (n=145), appendix (n=34), colon (n=59), pancreas (n=19), skin

(n=12), small bowel (n=59) and testis (n=59). For detailed metadata on all the biopsies, including technical specifics like sequencing depth and clonality, refer to **Supplementary Table 2**.

9.2 EXPERIMENTAL APPROACHES

9.2.1 DNA isolation

DNA extraction was performed on fresh-frozen tissues from primary tumours and their matched normal counterparts for sequencing analysis. We employed the AllPrep DNA/RNA Mini Kit (Qiagen, California, USA) for genomic DNA isolation. For long-read sequencing purposes, we used the Short Read Eliminator XS buffer (Circulomics, Maryland, USA) to remove DNA fragments shorter than 5 Kb. The DNA was then purified using Agencourt AMPure XP magnetic beads (Beckman Coulter, California, USA), following the manufacturer's instructions. DNA integrity and purity were assessed prior to advancing to subsequent analyses.

Note: *DNA isolation was conducted by our lab technicians, specially by Ana Pequeño and Jorge Rodriguez-Castro (University of Santiago de Compostela, Spain).*

9.2.2 Retrotransposition screening

We conducted a screening within our primary tumour cohort to identify samples exhibiting high rates of somatic retrotransposition. For this, tumour DNA samples were subjected to shallow whole-genome sequencing at Macrogen (Seoul, South Korea). This sequencing employed Illumina PCR-free, paired-end libraries with a 350 bp insert size and 150 bp sequencing reads. This approach resulted in an average coverage of 11.7x, with a 95% confidence interval ranging from 11.5x to 11.9x. Alignments to the reference human genome (assembly GRCh38) were performed using bwa mem v0.7.17[123] and refined with Samtools v1.12[152]. To ensure data quality, duplicate reads were identified and discarded from subsequent analyses using Biobambam2 v2.0.87[153].

The resulting BAM files were then analysed using xTea[98] for detecting L1 retrotransposition events. To confidently identify somatic events, given the absence of genomic data from adjacent tissues at this stage, we genotyped the resulting L1 insertions using MEIGA-SR and excluded those detected in the tumour of any other donor within the cohort. Additionally, we excluded events documented in the comprehensive database of retrotransposon polymorphisms generated within the framework of the 1000 Genomes Project[126]. For our in-depth analysis, we selected all primary tumours exhibiting at least 100 L1 insertions, a number indicative of high retrotransposition rates, in our cohort. This criterion led to the selection of ten tumours for subsequent analysis, including five HNSC (PD0266a, PD0270a, PD0274a, PD0277a and PD0287a), four LUSC (PD0307a, PD0312a, PD0319a and PD0331a) and one COAD (PD0351a).

9.2.3 Whole-genome sequencing

To achieve a comprehensive genomic characterization, we conducted Illumina short-read sequencing with a 30x coverage on the ten tumours selected in our screening, along with their matched normal counterparts. Sequencing data were generated and analysed following the procedures described in the previous section. In addition, we performed on-site ONT long-read sequencing on the same samples. In brief, we initiated the process with an end-repair and dA-tailing step using the NEBNext End Repair/dA-tailing module (NEB). Subsequently, we constructed whole-genome libraries for each unsheared DNA sample using the Oxford Nanopore Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies Ltd). These libraries were loaded into MinION R9.4 flow cells (FLO-MIN106, Oxford Nanopore Technologies Ltd) and sequenced to a minimum depth of 30x. MinION devices were controlled by the MinKNOW software v18.12.09. High-accuracy basecalling was performed in real-time using the GPU-dependent Guppy software v2.3.1. Generated fastq files were aligned on the reference genome (GRCh38) with Minimap2 v2.24[122], and the resulting alignments underwent sorting and quality-based filtering with Samtools v1.12[152].

Note: *ONT sequencing experiments were conducted by our lab technicians, specifically Ana Pequeño and Jorge Rodriguez-Castro (University of Santiago de Compostela, Spain). Data handling of ONT sequencing experiments was carried out by our bioinformatics technicians, specially by Javier Temes (University of Santiago de Compostela, Spain).*

9.2.4 Fluorescent in situ hybridization

To validate our findings, we included an additional lung adenocarcinoma cell line, NCI-H2009, and its corresponding normal counterpart, NCI-BL2009, in our study. We conducted FISH experiments on the NCI-H2009 cell line to assess the occurrence of a reciprocal translocation between chromosomes 3 and 6 mediated by the somatic integration of L1. We employed commercially available whole chromosome painting (WCP) probes targeting chromosomes 3 and 6 labelled with either green or red fluorescent dyes (FWCP-03 and FWCP-06, respectively; Creative Bioarray). Metaphase spreads were obtained from NCI-H2009 cells following standard techniques, pre-treated with RNAse and pepsin and subjected to dual FISH experiments as per the provider's recommendations. After a counterstaining with DAPI (0.14 pg/ml), photographs were taken for each individual colour with a Nikon Eclipse-800 fluorescence microscope (Tokyo, Japan) equipped with a DS-Qi1Mc CCD camera (Nikon) and controlled the using NIS-Elements software (Nikon). The resulting images were merged and processed using Adobe Photoshop CS6 (San Jose, CA, USA).

Note: *The FISH experiments were conducted in their entirety by Daniel García-Souto (University of Santiago de Compostela, Spain).*

9.3 BIOINFORMATIC METHODS

9.3.1 Haplotype-phasing of long-reads

We used WhatsHap[154] to phase germline polymorphisms and reconstruct the parental haplotypes. The phasing pipeline utilised in this study can be summarised in three key steps:

i. Variant calling: To identify germline polymorphisms, we employed the variant caller Freebayes[155] on the Illumina short-read sequencing data obtained from matched-normal samples. Both SNPs and short indels were considered. The command used was the following:

```
freebayes -f $ref \
-C 5 \
$bam_illumina > $snv_vcf
```

ii. Graph construction and phasing inference: Utilizing ONT long-read sequencing data, we employed Whatshap to first construct a graph that captures the different ways the germline variants detected in (1) can be combined, and then perform phasing inference by assigning the most likely haplotype to each parental chromosome. The command used was the following:

```
whatshap phase --indels \
--tag=PS \
--ignore-read-groups \
-o $snv_phased_vcf \
--reference=$ref \
$snv_vcf $bam_ont
```

iii. Output: The final step consisted of using WhatsHap to add phasing tags to each alignment that could be confidently associated with a parental haplotype. This step was performed on the alignments included in the ONT long-read sequencing data derived from both the tumour and matched normal tissues.

```
whatshap haplotag --ignore-read-groups \
-o $bam_ont_phased \
--reference $ref \
$snv_phased_vcf $bam_ont
```

Phased datasets were employed to precisely estimate the VAFs of retrotransposition events, achieving haplotype-level resolution. Furthermore, these datasets were instrumental in differentiating alleles originating from the same source L1 locus, significantly enhancing the accuracy in depicting the activity landscape of the source L1 elements.

9.3.2 Reconstruction of source L1 elements repertoire

To comprehensively characterize the activity patterns of source L1 elements in our tumour samples, we aimed to reconstruct the full repertoire of potentially active L1 elements for each donor, achieving haplotype-level resolution. To accomplish this, we utilized ONT phased BAM

files obtained from matched-normal samples. We employed two different approaches depending on whether the source elements were present or absent on the reference genome.

For L1 elements present in the reference genome, referred to as Ref-L1, the complete reconstruction of their sequences was performed using the *'refSrc seqs.py'* pipeline included in the MEIGA suite. In brief, we first collected alignments that spanned the L1 element and its adjacent regions. These reads were then segregated based on their haplotype tags and mapped to the L1Hs consensus sequence. For each distinct haplotype, the resulting alignments were aggregated and assembled using Wtdbg2[120]. Subsequently, the assembled sequences underwent a polishing step with Racon[121], and were finally realigned to the L1Hs consensus to identify the precise breakpoints of the source L1 element.

For source L1 elements absent in the reference genome, termed nRef-L1, we adopted a distinct approach. As these elements manifest as insertions in the genome, we employed MEIGA, using its default parameters, to identify full-length solo L1 insertions in the matched-normal tissue. In cases of homozygous nRef-L1 loci, we re-ran MEIGA on the target regions employing the *'*--hpTag*'* parameter for the reconstruction of haplotype-specific sequences. For heterozygous nRef-L1s, we relied on the insertion sequences initially reconstructed by MEIGA. Using the haplotype-aware version of MEIGA could have resulted in reduced sensitivity, due to the exclusion of reads where haplotype assignment is uncertain.

9.3.3 Somatic retrotransposition calling on long-reads

To characterize somatic retrotransposition patterns using long-read sequencing data, we employed MEIGA. Our approach involved conducting a paired analysis using ONT BAM files from both tumour and matched-normal tissues. For this analysis, we customized the repertoire of source L1 elements for each donor, which improved the precision in attributing activity to specific L1 loci. The specifics of how these donor-specific source L1 element repertoires were obtained are thoroughly outlined in the previous section. Each repertoire was then supplied to MEIGA, employing the '--transduction-search' parameter. For all other input parameters, the defaults were maintained. The command used was as follows:

```
# A) Set sample specific input
bam=/path/to/bam
normalBam=/path/to/normalBam
outDir=/path/to/output
mkdir -p $outDir

# B) Set general input
meigaDir=/path/to/meiga
reference=/path/to/ref/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna
databases=/path/to/databases
srcL1==/path/to/donor_L1_repertoire
technology=ont
threads=8

# C) Launch MEIGA
python $meigaDir/MEIGA-LR.py \
$bam \
$technology \
$reference \
$databases \
--transduction-search $srcL1 \
```

```
--normalBam $normalBam \
-p $threads \
-o $outDir
```

9.3.4 Source inference of solo-L1 retrotranspositions

With the advent of long-read sequencing data, we can now comprehensively reconstruct integrated retrotransposon sequences and create a detailed catalogue of potentially active L1 elements. Such data facilitate the identification of diagnostic polymorphisms that can be used for tracing somatic solo-L1 insertions back to their source L1 elements. To accomplish this, we employed the source inference method developed by Martin Santamarina (University of Santiago de Compostela, Spain). Utilizing a series of pairwise alignments, this method identifies diagnostic nucleotides and uses them to associate candidate insertions with specific source L1 elements, based on the likelihood that the shared set of variants is the only plausible explanation.

The accuracy of this inference method is noteworthy. Concordance assessment with simulated solo-L1 insertions yielded a high value of 99.9%. However, assignation rates varied across different L1Hs subfamilies, with Ta-1 at 36.8%, Ta-0 at 47.1% and pre-Ta at 59.6%. When evaluated with real data involving L1-partnered transductions, concordance estimates exceeded 89.9%, accompanied by assignation rates of 27.8%.

In this study, given the method reliance on the precise reconstruction of sequences for both somatic solo-L1 insertions and candidate source L1 elements, we employed MEIGA to generate these sequences. For each patient, somatic L1 insertions were identified during the tumour-normal paired analysis, while candidate source L1 elements were reconstructed through an analysis of the matched normal tissue. To mitigate the impact of ONT sequencing errors, the analysis focused exclusively on profiling base substitutions as diagnostic variants. We then executed the inference method using the default parameters.

9.3.5 Detection of polyadenylation signals

APARENT[132] is a deep neural network designed to detect and score all polyadenylation sites within a given DNA sequence. In this study, we utilised this software to identify both canonical and alternative polyadenylation signals and to estimate their relative strengths. For each tumour sample, APARENT was applied to the reconstructed sequences of source L1 elements along with their unique downstream sequences. Subsequently, this information was correlated with the tendency of each source L1 element to induce solo insertions and partnered transductions.

9.3.6 SV variant calling on long-reads

To characterize somatic SVs in ONT long-read sequencing data, we employed the variant caller Sniffles[134]. Owing to the absence of paired analyses in Sniffles, we implemented a variant calling approach specifically designed to effectively filter out germline variants from the identified SVs. First, we conducted independent SV calling on both tumoral and matched normal samples avoiding tandem repeat regions as per developers' recommendations. To that end, the following commands were used:

```
# SV calling on tumours
sniffles --input $bam --snf $snf --non-germline --allow-overwrite --
output-rnames --reference $ref --threads $cpus --tandem-repeats
$tandem_repeats

# SV calling on matched-normals
sniffles --input $normalBam --snf $normalSnf --allow-overwrite --
output-rnames --reference $ref --threads $cpus --tandem-repeats
$tandem_repeats

# SV genotyping
sniffles --input $snf $normalSnf --vcf $vcf --output-rnames --allow-
overwrite --threads $cpus
```

Subsequently, we only retained SVs that were supported by at least four reads in the tumoral samples and were absent in their corresponding matched normal counterparts. To filter out variants linked to retrotransposon activity, we conducted a comparative analysis between the somatic calls from Sniffles and MEIGA. Additionally, we introduced a further filtering step to exclude BNDs identified by Sniffles that were located downstream of source L1 elements. These BNDs are more likely indicative of L1 transductions rather than actual translocations and therefore, were excluded from our analysis.

9.3.7 Somatic retrotransposition calling on short-reads

For analysing somatic retrotransposition in healthy tissues, we employed MEIGA-SR in paired mode, using matched samples from a different tissue, typically muscle. While MEIGA-SR was run using its default parameters, an additional filtering step was implemented downstream. Somatic retrotransposon insertions were retained if meeting the following criteria:

i. ≥6 reads supporting the insertion.

ii. ≥1 discordant reads supporting each cluster.

iii. ≥1 clipped reads supporting a poly(A/T) tail longer than 10 bp.

Subsequently, somatic retrotransposition insertions identified in all microbiopsies from a given donor were clustered using a breakpoint window of ±30 bp to generate a list of non-redundant insertions. This curated list was then re-genotyped across all the donor's microbiopsies. We retained only those re-genotyped insertions that were supported by more than 4 discordant reads. Finally, all calls were confirmed through visual inspection using IGV.

9.3.8 VAF estimation of somatic retrotranspositions

To characterize the patterns of somatic retrotransposition in short-read sequencing data, we leveraged the coding developed for MEIGA and created an implementation for short-read sequencing data, named MEIGA-SR. This method was used within the timing analysis to estimate the VAFs of somatic retrotranspositions in our short-read sequencing dataset. To increase sensitivity in identifying signs of retrotransposition, we modified relevant filters and

incorporated them in a genotyper mode. These adaptations encompassed the following criteria to retain a call:

i. ≥3 reads supporting the insertion.

ii. ≥20 total reads in region.

iii. ≥50 MAPQ in region.

iv. TSD identified and smaller than read size (150 bp).

Employing this approach, we genotyped both somatic retrotransposon insertions and retrotransposon-mediated rearrangements identified by MEIGA in the paired analysis of the long-read sequencing dataset. The genotypes obtained were then utilized in the subsequent timing analysis.

9.3.9 Variant calling on short-reads

We used Illumina whole-genome sequencing data to comprehensively characterise somatic mutations within the selected tumours. To ensure a robust and standardised analysis, we applied established algorithms commonly used in the field for tumour-normal paired analysis. For the identification of point mutations and indels, we employed MuTect2[156] following GATK's best practices. Somatic SVs were called using SvaBa[157] in paired mode with default parameters. To detect copy number aberrations, determine purity and establish ploidy profiles, we used the Battenberg[158]. Minor adjustments were made to the Battenberg algorithm to ensure its compatibility with the GRCh38 human genome assembly. To improve the accuracy of breakpoints, we integrated SvaBa SV calls into the Battenberg runs, excluding inversions as they do not indicate a change in copy number.

9.3.10 Timing of somatic retrotranspositions

The rationale behind timing methods in the context of tumour evolution is simple. When a mutation occurs in a region with copy number gains, its timing can be estimated by counting point mutations within duplicated regions in the tumour sample. Mutations that happened before the duplication event are duplicated along with the chromosomal region, while mutations occurring after the duplication or on a non-duplicated chromosome remain in single copy. We applied this rationale to analyse Illumina whole-genome sequencing data using a previously established method, mutationTime.R[143,159].

We first used MEIGA-SR on genotyper mode to assess the VAFs of retrotransposition events as described in section: '9.3.8. VAF estimation of somatic retrotranspositions'. Estimated VAFs, together with tumour purity and copy number states from Battenberg, were used as input for mutationTime.R for relative timing estimates. To enable the timing of subclonal SNVs and retrotranspositions, we incorporated a pseudo-subclone into the analysis, assumed to be present in 30% of the cells in the biopsy. This adjustment was necessary as the sequencing coverage of our samples was generally too low to accurately infer subclonal SNV structures. Furthermore, we utilized CpG>CpT transitions identified through Mutect2 as a mutational clock to determine

the real-time timing of WGDs using an approach similar to that outlined by Gerstung *et al.*[143]. By doing so, we were able to place specific retrotransposition events within a broad time-frame window, measured in years, based on whether they occurred before or after the WGD event.

In the multisample case of donor PD0270, a retrotransposition event was classified as subclonal if it was either missing or identified as subclonal in at least three out of six samples. Conversely, it was classified as clonal early or clonal late if at least four of the six samples had this specific timing classification. If none of these criteria were met, the retrotransposition event was then classified as clonal NA.

Note: *The MutationTime.R analysis and the timing of WGD events were kindly conducted by Toby Baker (The Francis Crick Institute, UK).*

9.3.11 Timing of WGD events

For each tumour, the timing of the WGD event was inferred using the proportion of clonal clock-like CpG>TpG transitions across varying allelic copy numbers, known as their multiplicities. Our analysis was restricted to genomic regions where the highest number of copies of either parental allele, termed the major copy number, was two. This restriction assumes that all gains in these regions are a result of the WGD event. We postulated a linear acceleration, estimated at 5x, in the CpG>TpG mutation rate, occurring at a time uniformly distributed between 1 and 15 years prior to diagnosis. Given a sampled acceleration time point, we identified the WGD timing that corresponded to the maximum likelihood of the proportion of multiplicity one and two clocklike SNVs in genomic region. The likelihood was then marginalised over the fraction of multiplicity one mutations that were subclonal, and the multiplicity proportions were adjusted to compensate for our limited ability to detect SNVs with fewer than three supporting reads. The uncertainty in our WGD timing estimates was quantified by bootstrapping over the tumour SNVs and resampling the acceleration time point.

As we are blind to subclonal mutations below our detection limit, this analysis is expected to show a slight bias towards a later WGD. The WGD in sample PD0312 was not timed as it had a modal major copy number state of greater than two, suggesting the occurrence of more than one WGD, which was unable to be timed using this method.

9.3.12 Simulation of retrotransposition events

To establish a robust benchmarking framework, we utilized ME-simulator to generate in silico whole-genome datasets that mimic the complexities and characteristics of real retrotransposition activity. ME-simulator is an algorithm developed by Martin Santamarina (University of Santiago de Compostela, Spain) that has been included in the MEIGA suite. It functions by placing retrotransposon sequences at predetermined positions within a mock reference genome and simulating reads based on that reference. To generate Illumina short-read sequencing datasets, ART[160] is utilized internally, while PBSIM[161] is employed for obtaining ONT long-read simulated data.

Utilizing ME-simulator with its default settings, we simulated a total of 6,420 retrotransposon insertions, distributed as follows: 1,000 Alu-solo; 1,000 SVA-solo; 1,000 L1-solo; 1,710 L1-partnered and 1,710 L1-orphan. Specifically, for each of the 114 source L1 elements active in the PCAWG dataset, we simulated 15 L1-partnered and 15 L1-orphan transduction events. In addition, we simulated 20 deletions, 20 duplications, 20 inversions and 10 translocations, all L1-mediated. This entire set of retrotransposition events was accurately replicated in both short-read and long-read sequencing datasets. In the short-read simulation, we set a sequencing depth of 30x and a VAF of 50%. For the long-read simulation, retrotransposition events were simulated at various VAFs, including 15%, 30% and 50%, and a sequencing depth of 33.7x, which corresponds to the median depth of our ONT cohort.

Note: *Simulations were conducted by Martin Santamarina (University of Santiago de Compostela, Spain).*

9.3.13 Evaluation of retrotransposition callers

We used the ONT simulated datasets to conduct an unbiased evaluation of retrotransposition callers for long-reads, including rMETL[95], xTea[98] and PALMER[96], as well as our own method, MEIGA. Each caller was executed using default parameters to ensure a consistent evaluation and fair comparison of their capabilities. Key metrics such as precision, recall and computational efficiency were assessed to evaluate the performance of each tool. Recall metrics did not consider identity assignation, and a 150 bp offset was allowed for intersecting the calls. Furthermore, we used Blat[162] to evaluate the reconstruction of the inserted sequences by analysing observed-to-expected length ratios and the percentage of sequence identity.

We utilized the short-read simulated dataset to conduct a precise assessment of various retrotransposition callers specifically designed for short-read analysis. This included evaluating tools such as MEIGA-SR, TraFiC[68,92] and the xTea module tailored for short-reads[98]. Each method was executed using its default settings. Furthermore, this dataset was employed to evaluate the accuracy of MEIGA-SR in estimating the VAFs of retrotransposition events, as thoroughly detailed in the section titled ‘9.3.8. VAF estimation of somatic retrotranspositions’.

Note: *Benchmarking analysis was collaboratively performed by myself and Nuria Espasandin (University of Santiago de Compostela, Spain).*

RESULTS AND DISCUSSION

10 RESULTS AND DISCUSSION

10.1 CHARACTERIZING RETROTRANSPOSITION IN THE CONTEXT OF LONG-READS

The availability of computational tools for detecting retrotransposon insertions from long-read sequencing has been limited. When we initiated this project, there were no published methods available for this purpose. Nevertheless, for the last four years, several tools have been developed and publicly released, including xTea, rMETL and PALMER. However, despite these advancements, these tools were primarily designed for the identification of canonical insertions of L1, *Alu* and SVA elements, with limited capacity to detect non-canonical insertions such as pseudogenes or rearrangements resulting from aberrant retrotransposition. Of note, none of the currently available tools were developed for identifying somatic retrotranspositions from paired analyses* of tumour and matched normal samples, a critical requirement in cancer research. Hence, to overcome these limitations and gain a more thorough insight into the impact of retrotransposons on the cancer genome, we developed a bioinformatics algorithm called the Mobile Element Integrations Genome Analyzer (MEIGA).

MEIGA is a robust and versatile tool for identifying and characterizing somatic retrotranspositions from long-read sequencing data in the context of paired analysis. Our method detects retrotransposon insertions, as well as various genomic rearrangements resulting from the aberrant integration of retrotransposons, including deletions, duplications, inversions, translocations and more complex SVs. MEIGA categorizes retrotransposition events into five distinct categories based on sequence identity:

i. Solo-retrotranspositions, which include full-length or partial sequences of L1, *Alu* and SVA elements.

ii. Partnered transductions, in which an L1 element, along with a unique piece of genome downstream to it, are retrotransposed.

* Paired analyses involve comparing tumour mutations with the patient's normal genetic background in control tissues, such as blood or adjacent healthy tissue. This approach is crucial in cancer research as it helps identify genetic alterations that are somatically acquired and specific to the tumour.

iii. Orphan transductions, in which only the unique sequence downstream to an active L1 is retrotransposed, without the cognate L1.

iv. Processed pseudogenes (PSDs), which include events where mRNA from nuclear genes is retrotransposed.

v. Solitary polyadenylate tracts, which encompass poly(A/T) tracts resulting from truncated retrotranspositions.

In paired analyses, MEIGA effectively differentiates between somatic and germline events. Moreover, this method is specifically designed to accurately reconstruct inserted sequences and to identify key structural features within them. These features include, for example, the TSD and the poly(A/T) tail, which are hallmarks of the canonical mechanism of integration known as TPRT. Overall, these capabilities render MEIGA well-suited for comprehensively characterising the mutational patterns mediated by retrotransposition within tumours.

10.1.1 A comprehensive overview of MEIGA

MEIGA takes as input a BAM file containing aligned reads on a reference genome, specifically generated from long-read sequencing platforms like ONT and PacBio. While MEIGA is tailored to work with the human reference genome, assembly GRCh38, it can be easily adapted for other assemblies or species if repeat annotations are available. The MEIGA workflow involves five essential steps for accurately identifying retrotransposition events: (a) Recruitment of supporting reads; (b) Read clustering; (c) Identifying insertions and rearrangements mediated by retrotransposition; (d) Determining insertion and bridge structures; and (e) Applying filtering criteria. These well-defined stages collectively provide MEIGA with a high degree of versatility to meet specific research requirements.

a) Recruitment of supporting reads:

In an initial stage, MEIGA identifies reads that provide evidence for various types of SVs within the reference genome. To accomplish this, MEIGA examines the CIGAR strings of BAM alignments and distinguishes between two types of reads: (1) spanning reads, which include an alignment that covers the full length of an insertion, and (2) split reads, which cover one of the two breakpoints delimiting an insertion or other type of SV. Within the split reads, two alignment segments are identified: the anchor segment, which aligns onto the region of the reference genome immediately adjacent to the relevant SV; and the clipped segment, which supports either an insertion larger than the clipping length or another type of SV.

To ensure the precision of MEIGA calls, only read alignments with a mapping quality (MAPQ) exceeding 20 are considered by default in the initial step of read retrieval. In addition, the algorithm is preconfigured to identify SVs larger than 50 bp by default, which helps to reduce the error rate associated with the loss of base-calling accuracy observed for short indels in ONT data. When MEIGA is executed in paired mode, recruitment of supporting reads is simultaneously performed in both the tumour and matched normal sequencing datasets.

b) Read clustering:

The read clustering process comprises two consecutive steps: primary clustering and meta-clustering. In primary clustering, a cluster is formed with a default requirement of at least two distinct reads. During primary clustering, split reads are grouped based on two conditions: (a) their anchor segments share the same clipping orientation, and (b) their distance to the closest breakpoint within the cluster is by default ≤50 bp. Similarly, spanning reads are clustered together if their embedded insertions are closer than 250 bp by default. Then, spanning clusters undergo a correction step to resolve insertion fragmentation caused by alignment contiguity defects. This correction process includes merging shorter insertions and redefining insertion boundaries for supporting reads displaying insertion fragmentation.

In the next stage, split clusters and spanning clusters undergo a meta-clustering process with the objective of grouping clusters supporting the same SV. To conform a metacluster, it is required by default a reciprocal distance ≤500 bp between clusters and a minimum of three supporting reads in total. The meta-clustering process yields two types of metaclusters: insertion (INS) and break-ends (BND) metaclusters. INS metaclusters provide support for genomic insertions entirely encompassed by the reads supporting the metacluster, whereas BND metaclusters offer support for either long insertions not fully covered by those reads, or an alternative type of SV. At this step, a set of default filters is applied to refine the metacluster calls, which include (i) setting a maximum limit of 500 reads that can compose a metacluster, and (ii) excluding metaclusters providing support for insertion lengths with a coefficient of variation* greater than 40%.

c) Identifying retrotransposon insertions:

From each INS metacluster, MEIGA extracts the relevant inserted sequences, along with sequences from the upstream and downstream flanking regions extending up to 2.5 Kb in length. Subsequently, MEIGA employs wtdbg2[120] to derive a consensus from the extracted sequences, which is then polished using Racon[121]. In the rare event that consensus construction fails, MEIGA initiates the polishing step by utilizing the inserted sequence from the read containing the insertion whose length is the closest to the median. Next, to precisely define the insertion sequence and the insertion-genome junctions, the polished sequence undergoes alignment to the reference region where the insertion was located using minimap2[122].

In the subsequent step, MEIGA aims to determine the identity of the insertions by performing a whole-genome alignment, followed by an annotation step. For the whole-genome alignment, MEIGA maps the predicted inserted sequences using both bwa-mem[123] and minimap2[122]. We employ this dual approach because bwa-mem excels at characterizing short alignments that may be missed by minimap2, while minimap2 is better suited for handling longer alignments that may become fragmented when using bwa-mem.

Next, MEIGA examines the genomic annotations linked to the alignment positions retrieved for the inserted sequences, in order to identify particular categories of interest related to retrotransposon activity. These categories encompass alignments that match retrotransposon repeats, poly(A/T) repeats, sequences positioned downstream of source L1 elements, which

* The coefficient of variation is a statistical metric of the relative dispersion of data points within a data series in relation to their mean.

serve as indicators of transductions, and exonic sequences, which point to PSD mobilizations. Only INS metaclusters that fall into one of the mentioned categories are retained.

d) Identifying retrotransposition-mediated junctions:

BND metaclusters can be indicative of rearrangements mediated by retrotransposition. These rearrangements typically present as junctions between two distant breakpoints in the genome, with retrotransposition bridges connecting them. In our approach, we expect retrotransposition-mediated junctions to be supported by a single BND metacluster at each genomic breakpoint, both intricately linked by a retrotransposed sequence known as the retrotransposition bridge.

To identify these events, MEIGA first examines the longest clipping that supports each metacluster and performs a whole-genome alignment using minimap2[122]. Subsequently, our algorithm examines the alignments adjacent to the beginning of the clipping segment to identify the presence of a retrotransposition bridge. This detection is carried out using genomic annotations, as detailed in the previous step. Only BND metaclusters where a candidate bridge is identified are retained for further analysis.

MEIGA then continues to scan the alignments next to the candidate bridge, which we refer to as anchor alignments. In this phase, our primary aim is to detect pairs of BND metaclusters in which the coordinates of one metacluster overlap with the coordinates of the anchor alignments from the other metacluster. When these pairs are identified, they are classified as junctions and are retained for further analysis if they satisfy the conditions of sharing a minimum of two reads and their respective bridges exhibit the same identities. Junctions can take two forms: intrachromosomal, which are often linked to DNA loss or gain; or they can be interchromosomal, indicating a translocation event.

Finally, MEIGA aims to accurately reconstruct the junction bridge sequence, following a similar approach to the one used for the INS metaclusters. First, MEIGA gathers the supporting reads and extracts bridge sequences together with up to 2.5 Kb of the flanking regions. Next, using wtdbg2[120], it creates a consensus sequence, which is further refined with Racon[121]. If consensus construction is unsuccessful, the process begins by extracting the bridge sequence from the event that has both the longest clipped and anchor segments. Then, MEIGA accurately determines the breakpoints of the bridge by executing a minimap2[122] alignment targeting both genomic sides of the junction.

e) Refining insertion and bridge structures:

To acquire a more detailed understanding of the structural features of retrotransposition events, MEIGA employs minimap2[122] to align the sequences of insertions and junction bridges with the consensus sequence of the assigned candidate retrotransposon for each event. This approach provides a precise means of assessing various structural attributes, such as retrotransposon subfamily, internal rearrangements (e.g., inversions and deletions), event size, presence of poly(A/T) tail and TSDs, retrotransposition mechanisms and other relevant characteristics.

Following this, MEIGA conducts an additional read gathering step with less stringent criteria to refine the counts of supporting reads. After this, in somatic analysis, retrotransposition events are categorized as somatic when no supporting reads are found in the matched normal sample. Otherwise, they are classified as germline. Next, MEIGA proceeds to annotate the genomic

regions where the retrotransposition events are located, which is required for the final filtering step.

f) Applying filtering criteria:

After full characterization of the retrotransposed sequence, a final set of filters is applied to ensure specificity. Candidate insertions and junctions are retained if the following conditions are met:

i. The minimum and maximum number of supporting reads ranges from 3 to 500 reads.

ii. The retrotransposed sequence is not entirely composed of expanded repeats.

iii. The retrotransposed sequence does not comprise a low-complexity sequence, with the exception of poly(A/T) tracks.

iv. There are no incompatible identities assigned, such as a combination of L1 and *Alu* identities.

v. The proportion of the retrotransposed sequence with an assigned identity exceeds 40%.

vi. The ratio of spanning reads versus clipping reads is consistent given the insertion size.

Finally, retrotransposon insertions are reported in VCF format. INS metaclusters that do not meet the filtering criteria are also documented by default in a separate VCF file. Junctions supporting retrotransposon-mediated rearrangements are documented in BEDPE format. Filtered BND metaclusters with strong support are also documented in a separate file. In the launch command, the *extraFields* argument can be activated to obtain a more detailed output.

Note: *MEIGA was collaboratively developed, primarily with Bernardo Rodriguez-Martin, and in conjunction with Martin Santamarina, Javier Temes and Eva Garcia, all of whom were affiliated with the University of Santiago de Compostela at the time. My direct involvement encompassed filtering and refining the insertions branch, while assumed the lead role in developing the rearrangements-detection component. Currently, I am responsible for overseeing code maintenance.*

10.1.2 MEIGA outperforms previous retrotransposition callers

We evaluated the accuracy of our algorithm by generating a mock genome containing 6,420 retrotransposon insertions from various families and insertion types. Subsequently, we simulated long reads from this in silico genome, incorporating different variant allele frequency (VAF) levels of 0.5, 0.3 and 0.1 for the relevant insertions. We then ran MEIGA alongside three

additional reference pipelines —xTea, rMETL and PALMER— to detect retrotranspositions within these genomes.

The results indicated that while other methods achieved a recall rate ranging from 44.45% (PALMER) to 60.29% (xTea) for L1 insertions, MEIGA recall rate reached 96.04% (**Fig. 5a**). Additionally, MEIGA attained the highest precision at 99.91% for L1s, slightly surpassing the other three methods: xTea (99.89%), rMETL (97.36%) and PALMER (99.8%). For *Alu*s and SVA elements, recall rates exhibited no relevant differences between the methods (range=[95.9, 99.9]). However, precision rates were slightly lower for *Alu*s in the alternative methods (xTea: 93.66%, rMETL: 92.59%, PALMER: 95.23%) compared to MEIGA (99.8%), while SVAs showed similar precision (range=[95,100]).

A detailed analysis of L1 recall rates, categorized by different L1 insertion types, revealed that all algorithms effectively identified solo insertions (**Fig. 5b**). However, substantial differences arose in their ability to detect orphan transductions, a major category of retrotransposition that typically constitutes 12-13% of somatic retrotransposon insertions[68,92]. MEIGA achieved a recall rate of 97.6%, while xTea, rMETL and PALMER scored 5.03%, 7.84% and 0.7%, respectively. We also noticed that identifying L1-partnered transductions, which account for 6-12% of all somatic retrotransposon insertions[68,92], presented challenges for rMETL and PALMER. Regarding L1-transductions, MEIGA and xTea achieved recall rates of 92.28% and 95.26%, while rMETL and PALMER scored 78.77% and 57.89%, respectively.

Of note, MEIGA performed an accurate reconstruction of the simulated insertions for all retrotransposition types, resulting in an average pairwise identity percentage of 99.72% and a mean observed-to-expected length ratio (O/E) of 0.99 (**Fig. 5c; Supplementary Table 3**). Additionally, the evaluation of the pipelines performance under different levels of VAF, showed that MEIGA notably outperforms other methods in handling subclonality for all insertion types (**Fig. 5d**).

When assessing the running time and memory usage of each method, rMETL emerged as the fastest with the lowest memory requirements, only 0.83 hours per genome and a maximum memory peak of 3.7 GB under the simulated conditions of 50% VAF (**Table 1**). MEIGA ranked the second, with an average processing time of 1.56 hours and a memory usage of 26.2 GB. xTea performed the third, with a running time of 6.1 hours and utilizing 11.7 GB of memory. The last, PALMER, required separate runs for each retrotransposon family, lacked threading, and exhibited notably slow performance, averaging more than 68 days. Of note, all methods scaled effectively across various VAFs (**Table 1**).

As there are no other reference pipelines for the detection of retrotransposon-mediated rearrangements, we exclusively evaluated the performance of MEIGA in detecting such events. We generated a mock cancer genome in which we simulated deletions, duplications, inversions and translocations, all mediated by L1, at 50% VAF and a sequencing depth of 30x. MEIGA was executed with default settings, resulting in no false positives and achieving the following recall rates: 20/20 for deletions, 18/20 for duplications, 17/20 for inversions and 9/10 for translocations. It is worth noting that two missing inversions and one missing translocation were detected but discarded in the final filtering step; and that only one event was incorrectly assigned to a different SV type. Overall, these results confirm the high accuracy and reliability of MEIGA for the detection of retrotransposon-mediated rearrangements.

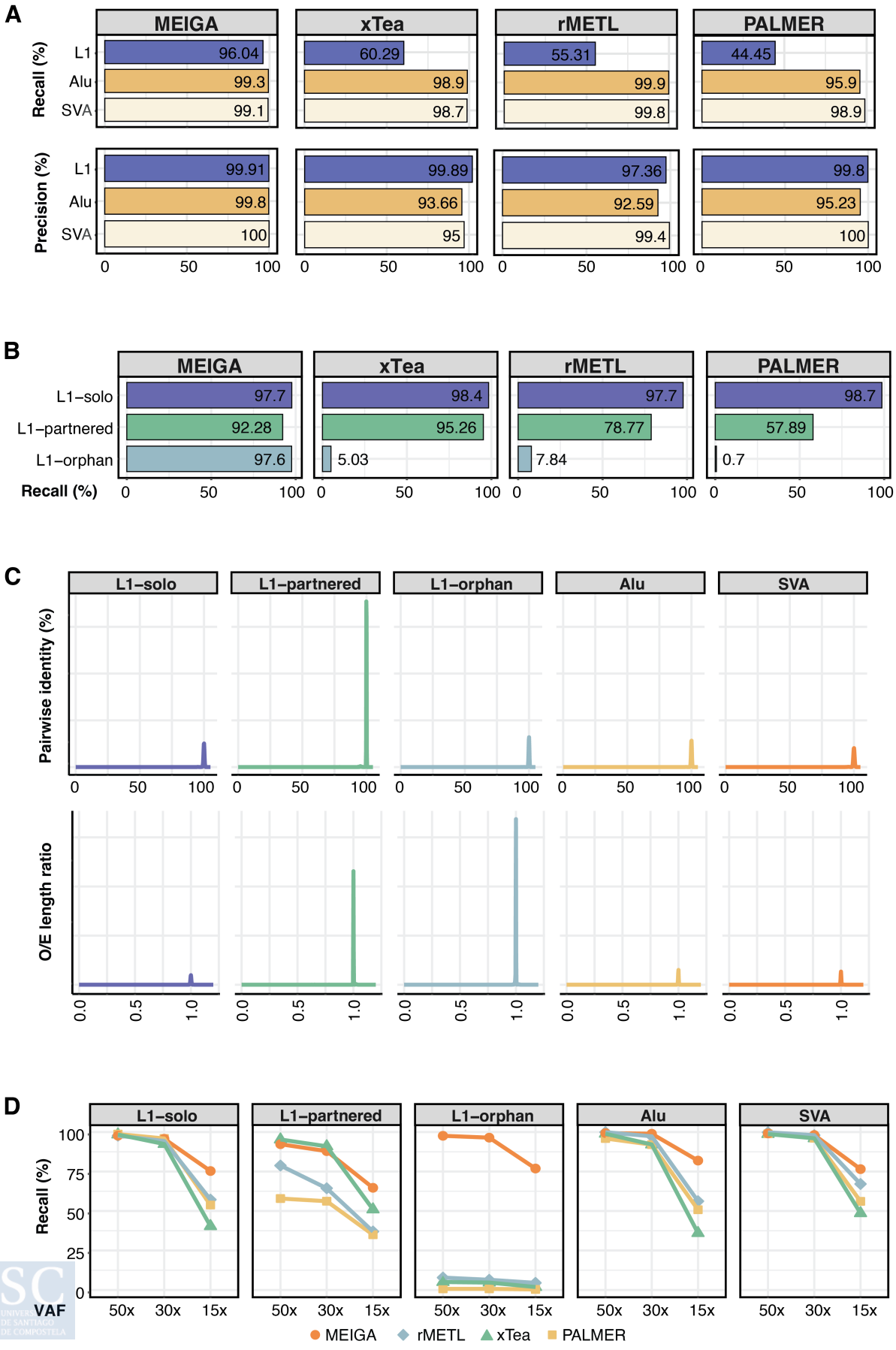


Figure 5: MEIGA outperforms alternative retrotransposition callers for long-reads. The benchmark dataset comprised 6,420 retrotransposon insertions representing various families and insertion types: 1,000 solo-*Alu;* 1,000 solo-SVA; 1,000 solo-L1; 1,710 partnered-L1 and 1,710 orphan-L1. For each active source L1 element in PCAWG (n=114), we simulated 15 partnered-L1 and 15 orphan-L1 events. **(a)** MEIGA demonstrates superior recall, particularly for L1 (99.3%, 96.04% and 99.1% for *Alu,* L1 and SVA, respectively). **(b)** Detailed analysis of L1 recall and precision rates, categorized by different L1 insertion types, revealed that all algorithms effectively identify solo insertions, but substantial differences arose in their ability to detect L1-transductions. **(c)** Statistics of the sequence reconstructions performed by MEIGA, including pairwise identity percentage and observed-to-expected length ratio (O/E). For overall insertions, MEIGA showed an average pairwise identity percentage of 99.72% and a mean observed-to-expected length ratio (O/E) of 0.99. **(d)** Evaluation of the pipelines performance under different levels of VAF (50x, 30x and 15x), revealing that MEIGA excels other methods in handling subclonality for all insertion types. L1: Long Interspersed Nucleotide Element 1; SVA: SINE-VNTR-*Alu*; VAF: Variant Allele Frequency.

Overall, the benchmarking results demonstrate that MEIGA outperforms currently available pipelines for detecting retrotransposition events. MEIGA excels in the accurate detection of transductions and stands as the sole method suitable for identifying retrotransposon-mediated rearrangements. It also offers a unique capability for paired tumour-normal analysis. In terms of performance, MEIGA ranks as the second-best pipeline, but its efficient running time and memory requirements make it well-suited for systematically processing large cohorts of cancer genomes. In summary, MEIGA is the only tool capable of comprehensively characterizing a wide range of mutational patterns mediated by retrotransposons, making it a valuable asset in the field of retrotransposition analysis.

METHOD	VAF (X)	TIME (D-H:MIN:SEC)	MAX-MEM (GB)
MEIGA	50	02:55:13	30.16
	30	01:33:26	26.20
	15	01:02:14	22.57
xTea	50	04:17:44	11.28
	30	06:05:49	11.56
	15	04:13:51	11.66
rMETL	50	00:23:04	3.79
	30	00:49:40	3.66
	15	00:48:59	3.55
PALMER	50	89-20:35:28	72.69
	30	68-21:33:26	70.75
	15	23-17:17:22	68.32

Table 1: Time and memory usage of various retrotransposition callers. D: Days; H: Hours; MIN: Minutes; SEC: Seconds; MAX-MEM: Maximum Memory peak; VAF: Variant Allele Frequency.

10.2 THE LANDSCAPE OF CANCER RETROTRANSPOSITION IN LIGHT OF LONG-READS

Somatic retrotransposition occurs in approximately 35% of all cancer genomes, with epithelial tumours exhibiting a particularly high incidence of this mutational process[68]. The PCAWG study[68], revealed that four cancer types —ESAD, HNSC, LUSC and COAD— account for over 70% of all somatic retrotranspositions observed across 38 distinct cancer subtypes. This is a significant finding considering these samples constituted only 9% of the analysed cohort. Additionally, it has been shown that somatic retrotransposition can promote cancer driver events, particularly through aberrant retrotransposition, playing a significant role in the onset and progression of certain tumours[68,79,124,125].

However, the scope of previous studies on cancer retrotransposition has been hindered by the limited library size of Illumina sequencing. Notably, this constraint has restricted our understanding of the full extent and impact of somatic retrotransposition in human cancers. Hence, the emergence of long-read sequencing technologies offers a promising avenue for a more in-depth exploration of this and other mutational processes active in cancer.

10.2.1 Characterizing tumour genomes with high RT rates using long-read sequencing

To identify a set of tumours with high rates of somatic retrotransposition, we conducted a prospective screening on a cohort of 150 cancer patients, provided by the Basque Biobank (Spain) and the Galician Foundation of Genomic Medicine (Spain). This cohort comprised 50 patients suffering from HNSC, 50 from LUSC and 50 from COAD. Our preliminary analysis included shallow whole-genome sequencing of these tumours using short-reads, while matched adjacent tissues were preserved for subsequent analysis. This data was then utilized to identify somatic retrotransposition events within the tumours.

We employed the xTea[98] algorithm to identify L1 insertions in the short-reads dataset. To confidently identify somatic retrotranspositions, given the absence of genomic data from adjacent tissues at this stage, we genotyped the resulting L1 insertions using MEIGA-SR and excluded those detected in the tumour of any other donor within the cohort. Additionally, we excluded events documented in the comprehensive database of retrotransposon polymorphisms generated within the framework of the 1000 Genomes Project[126]. Our strategy led to the identification of a subset of 10 tumours with more than 100 retrotranspositions presumed to be somatic. These high-retrotransposition-rate tumours included five HNSC, four LUSC and one COAD (**Fig. 6a**).

To gain deeper insights into the mechanisms of cancer retrotransposition, we performed whole-genome long-read sequencing using ONT on the 10 high-retrotransposition rate tumours along with their corresponding non-tumoral adjacent tissues. This sequencing approach yielded a median coverage and N50 read size of 33.7x and 19.9 kb for the tumours, and 31.1x and 18.7 kb for the adjacent tissues (**Supplementary Fig. 1a)**.

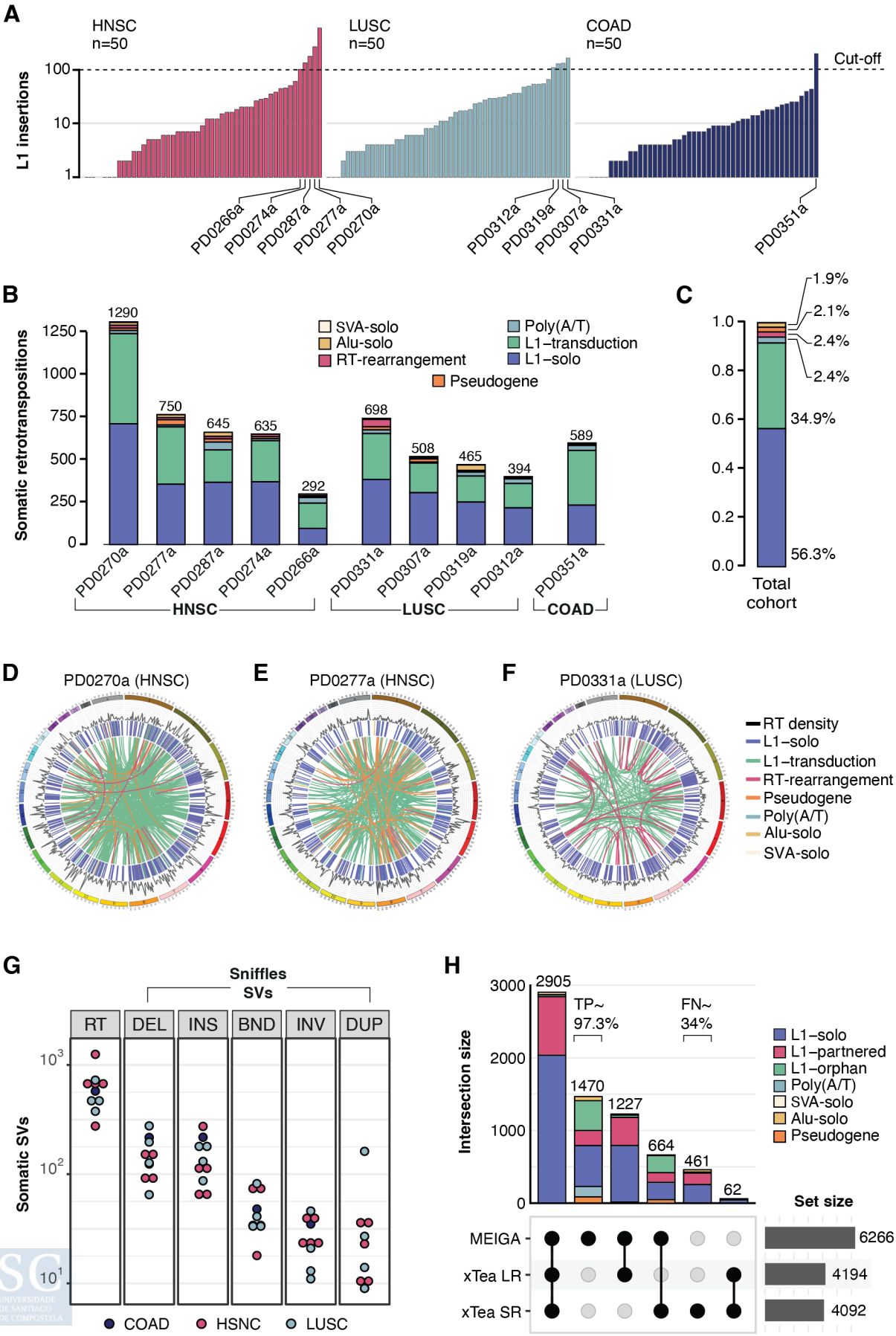


Figure 6: The landscape of somatic retrotransposition in the light of long reads. (a) Number of somatic retrotransposition (RT) events detected in 150 cancer genomes through Illumina whole-genome shallow sequencing. This includes 50 HNSC tumours, 50 LUSC and 50 COAD. **(b)** Number of somatic RT events identified in 10 cancer whole genomes sequenced with long reads using ONT. Events are classified into seven RT types: Alu-solo, SVA-solo, L1-solo, L1-transductions, poly(A/T), processed pseudogenes (PSDs) and RT-mediated rearrangements (RT-Rearrangements). **(c)** Proportion of somatic RT events attributed to each of the seven categories mentioned in (b) across the 10 tumours cohort. **(d)** Circos plot showing 1,311 somatic events in the HNSC tumour PD0270a. **(e)** The HNSC tumour PD0277, with the highest number of PSD insertions, accounting for one-quarter of the total PSDs detected in the cohort (31 out of 133). **(f)** The LUSC tumour, PD0331a, bearing 49 cases of retrotransposon-mediated rearrangements. **(g)** Number of structural variants in the set of 10 tumours sequenced with long reads, classified into RT events, identified using MEIGA, and those not mediated by RT (non-RT), identified by Sniffles. The non-RT variants were further classified into specific types: deletions (DEL), duplications (DUP), insertions (INS), inversions (INV) and break-ends (BND). **(h)** Comparative analysis of the number of somatic RTs detected with three reference methods across the set of 10 tumours from our cohort, including MEIGA, the xTea module for long-reads (xTea-LR) and the xTea module for short-reads (xTea-SR); RT types are the same mentioned in (b). We also indicate the estimated fraction of true positives (TPs) from MEIGA private calls and the fraction of false negatives (FNs) from xTea specific calls. COAD: Colorectal Adenocarcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; L1: Long Interspersed Nucleotide Element 1; LUSC: Lung Squamous Cell Carcinoma; RT: Retrotransposition; SVA: SINE-VNTR-*Alu*.

Then, we ran MEIGA on the long-read data from the 10 relevant tumours and their adjacent tissues (for full details on the MEIGA pipeline, please refer to section: '10.1. Characterizing retrotransposition in the context of long-reads'). This analysis retrieved a total of 6,266 somatic retrotransposon insertions and confirmed high rates of somatic retrotransposition in all tumours (median=622.5, range [296-1311]) (**Fig. 6b**). Overall, L1-solo elements were the most common type of insertion (56.3%, 3611), followed by L1 transductions (34.9%, n=2240), poly(A/T) tracts (2.4%, n=157), PSDs (2.1%, n=133), *Alu*-solo (1.9%, n=119) and SVA-solo (<0.1%, n=6) (**Fig. 6c**). Remarkably, in one exceptional HNSC tumour, PD0270a, we identified 1,311 somatic retrotranspositions, accounting for 20% of all events within the cohort (**Fig. 6d**).

L1-transductions, which entail the mobilization of unique DNA sequences downstream of the source L1 element, accounted for ~35% (n=2240) of the total somatic activity in our cohort. This represents a relevant increase compared to prior cancer retrotransposition studies that relied solely on short-read sequencing methods[1,2], which reported a transduction frequency of 19%-24%. Notably, this discrepancy arises from the identification of partnered transductions with exceptionally short transduced regions (i.e., < 150 bp), referred to as microtransductions (**Supplementary Fig. 2**). In this study, microtransductions constituted 43% (964 out of 2,240) of all transductions identified. Remarkably, owing to their short transduced regions, prior research on cancer retrotransposition[68] might have inaccurately classified microtransductions as solo-L1s.

While PSDs constituted a small fraction of the total somatic retrotranspositions in the cohort, one HNSC tumour, PD0277a, exhibited substantial mobilization of PSDs. Notably, this tumour accounted for a quarter of the total PSDs detected in our cohort (31 out of 133; **Fig. 6e**). A similar pattern was observed in the PCAWG dataset, where a single pancreatic adenocarcinoma

tumour harboured up to 70 PSDs. These PSD-rich patterns strongly indicate a high level of promiscuous activity of the L1 machinery within certain samples.

In addition to the events mentioned above, our analysis identified 152 instances in which a retrotransposed sequence, typically L1, bridges genomic rearrangements likely caused by aberrant integration (**Fig. 6b,c**). Retrotransposon-mediated rearrangements, including breakpoint junctions of deletions, duplications, translocations and more complex events, constituted 2.4% (152 out of 6,418) of all somatic retrotranspositions in the cohort. In a notable LUSC tumour, PD0331a, we identified 49 cases of retrotransposon-mediated rearrangements (**Fig. 6f**), which accounted for half of all the rearrangements discovered in the PCAWG cohort (total=96) and represented the highest number of retrotransposon-mediated rearrangements observed in a tumour (PCAWG: median=0, range=[0,9])[68]. Retrotransposon-mediated rearrangements are fully addressed in a subsequent section.

To gain a more comprehensive understanding of the impact of retrotransposition on the overall load of somatic SVs, we conducted an analysis using Sniffles. Sniffles is a reference SV caller designed for recognizing classical SVs —those not mediated by retrotransposons— in long-read sequencing data. When compared to other types of SVs, somatic retrotranspositions emerged as the predominant category of variation across all the tumours examined, constituting between 48%-74% of the total SV burden (**Fig. 6g**).

10.2.2 Assessing MEIGA findings on high RT tumours sequenced with long-reads

To evaluate our findings, we ran an additional long-reads pipeline, xTea, which achieved the second-best result in the benchmarking analysis (details in the section '10.1.2. MEIGA outperforms previous retrotransposition callers'), on our tumour dataset. In addition, we performed Illumina paired-end sequencing on both the tumour and matched-normal tissues at a median depth of 30x and utilized the xTea module for short-read analysis in paired mode.

In the comparison of MEIGA results with those from the two independent xTea pipelines, we observed that the majority (76.5%; 4,796 out of 6,266) of MEIGA calls were consistently confirmed by at least one additional long or short-read algorithm (**Fig. 6h**), suggesting they represent genuine retrotransposition events. For the remaining 1,470 MEIGA-specific retrotranspositions, we visually inspected a random selection of 300 insertions using the Integrative Genomics Viewer (IGV). The following criteria had to be accomplished for a candidate insertion to be classified as a real somatic retrotransposition: (i) presence of three or more supporting reads in the tumour long-read data; (ii) lack of any supporting reads in the matched normal long-read data; (iii) presence of at least one of these retrotransposition hallmarks: TSD, poly(A/T) tail or retrotransposon sequence.

The IGV inspection confirmed that 97.3% of MEIGA private retrotranspositions represent true positive events (292 out of 300). Notably, about half of these events (46.4%; 682 out of 1470) could be attributed to a sensitivity issue in the xTea pipelines to detect orphan transductions (n=442), poly(A/T) tracts (n=156) and PSDs (n=84). Regarding events exclusive to the xTea short-reads algorithm, our analysis indicated that only 34% of events (102 out of 300) inspected via IGV could be validated as real somatic events using the long-read sequencing data. As for the remaining 66%, it is worth noting that although there were false positives, the majority of these events were, in fact, subclonal variants unique to the short-read sequencing data. Due to

the lack of germline filtering in the xTea long-reads module, we opted to exclude its private events from the analysis, as these events were likely to correspond to germline variants.

In summary, MEIGA outperforms alternative methods not only in simulated data but also in real data, with an estimated sensitivity rate of 96.6% and a low false discovery rate of 0.6% (**Table 2**). For instance, in our cohort, MEIGA identifies 22.2% more genuine somatic retrotranspositions compared to both xTea pipelines (1,431 out of 6,446). These results underscore the reliability of MEIGA for the precise analysis of somatic retrotransposition patterns in tumours sequenced with long-read technologies.

Estimation of MEIGA sensitivity and False Discovery Rate (FDR):

```
RT events private to both xTea SR and xTea LR = 62
RT events private to xTea SR = 461
TP ratio of xTea SR private events = 102/300
FNs = 62 + 461 * (102/300)
FNs ~ 219


RT events private to MEIGA = 1470
FP ratio of MEIGA private events = 8/300
FPs = 1470 * (8/300)
FPs ~ 39


Total MEIGA calls = 6266
TPs = 6266 - 39 = 6227
TPs = 6227


Sensitivity = TP / (TP + FN)
MEIGA sensitivity = 6227 / (6227 + 219)
MEIGA sensitivity ~ 96.6%


FDR = FP / (FP + TP)
MEIGA FDR = 39 / (39 + 6227)
MEIGA FDR ~ 0.6%


MEIGA private TP = 1470 - 39 = 1431
Total TP: TP + FN = 6227 + 219 = 6446
TP private to MEIGA = 1431 / 6446
TP private to MEIGA ~ 22.2%
```

Table 2: Estimation of MEIGA sensitivity and FDR on real data. FDR: False Discovery Rate; FN: False Negative; FP: False Positive; TP: True Positive.

10.2.3 Long reads reveal the structure of somatic retrotranspositions to unprecedented resolution

While the majority of somatic L1 insertions exhibit internal rearrangements, including 5'-truncation and interstitial deletions and inversions, thoroughly characterising these structural

attributes has been challenging with previous sequencing technologies[11,63,79,92,127]. Low-throughput sequencing methods lack systematic applicability, and Illumina short-read sequencing is limited by its short library size. Recent studies have taken advantage of long-read sequencing to characterise internal rearrangements in germline L1 loci[98,126,128]. However, somatic insertions in the context of cancer remain largely unexplored in this regard.

We used MEIGA to precisely reconstruct the entire sequences of somatic retrotransposition events in the 10 tumours within our long-read sequenced cohort, thereby enabling us to characterise their structural attributes with unprecedented resolution. Firstly, our methodology enabled an accurate estimation of the lengths of the insertions (**Fig. 7a, Table 3**). The longest insertions in the cohort were L1-partnered transductions, showing an estimated median length of 1,094 bp. In contrast, the shortest insertions were poly(A/T) tails, with a median length of 75 bp. Notably, solo-L1 insertions showed a median length of 878 bp, indicating truncated derivatives, while processed pseudogenes exhibited a median length of 574 bp.

Insertion type	*Median*	*Mean*	*Min*	*Max*	*Std*
Poly(A/T)	75	83.4	51	590	46.7
Alu-solo	329	317.2	106	440	55.7
Pseudogene	574	832.1	96	3,646	712.9
L1-orphan	464	623.2	56	3,861	521.0
L1-solo	878	1,284.6	55	6,671	1,205.5
L1-partnered	1,094	1,464.8	112	7,421	1,228.9

Table 3: Length distribution of somatic retrotransposition events. Max: Maximum; Min: Minimum; Std: Standard deviation.

The analysis of interstitial rearrangements revealed that approximately 46% (1,661 out of 3,611) of solo-L1 insertions in the cohort contained an internal inversion, ranging from 35% and 55% depending on the specific tumour sample. Notably, we observed that solo-L1 insertions with inversions are longer (median=1,480 bp) compared to their non-inverted counterparts (median=503 bp; Wilcoxon rank-sum test, $p<0.05$; **Fig. 7b**). Interestingly, we noticed no specific inversion hotspot within the L1 sequence; instead, inversions predominantly occurred closer to the midpoint of the inserted sequence (median relative position=42.5%, **Fig. 7b**). These observations collectively suggest that the inversion point significantly influences the ultimate length of the retrotransposition event.

As we delved further into our analysis of the sequence structure of L1-solo insertions, we found intricate patterns involving reverse transcriptase jumps. For instance, in tumour PD0270a, we observed a case of an L1 insertion with an internal inversion coupled to a reverse transcriptase jump, which involved a 3-kb leap to the 3' end of the L1 mRNA template (**Fig. 7c**). Furthermore, we identified more intriguing patterns suggestive of template jumping, indicating that the reverse transcriptase can switch templates during the retrotransposition process. In a remarkable example within tumour PD0270a, we found compelling evidence of end-to-end L1 template switching (**Fig. 7d**), resulting in an insertion formed by two distinct precursor L1 mRNAs, as confirmed by sequence analysis. Altogether, these patterns suggest a scenario in which L1 insertions can result from a single L1 molecule or involve multiple molecules, leading to either simple structures or intricate chains of internal rearrangements within the context of

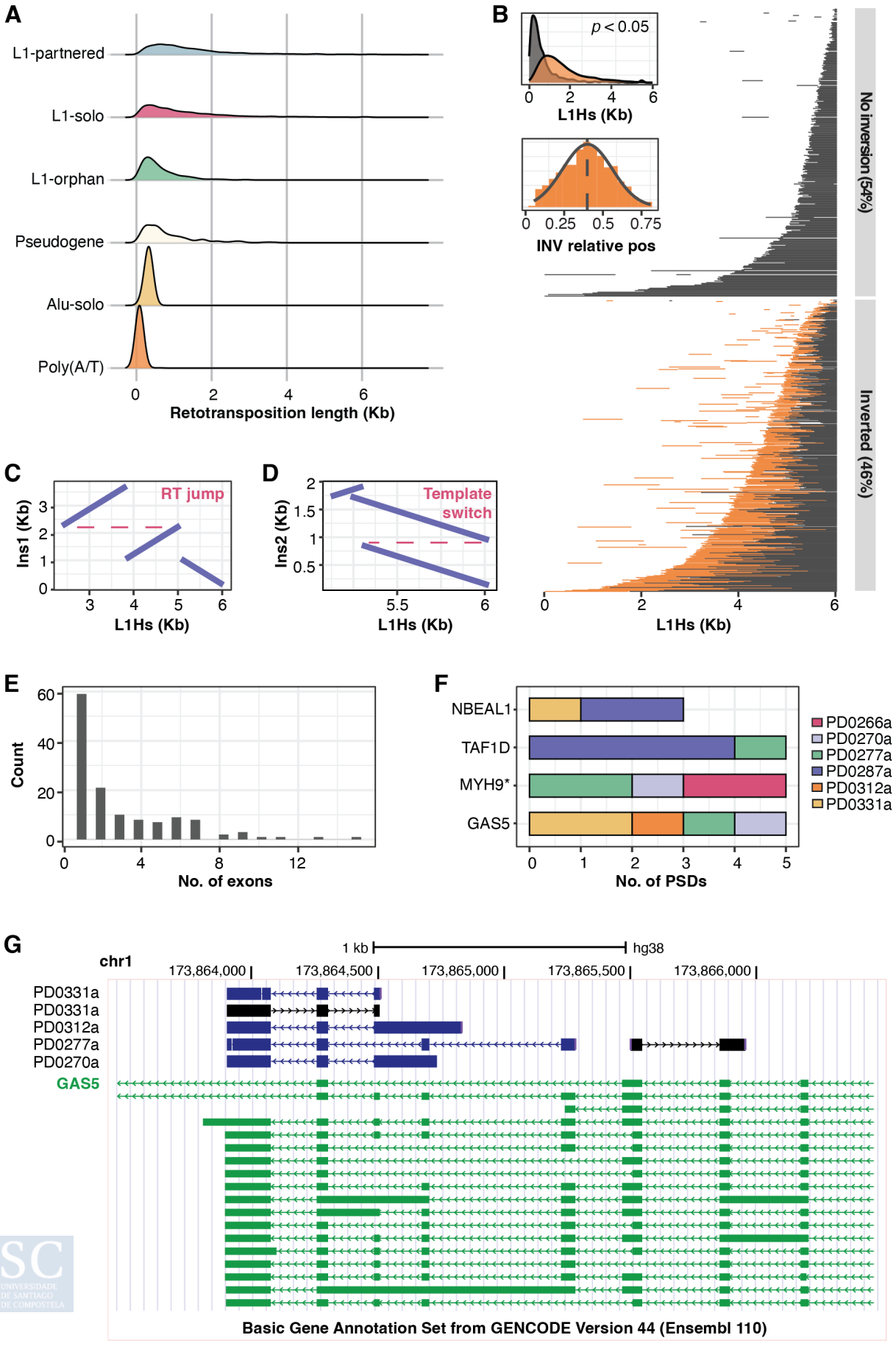

A
L1-partnered
L1-solo
L1-orphan
Pseudogene
Alu-solo
Poly(A/T)
0 2 4 6
Retotransposition length (Kb)
B
p < 0.05
0 2 4 6
L1Hs (Kb)
0 0.25 0.5 0.75
INV relative pos
No inversion (54%)
Inverted (46%)
0 2 4 6
L1Hs (Kb)
C
RT jump
Ins1 (Kb)
0 1 2 3
3 4 5 6
L1Hs (Kb)
D
Template switch
Ins2 (Kb)
0.5 1 1.5 2
5.5 6
L1Hs (Kb)
E
Count
0 20 40 60
0 4 8 12
No. of exons
F
NBEAL1
TAF1D
MYH9*
GAS5
0 1 2 3 4 5
No. of PSDs
PD0266a
PD0270a
PD0277a
PD0287a
PD0312a
PD0331a
G
chr1
1 kb
hg38
173,864,000
173,864,500
173,865,000
173,865,500
173,866,000
PD0331a
PD0331a
PD0312a
PD0277a
PD0270a
GAS5
Basic Gene Annotation Set from GENCODE Version 44 (Ensembl 110)

Figure 7: Long reads reveal the structure of somatic retrotranspositions to unprecedented resolution. (a) Insertion length distribution of somatic retrotransposition events across six categories: Alu-solo, L1-solo, L1-partnered transductions, L1-orphan transductions, poly(A/T) insertions and processed pseudogenes (PSDs). SVA-solo insertions were excluded from the analysis due to a low sample size (n=6). **(b)** Analysis of interstitial rearrangements of L1-solo insertions within the human-specific L1 sequence (L1Hs). Single insertions are depicted as horizontal lines, with orange indicating inverted segments. The inserted sequences, characterized by wide 5' truncation, are classified into two groups: those presented above without an internal inversion (54%) and those below with an internal inversion (46%). The density plot above indicates that solo-L1 insertions with inversions are longer (median=1,480 bp) compared to their non-inverted counterparts (median=503 bp; Wilcoxon rank-sum test, $p<0.05$). The density plot below indicates no specific inversion hotspot within the L1 sequence; instead, inversions predominantly occur closer to the midpoint of the inserted sequence (median relative position=42.5%). **(c)** Detailed structure analysis of a L1-solo insertion revealed an internal inversion coupled to a reverse transcriptase jump, which involved a 3-kb leap to the 3' end of the L1 mRNA template. **(d)** Detailed structure analysis of another L1-solo insertion showed a pattern suggestive of end-to-end L1 template switching, resulting in an insertion formed by two distinct precursor L1 mRNAs. **(e)** Number of exons identified in 133 PSDs detected during the somatic analysis across a set of 10 tumours with high retrotransposition rates. Approximately 44% of the PSD insertions (59 out of 133) featured a single exon, typically representing the 3'-UTR, while the rest comprised insertions involving multiple exons. **(f)** Four genes—*GAS5, MYH9, TAF1D* and *NBEAL1*—were recurrently mobilized by retrotransposition to generate PSD insertions. The number of insertions per source gene is indicated along with the tumours where they were observed. **(g)** In the case of *GAS5*, we identified up to five different PSD insertions. Although 5' truncation was a common feature in these instances, various splicing variants can be identified with the GENCODE database of transcript variants. The alignment was conducted using Blat. L1: Long Interspersed Nucleotide Element 1; mRNA: Messenger RNA; PSD: Processed Pseudogene; SVA: SINE-VNTR-*Alu*; UTR: Untranslated Region.

cancer. Conducting a comprehensive analysis of these patterns in future projects will be of paramount importance to better understand the underlying molecular mechanisms.

We also characterized the structure of PSD insertions, which are by-products of L1 activity generated when mRNA molecules from nuclear genes are retrotranscribed and integrated using the L1 protein machinery. Our approach allowed us to accurately reconstruct the sequences of the 133 detected PSDs, confirming that all the insertions presented a poly(A/T) tail and a target site duplication, both of which are hallmark features of TPRT retrotransposition. This not only facilitated the resolution of their internal structure but also enabled the identification of the alternatively spliced transcripts from which those PSDs originated.

All analysed PSDs exhibited 5' truncation. Approximately 44% of the PSD insertions consisted of single exons (59 out of 133), typically comprising a solitary 3'-UTR, while the remaining PSDs represented insertions involving multiple exons (**Fig. 7e**). Our analysis revealed that PSDs undergo internal sequence rearrangements, similar to regular L1 retrotranspositions, including 5' truncation and interstitial inversions, providing further evidence that these events represent genuine retrotranspositions mediated by the L1 machinery. In one notable example from tumour PD0307a, we identified a 3,646 bp-long insertion containing 15 exons from the *SIPA1L2* gene, representing the largest PSD insertion in our dataset (**Supplementary Fig. 3**). This event exhibited 5' truncation, resulting in the removal of the first exons of the source transcript, along with an internal inversion affecting exon 18 (GENCODE transcript ID: ENST00000674635.1).

Four genes were recurrently mobilized by retrotransposition to generate PSD insertions: *GAS5, MYH9, TAF1D* and *NBEAL1* (**Fig. 7f**). In the case of *GAS5*, we identified up to five different PSD insertions in our cohort. Although 5' truncation was a common feature in these instances, resulting in a reduction of available exons, we detected at least three different splicing variants (**Fig. 7g**). Additionally, we found up to nine instances where PSDs involved the retrotransposition of cancer-related genes, including *MYH9, MYH11, DEK, LEPROTL1, MSN, SMARCD1, SMARCE1, TPM3* and *XPO1*, as documented in the COSMIC database[129]. While the significance of this finding remains uncertain, it is worth noting that PSDs have the potential to mediate the repression of the source gene expression through siRNAs[130,131]. However, we lacked the necessary expression data to test this hypothesis.

10.2.4 Unveiling a novel panorama of source L1 elements activity

Because L1 transductions bear unique genomic sequences that were adjacent to the L1 element of origin, it is possible to unambiguously identify the source L1 element whence they derive[3]. Following this rationale, previous studies based on short-read sequencing data identified 124 active source L1 elements in human cancer[1,2]. A recent work from The Human Genome Structural Variation Consortium using long reads identified 142 active L1 elements in the germline inferred from L1 transductions[4]. Similarly, we ran a transduction-based approach to explore the activity of L1 elements in our long-reads cohort, finding 92 active L1 loci. However, these analyses did not address the element-of-origin of somatic solo-L1s, which constitute the major class of somatic retrotranspositions. Additionally, we encountered instances of L1 microtransductions for which their transduced regions did not allow an unequivocal identification of the source element from which they originated, representing 9% (200 out of 2,240) of the total transductions in our cohort.

Overall, these findings suggest that the number and the activity levels of source L1 elements in cancer may have been underestimated. To tackle this issue, we harnessed the capabilities of long-read sequencing for the precise reconstruction of target sequences. We developed an innovative approach based on identifying diagnostic nucleotides unique to specific source elements within retrotransposed sequences. This strategy leverages the fact that genetic variants are transferred from a source element to its derivative copies during the processes of transcription and retrotranscription. As a result, the identification of these diagnostic variants allows for the potential tracing of each somatic insertion back to its element-of-origin.

Our approach based on diagnostic nucleotides begins by reconstructing the allele sequences of potentially active L1 loci in the non-tumoral, adjacent tissue of each patient. Detailed information can be found in the Materials and Methods section: '9.3.2. Reconstruction of source L1 elements repertoire'. In brief, to reconstruct sequences at haplotype-level resolution, we utilized WhatsHap for phasing our long-read sequencing data (**Supplementary Fig. 1**). Then, we employed a MEIGA module, which incorporated Wtdbg2[120] and Racon[121], to obtain polished reconstructions of source L1 sequences from the phased data.

As a proof of concept, we utilized the L1 reconstructed sequences from donor PD0270 to perform a multiple sequence alignment with MUSCLE, followed by a phylogenetic analysis using the Neighbour-Joining method. This analysis demonstrated the presence of nucleotide substitutions that enabled discrimination among distinct source L1 elements (**Fig. 8a**).

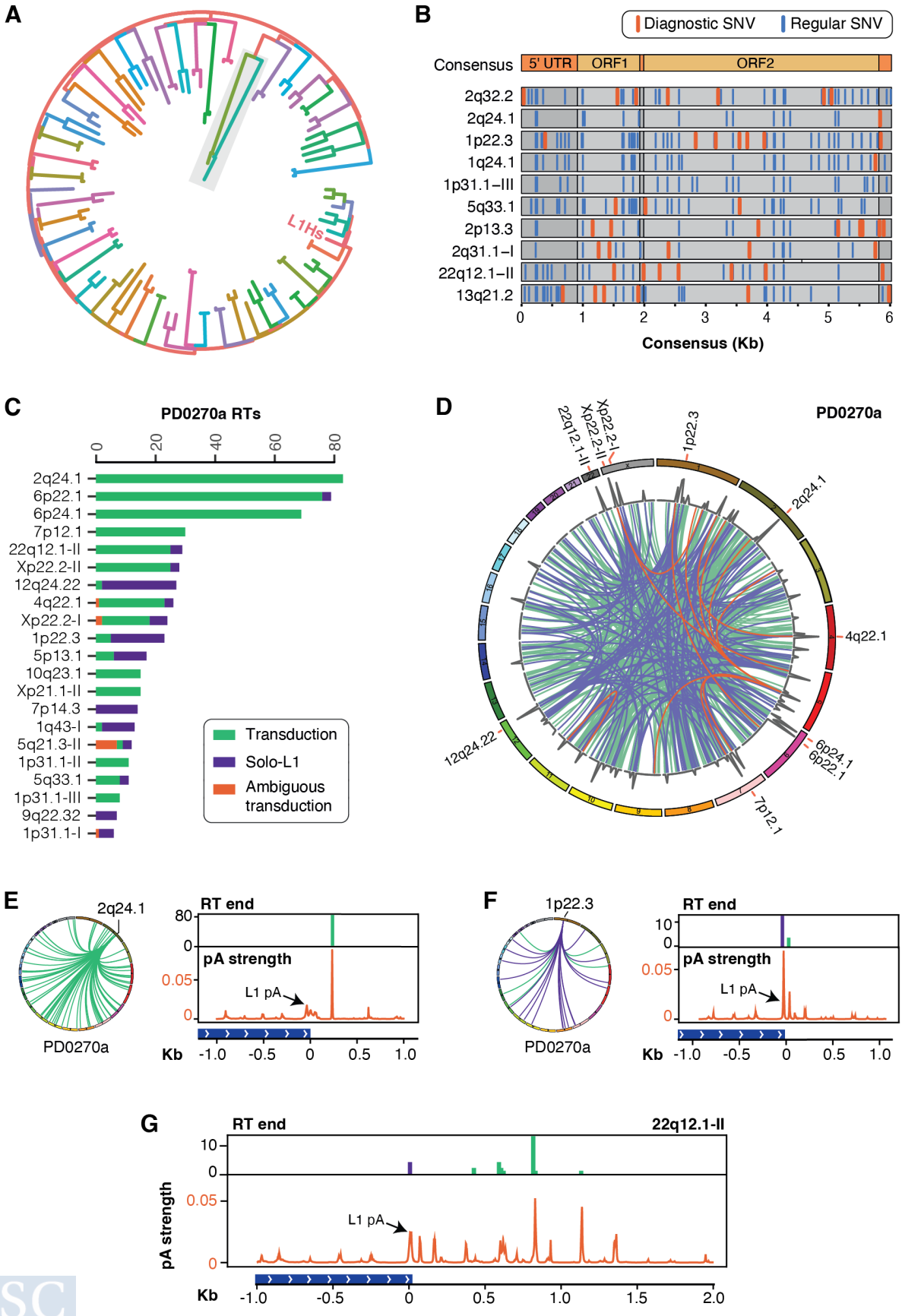


Figure 8: The dynamics of source L1 elements revealed by long reads. (a) Neighbour-Joining

phylogenetic analysis illustrating 100 randomly reconstructed L1 sequences from donor PD0270. The analysis showcases the presence of diagnostic nucleotides facilitating discrimination among distinct source L1 elements. Notably, alleles originating from the same locus, marked with the same colour, often exhibit private variants but consistently cluster together, validating the accuracy of the sequence reconstruction. A grey rectangle designates a pair of sequences excluded from subsequent analysis due to more than 1% nucleotide sequence divergence. **(b)** Multiple sequence alignment of source L1 elements from donor PD0270 aimed at identifying specific diagnostic nucleotide substitutions. Ten notable source L1 element candidates are illustrated, highlighting both the diagnostic and regular SNVs identified. **(c)** Number of L1-insertions attributed to each source element in donor PD0270a. The source element inference approach based on diagnostic nucleotides unveils hidden source element activity and facilitates the identification of novel, previously unreported source elements that predominantly propagate by generating solo-L1 insertions. **(d)** Circos plot showing all L1 insertions traced to a specific source element in tumour PD0270a. Notably, this includes 149 L1-solo insertions and 12 ambiguous transductions whose origin could not be traced through a transduction-based strategy. **(e)** Circos plot from tumour PD0270a showing that the source L1 element situated at 2q24.1 exclusively propagates through the generation of somatic transductions (n=83, green), ending precisely 234 bp downstream from the L1 locus (right, top panel). Analysis of polyadenylation (PA) signals (right, bottom panel) indicated the lack of a canonical PA signal at the end of the source L1. However, an alternative PA site was identified 234 bp downstream from the 3' end of the L1, aligning seamlessly with the endpoint of the derived transductions. **(f)** Circos plot from tumor PD0270a shows that the source L1 element situated at 1p22.3 primarily resulted in solo insertions (n=18) compared to transductions (n=5). Analysis of APARENT results demonstrated a stronger canonical PA signal at the end of the source L1 element in comparison to the first alternative PA site predicted. Notably, this predicted alternative PA site closely matched the endpoint of the transductions. **(g)** In tumour PD0270a, the source element located at 22q12.1-II, exhibiting a dispersed pattern of transduction endpoints (top panel), demonstrates a close correlation between the observed distribution of transduction endpoints and the intensity of APARENT-predicted PA signals (bottom panel). PA: Polyadenylation signals; L1: Long Interspersed Nucleotide Element 1; SNV: Single-Nucleotide Variant.

Interestingly, this approach revealed that different alleles originating from the same locus often displayed private variants but consistently clustered together, supporting the accuracy of the sequence reconstruction.

Next, we employed the reconstructed L1 repertoire from each patient to associate somatic solo-L1 insertions with their source elements using diagnostic nucleotides. To achieve this, we used a source inference method developed by Martin Santamarina at the University of Santiago de Compostela. Briefly, this method involves the following steps: (i) conducting a multiple sequence alignment of the source element candidates to pinpoint specific diagnostic nucleotide substitutions (**Fig. 8b**); and (ii) genotyping these diagnostic nucleotides within the set of somatic retrotranspositions identified by MEIGA in the paired analysis.

We applied this source element inference algorithm to the genomes of the 10 patients in our cohort, successfully attributing 25.1% (899 out of 3,580) of L1-solo events to specific source elements. In the case of ambiguous transductions, particularly microtransductions, the assignment rate was 19% (38 out of 200). For instance, in the case of HNSC patient PD0270a, the pipeline identified a total of 1,919 diagnostic positions along the L1 consensus sequence. On average, there were 8.6 diagnostic substitutions per candidate source across 261 elements. This analysis enabled us to link 149 derived solo insertions and 12 ambiguous transductions to their respective source L1s (**Fig. 8c, d**).

Notably, these findings address a substantial fraction of the total somatic retrotranspositions whose origin could not be traced through a transduction-based strategy, fundamentally altering our understanding of the landscape of active source L1 elements. Analysing the activity patterns of distinct source L1 elements, we discovered that certain L1 loci primarily propagate by generating somatic transductions, while others predominantly generate solo-L1s. To investigate the causes of such patterns, we utilized the reconstructed sequences of L1 loci and their downstream regions to examine polyadenylation (PA) signals associated with each specific L1 locus. For this analysis, we employed the APARENT neural network[132] to predict PA sites.

In one remarkable example in the HNSC tumour PD0270a, the source L1 element located at 2q24.1 generated 83 derived copies, all of which were transductions with no solo insertions (**Fig. 8e**). Notably, the majority of these transduction events terminated 234 bp downstream from the L1 locus. Examination of the APARENT results revealed the absence of a canonical PA signal at the end of the source L1 element. Instead, the first alternative PA site was predicted to be 234 bp downstream from the 3' end of the L1, aligning perfectly with the endpoint of the derived transductions.

In another example from the same patient, the source element located at 1p22.3 predominantly resulted in solo insertions (n=18) compared to transductions (n=5), with transductions ending 64 bp downstream (**Fig. 8f**). Notably, the analysis of PA signals in this case revealed a robust PA signal at the canonical end of the source element, followed by a weaker PA site at the transductions end. Furthermore, in the case of the source element at 22q12.1-II, which displays a more dispersed pattern of transduction endpoints, we observed that the usage of these endpoints closely correlated with the intensity of the PA signals (**Fig. 8g**). Collectively, this confirms that the strength of canonical and alternative PA signals plays a crucial role in influencing the transduction rate of each source element.

Note: *The source inference method was both developed and applied to our cohort by Martin Santamarina (University of Santiago de Compostela, Spain).*

10.3 THE NOVEL PANORAMA OF RT-MEDIATED STRUCTURAL VARIATION

Somatic retrotransposition can promote chromosomal rearrangements of different types and complexity as a result of an aberrant retrotransposition process. L1-mediated rearrangements have been observed to occur somatically with engineered L1s in cultured human cells[65] and naturally in human tumours[68]. Although preliminary work reported that L1-mediated genomic instability may be relevant, but occasional in cancer, the extent to which this mutational process impacts the structure of the cancer genome remains partially unresolved due to library size constraints of short reads. Here, we used MEIGA to conduct an extensive investigation into the patterns and mechanisms of chromosomal instability mediated by retrotransposition in our long-reads dataset.

10.3.1 MEIGA reveals numerous RT-mediated rearrangements in cancer genomes

Rearrangements mediated by retrotransposition manifest as junctions between two distant breakpoints in the genome with retrotransposon bridges connecting them. After curation, a total of 152 retrotransposition-mediated junctions were identified within our cohort. These junctions were classified into four primary categories, determined by whether they were interchromosomal or intrachromosomal events and the orientation of their breakpoints. Importantly, we did not consider associated copy-number changes in this classification. For intrachromosomal junctions, our naming convention was as follows: '*deletion*' for {+/-}, '*duplication*' for {-/+} and '*inversion*', which encompasses both head-to-head {+/+} and tail-to-tail {-/-} inversions. Interchromosomal junctions were categorised as '*translocations*'.

Deletions-like junctions were the most prevalent category, comprising 43% (n=66) of the total, followed by translocations at 30% (n=45), inversions at 18% (n=27) and duplications at 9% (n=14) (**Fig. 9a**). Notably, these numbers significantly differ from those reported in the PCAWG study[68], where only 90 deletions, one duplication, one translocation and one inversion were detected across 2,954 tumours. This suggests that all categories except deletions were likely underestimated in prior short-read studies.

Breakpoint junctions mediated by retrotransposition are present in all tumour samples, where they showed a considerable inter-patient variability in terms of both numbers and types (median=13.5; range=[3-49]; **Fig. 9b**; **Supplementary Table 4**). In a remarkable LUSC tumour, PD0331a, we identified 49 retrotransposition-mediated junctions, encompassing all four categories and making up 32% (49 out of 152) of the total junctions within the cohort. Despite L1-solo was the most frequent type of bridge (66%, n=101) in the cohort, we found that other types of retrotransposons can link the junction break-ends (**Fig. 9c**). Particularly noteworthy are two junctions mediated by processed pseudogenes, involving the *STK3* and *UBE2V2* genes, as well as an inversion mediated by an *Alu* element.

We observed that 90% of the deletions mediated by retrotransposition in our cohort span regions between 0.1 and 10 kb, closely resembling the distribution observed for deletions in the PCAWG dataset[68] (Wilcoxon rank sum test, p=0.42). However, inversions and duplications did not exhibit specific patterns in terms of size (**Fig. 9d**). Interestingly, this observation implies that distinct mechanisms are likely responsible for the formation of deletions compared to duplications and inversions.

Initially, we found no significant correlation between the number of insertions and other rearrangements mediated by retrotransposition. However, tumours PD0270a and PD0331a appeared as outliers regarding of the number of insertions and rearrangements, respectively. When excluding them, we observed a significant correlation, although it is important to note the limited sample size (Pearson correlation test, R=0.81, p=0.014; **Fig. 9e**).

10.3.1 MEIGA unveils a hidden landscape of balanced rearrangements mediated by retrotransposition

To analyse junctions within a wider genomic context, we further classified retrotransposition-mediated junctions based on the proximity of their breakpoints. Notably, we discovered that

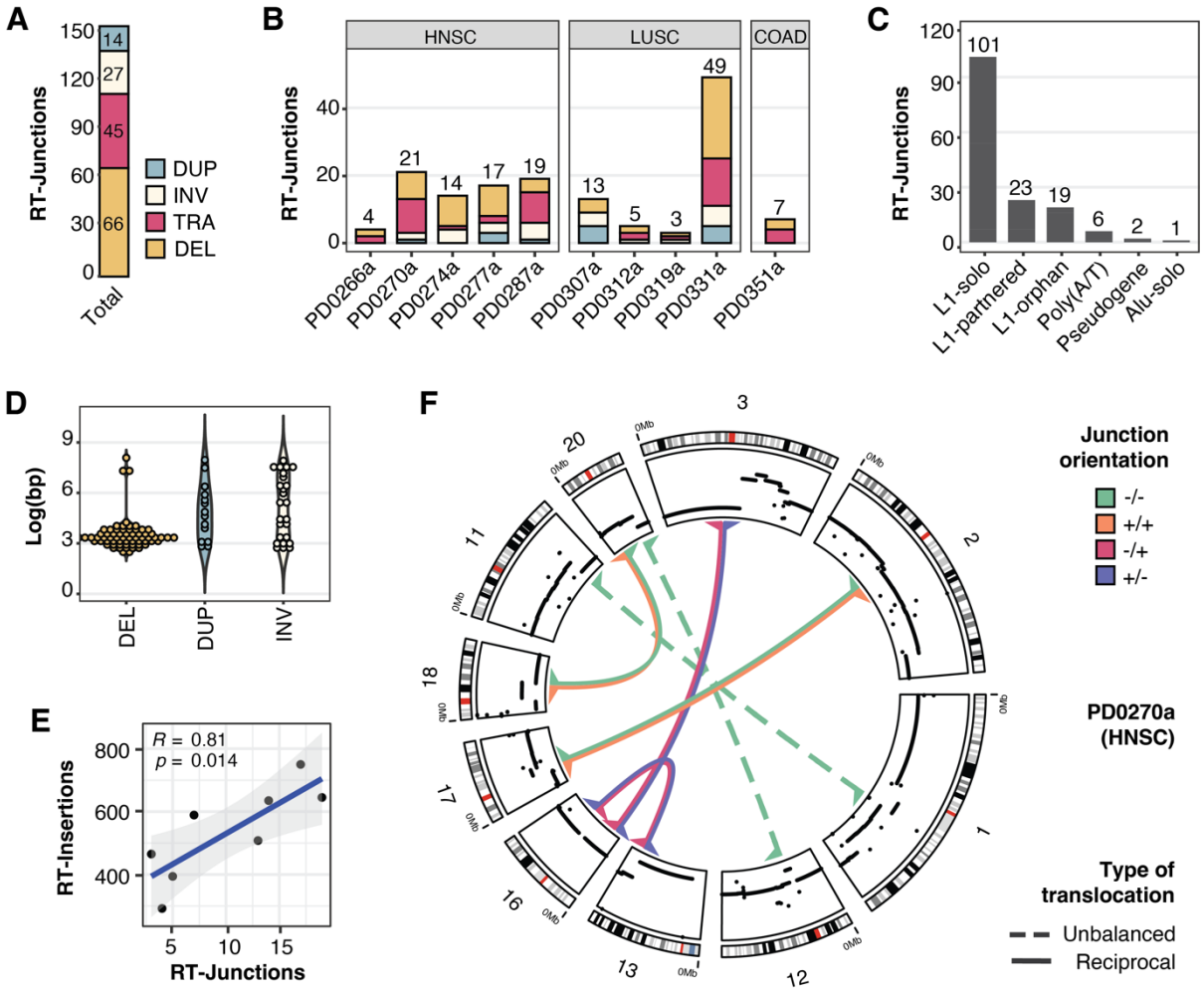


Figure 9: The landscape of RT-mediated structural variation in light of long-reads. **(a)** Total number of somatic breakpoint junctions mediated by retrotransposition (RT-Junctions) in our long-reads tumour cohort. These junctions are categorized into four primary types based on their chromosomal context and their breakpoints orientation. For intrachromosomal junctions, we used the following naming convention: deletions (DEL) for {+/-}, duplications (DUP) for {-/+} and inversions (INV), which encompasses both head-to-head {+/+} and tail-to-tail {-/-} inversions. Interchromosomal junctions were categorised as translocations (TRA). **(b)** Number of somatic RT-Junctions per tumour, with events classified into the junction types as described in (a). **(c)** Number of somatic RT-Junctions based on the type of RT event forming their bridge sequence. **(d)** Size distribution of genomic regions between both sides of the RT-Junctions, with events categorized into the three intrachromosomal types as detailed in section (a). **(e)** Correlation between the number of insertions and junctions mediated by retrotransposition (Pearson correlation test, R=0.81, *p*=0.014). Tumours PD0270a and PD0331a appeared as outliers regarding of the number of insertions and rearrangements, respectively, and were excluded from the analysis. **(f)** Circos plot showing all interchromosomal RT-junctions identified in the HNSC tumour PD0270a. Dashed links represent unbalanced events, while solid links denote reciprocal events. COAD: Colorectal Adenocarcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; L1: Long Interspersed Nucleotide Element 1; LUSC: Lung Squamous Cell Carcinoma; RT: Retrotransposition.

13.2% (20 out of 152) of these junctions involved 13 reciprocal translocations without concurrent copy number changes. Reciprocal translocations are chromosomal rearrangements characterized by a balanced exchange of genetic material between two non-homologous chromosomes, leading to the formation of two derivative chromosomes. In the cases we identified, these derivative chromosomes were linked by retrotransposon bridges at at least one of the two breakpoint junctions.

Importantly, we consistently observed this pattern in five out of the ten analysed tumours, underscoring its prevalence and the necessity for further investigation. For example, in a noteworthy case, tumour PD0270a, we uncovered a total of four reciprocal translocations, along with two unbalanced translocations, all mediated by retrotransposition (**Fig. 9f**). This unexpected discovery highlighted the potential role of retrotransposons in contributing to balanced rearrangements.

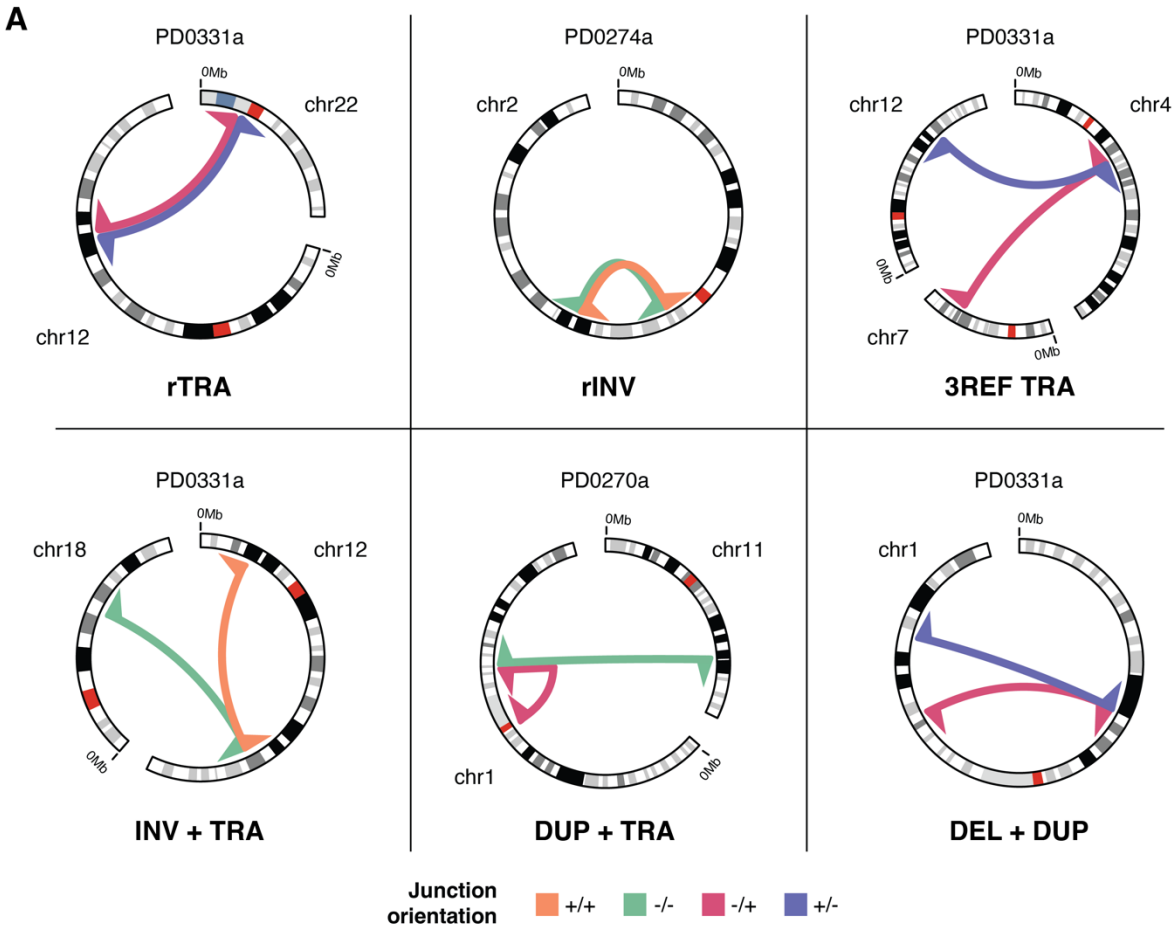
Beyond reciprocal translocations, we observed that retrotransposons could also mediate other forms of reciprocal rearrangements (**Fig. 10a,b)**, such as inversions (n=6), as well as more complex rearrangements resulting from double-stranded DNA breaks (DSBs). For instance, we observed two cases of retrotransposition-mediated translocations consisting of two breakpoint junctions joining 3 different chromosomes (**Fig. 10a,b)**. In these cases, both sides of a DSB were rescued by interchromosomal junctions to two distinct chromosomes. This intricate pattern was observed in two independent samples, tumours PD0287a (HNSC) and PD0331a (LUSC).

Additionally, we observed instances of rearrangements associated with DSBs that resulted in translocations coupled with intrachromosomal junctions, such as inversions (n=3) or duplications (n=1). We also observed an example where two intrachromosomal junctions occurred concurrently, leading to a 31 Mb deletion.

Remarkably, balanced rearrangements mediated by retrotransposition, such as reciprocal translocations and reciprocal inversions, closely resemble canonical retrotransposon insertions when examined using Illumina short-read sequencing (**Supplementary Fig. 3**). This similarity likely contributed to their misclassification in earlier cancer retrotransposition studies.

We experimentally validated our findings in an LUAD cell line, H2009, with high retrotransposition rates. Employing MEIGA, we identified a total of 17 deletions, three inversions, two translocations and one duplication, all attributed to retrotransposition activity. Among the two translocation junctions, we identified a reciprocal translocation between chromosomes 3 and 6 [t(3;6)(q24;q21)], involving two bridges of L1 (**Fig. 11a**). We began by examining available Micro-C data, which served to confirm the presence of the event (**Fig. 11b**).

Additionally, we conducted fluorescence *in situ* hybridization (FISH) using whole-chromosome probes targeting chromosomes 3 and 6 (**Fig. 11c**). The results verified the expected derivative chromosomes along with their estimated copy numbers. Intriguingly, we also noticed the existence of an additional derivative chromosome presumably resulting from a more complex rearrangement involving the derivative t(6:3). Further insights came from a spectral karyotyping (SKY) analysis of this cell line sourced from the Cellosaurus database[133]. The SKY analysis identified two reciprocal translocations, one of which matched the reciprocal



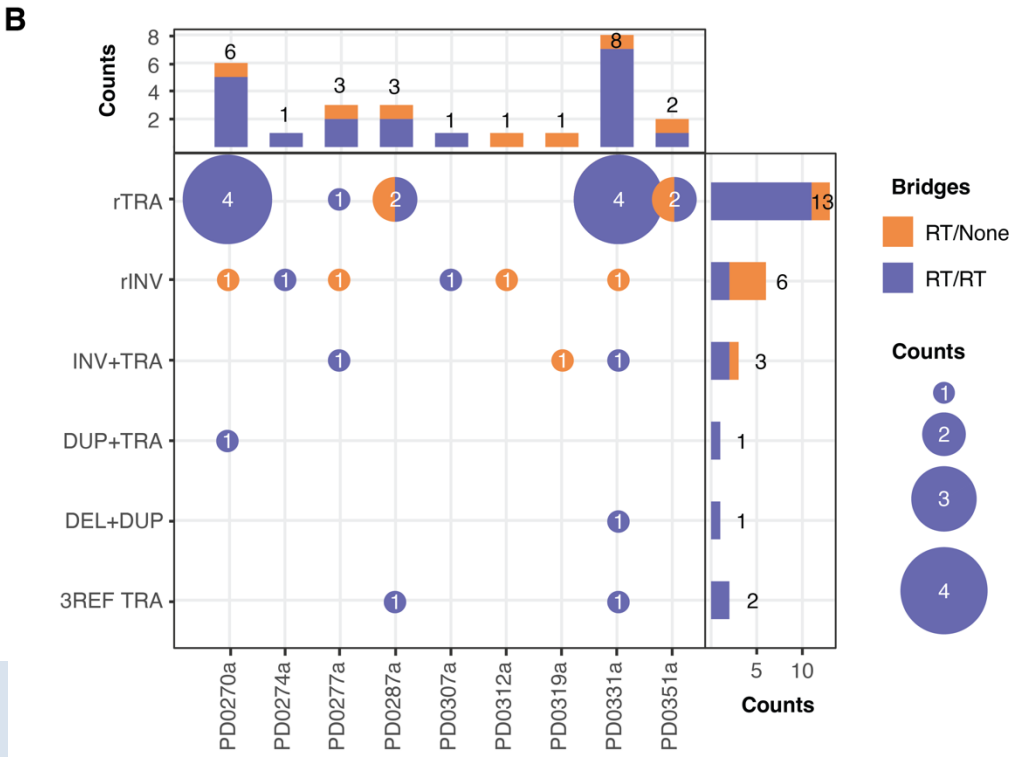


Figure 10. A hidden landscape of balanced rearrangements mediated by retrotransposition. (a) Aberrant retrotransposition events can mediate several forms of balanced rearrangements. These include reciprocal translocations (rTRA) and reciprocal inversions (rINV), as well as more complex rearrangements linked to double-strand breaks (DSBs). Such complex forms encompass scenarios where two interchromosomal junctions connect three different chromosomes (3REF TRA) at a given DSB, and rearrangements leading to translocations coupled to intrachromosomal junctions, such as inversions (INV+TRA) or duplications (DUP+TRA). Additionally, we observed a case where two simultaneous intrachromosomal junctions were used to rescue both sides of a DSB (DEL+DUP). **(b)** Number of RT-mediated rearrangements attributed to each of the six categories mentioned in (a) across the 10 tumours in the cohort. The legend specifies whether both junctions of the balanced rearrangement are characterized by an RT-bridge (RT/RT), or only one of them (RT/None). RT: Retrotransposition.

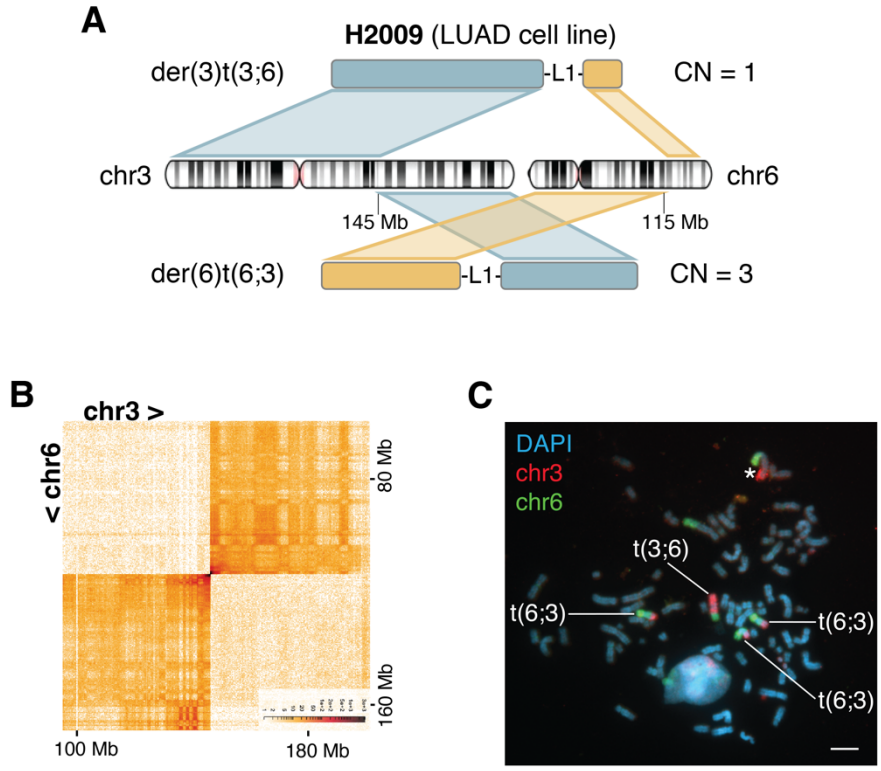


Figure 11. Validation of a reciprocal translocation using the LUAD cell line, H2009. (a) Schematic representation of a reciprocal translocation between chromosomes 3 and 6, noted as t(3;6)(q24;q21) and involving two bridges of L1, identified by MEIGA in the LUAD cell line, H2009. **(b)** Micro-C contacts of the two chromosomes involved in the rearrangement at 15 kb resolution. **(c)** FISH analysis performed on H2009 metaphases using whole chromosome probes. Co-hybridization with probes for chromosome 3 (labelled in red) and chromosome 6 (labelled in green) demonstrated the reciprocal translocation between these chromosome pairs, as well as the aneuploidy of one of the derivative chromosomes, specifically t(6:3). The asterisk indicates a derivative chromosome that results from a more complex rearrangement, likely involving the translocation t(6:3). DAPI staining (blue) was employed as a nuclear counterstain. The scale bar represents 100 µm. FISH: Fluorescence *In Situ* Hybridization; L1: Long Interspersed Nucleotide Element 1; LUAD: Lung Adenocarcinoma.

translocation we reported mediated by L1. Unfortunately, the other reciprocal translocation appeared to be centromeric and could not be located.

Altogether, this raises the question of whether retrotransposons are the primary drivers behind reciprocal translocations. We attempted to estimate the frequency of classical reciprocal translocations (e.g. those not mediated by retrotransposition) using Sniffles[134] on our long-reads dataset, but the results were inconclusive. It remained uncertain whether the absence of classical reciprocal translocations is due to limitations in Sniffles performance or if it accurately reflects their rare occurrence.

As an alternative approach, we assessed their frequencies using the PCAWG SV call set[68]. Observations from the PCAWG dataset indicated that classical reciprocal translocations as well as classical reciprocal inversions are infrequent occurrences within the analysed tumour types (Translocations: median=0, range [0,3]; Inversions: median=0, range [0,5]; **Fig. 12a,b**). Notably, we observed a significantly higher frequency of reciprocal translocations mediated by retrotransposons in our dataset when compared to classical reciprocal translocations in PCAWG (Wilcoxon rank sum test, p=0.0001121; **Fig. 12a**). In contrast, no significant difference was found in the case of reciprocal inversions (Wilcoxon rank sum test, p=0.4999; **Fig. 12b**).

Further analysis of our dataset is essential to comprehensively understand other mechanistic processes responsible for generating reciprocal chromosomal rearrangements. However, it is already clear that retrotransposons notably contribute to the promotion of reciprocal translocations and inversions, thereby playing a significant role in large-scale genomic instability in certain tumours.

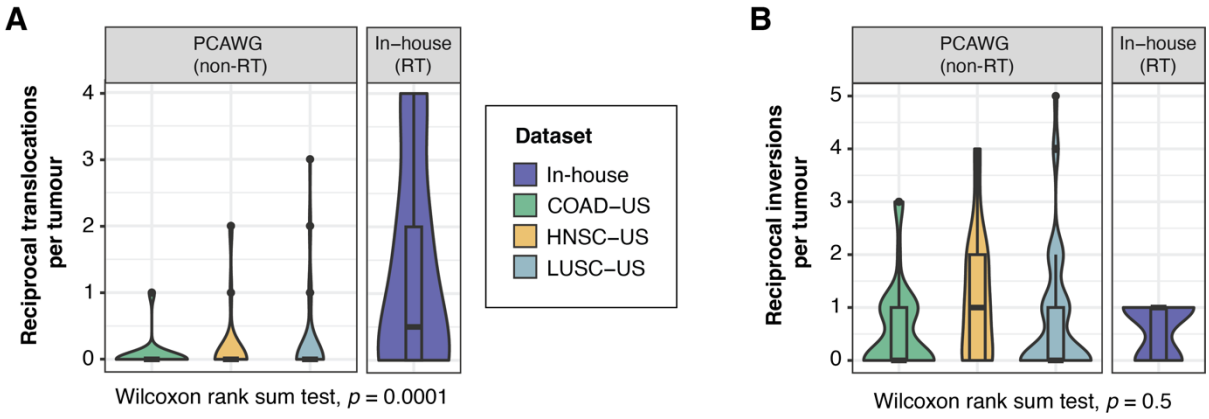


Figure 12. RT-mediated reciprocal translocations in our dataset are more frequent than classical reciprocal translocations in PCAWG. (a) Our dataset showed a significantly higher frequency of RT-mediated reciprocal translocations compared to classical reciprocal translocations (e.g. those not mediated by RT) in PCAWG, across the same tumour types (Wilcoxon rank sum test, p=0.0001121). **(b)** Notably, no significant difference was found in the case of reciprocal inversions (Wilcoxon rank sum test, p=0.4999). COAD: Colorectal Adenocarcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; LUSC: Lung Squamous Cell Carcinoma; PCAWG: Pan-Cancer Analysis of Whole Genomes; RT: Retrotransposition; US: United States.

10.3.2 Unravelling the mechanisms behind RT-mediated balanced rearrangements

We next explored the mechanism of formation of reciprocal translocations mediated by retrotransposition. For example, in the HNSC tumour PD0270a, we identified a translocation between chromosomes 18 and 20, noted as t(18,20)(q12.2;p12.1), involving two bridges of L1 between breakpoint junctions (**Fig. 13a**). The examination of the long-read sequences spanning the breakpoint junctions revealed a singular design with the following hallmarks.

First, each junction of the two derived chromosomes has its own, independent L1, as evidenced by transduced sequences originating from different source elements. Second, in each derivative, the presence of an endonuclease motif along with a poly(A/T) tail confirmed the utilization of the TPRT integration mechanism. Third, the duplications of the target site resulting from the two L1 insertion events are intertwined in such a way that each individual L1 event is flanked by two distinct TSD sequences. Consequently, the two copies of the TSDs are located on different non-homologous chromosomes, rather than being on the same chromosome.

In another case, within the LUSC tumour PD0331a, we detected a reciprocal translocation between chromosomes 12 and 22, noted as t(12,22)(q21.31;p11.2) (**Fig. 13b**). This translocation encompassed two breakpoint junctions involving L1-mediated transductions from different source elements, confirming that two distinct retrotransposon molecules were involved in its formation. We also confirmed the presence of two sets of retrotransposition hallmarks, including TSDs, poly(A/T) tails and endonuclease motifs. However, in this case, we observed that while one of the junction bridges had a poly(A/T) tail and an endonuclease motif at each end, the other bridge lacked these features. Additionally, the interleaved coordinates between the L1 inserts of each bridge suggested a complex pattern of internal inversions.

Twin priming, a variant of TPRT, is the accepted mechanism behind the formation of L1 internal inversions. During twin priming, the L1 endonuclease cleaves the second DNA strand before reverse transcription has been fully completed, resulting in the formation of a second overhang. This overhang subsequently anneals internally to the L1 RNA and primes reverse transcription from what is the point of inversion. Once the RNA is removed from the RNA/cDNA structure, the single-stranded cDNAs pair through microhomology-driven complementarity, forming the inversion junction, and the remaining DNA synthesis is completed. This entire process results in an internal inversion within the L1 insert.

Our data strongly suggests that, in this case, two independent cDNA/RNA structures on non-homologous chromosomes, which had undergone twin priming, were resolved by complementarity pairing, not within their own structure, but between each other. As a result, two adjacent intrachromosomal junctions with interleaved inversion coordinates on their bridges were formed, giving rise to the featured reciprocal translocation (**Fig. 13b**).

In another example, in the LUSC sample PD0331a, we discovered a reciprocal translocation between chromosomes 8 and 22, designated as t(8;22)(q22.2;q13.2) (**Fig. 13c**). Here, we observed that a retrotransposition intermediate that had undergone twin priming was resolved by pairing with another canonical TPRT intermediate structure. Of particular interest, the breakpoints on chromosome 22 overlapped with the *EP300* gene, a known tumour suppressor gene in colorectal cancer, presumably rendering a non-functional transcript.

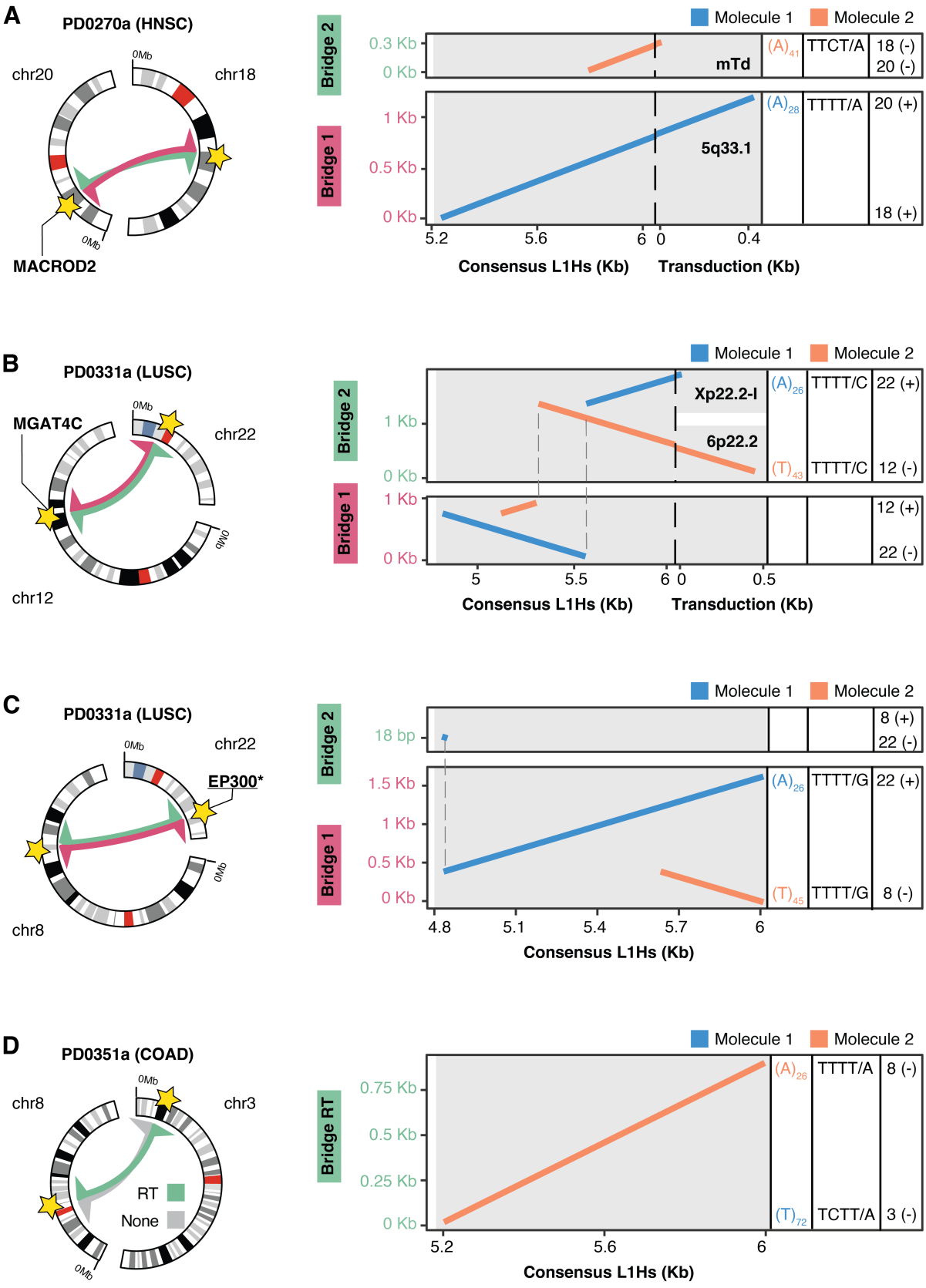


Figure 13. RT-mediated reciprocal translocations resulting from two independent retrotransposition events. (a) On the left, a circos plot shows the reciprocal translocation between

chromosomes 18 and 20, noted as t(18,20)(q12.2;p12.1), in the HNSC tumour PD0270a. Endonuclease motifs are marked with stars, and gene names at the junction breakpoints are indicated. On the right, sequence dot plots illustrate alignments to the L1Hs consensus sequence for both bridges, detailing aspects such as poly(A/T) tail lengths, endonuclease motifs and breakpoint junction orientations. The alignment colours indicate the molecule of origin. This reciprocal translocation was found to result from two independent L1 molecules, inferred from transduced sequences from different source elements and two distinct sets of poly(A/T) tails and endonuclease motifs. **(b)** Similarly, in the LUSC tumour PD0331a, a reciprocal translocation between chromosomes 12 and 22, t(12,22)(q21.31;p11.2), is analysed. Here, one junction bridge showed a poly(A/T) tail and endonuclease motif at each end, while the other lacked these features. Additionally, the interleaved coordinates between the L1 inserts of each bridge suggested a complex pattern of internal inversions. This data indicates that, in this case, two independent cDNA/RNA intermediate structures on non-homologous chromosomes, which had undergone twin priming, were resolved through complementarity pairing between each other, rather than within their own structures. **(c)** Additionally, an analysis in the same LUSC tumour PD0331a reveals a reciprocal translocation, t(8;22)(q22.2;q13.2), involving two independent retrotransposition intermediates, one that had undergone twin-priming and one that was formed through canonical TPRT. Notably, the breakpoints on chromosome 22 intersect with the EP300 gene, implicated as a tumour suppressor in colorectal cancer. **(d)** A similar analysis is shown for a reciprocal translocation t(3;8)(p24.3;q11.23) detected in the COAD sample PD0351a. Here, a retrotransposon sequence is present in only one of the two junctions. Nonetheless, the presence of two poly(A/T) tails and two endonuclease motifs suggests mediation by two independent retrotransposition events. COAD: Colorectal Adenocarcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; L1: Long Interspersed Nucleotide Element 1; LUSC: Lung Squamous Cell Carcinoma; mTD: Microtransduction; RT: Retrotransposition; TPRT: Target-Primed Reverse Transcription.

Significantly, in cases where the retrotransposon sequence is present only in one of the two junctions forming a reciprocal translocation, our findings suggest that such translocations may also be mediated by two independent retrotransposition events, as indicated by the presence of two poly(A/T) tails and two endonuclease motifs. An example of this is observed in a reciprocal translocation between chromosomes 3 and 8 within the COAD sample PD0351a, noted as t(3;8)(p24.3;q11.23) (**Fig. 13d**). In this instance, it appears that two intermediate structures of L1 retrotransposition, generated by canonical TRPT, paired through complementarity between the two L1 cDNAs. As a result, this pairing led to the formation of one junction involving two distinct L1 molecules, while the other junction lacks L1 sequence.

In 10 out of 13 cases, our data collectively support a model where reciprocal translocations seem to result from microhomology-driven annealing between two independent retrotransposition events. According to this model, when the retrotransposition bridges are not inverted, interchromosomal junctions were formed through annealing between the retrotranscribed cDNA of one retrotransposition event and the genomic 3' end of the other event. In cases of inversion, annealing occurred not with the genomic 3' end, but with the inverted cDNA of the other event. Additionally, when both bridges possess a poly(A/T) tail, the pairing took place between two events on the same strand. Conversely, when pairing occurs between events on opposing strands, it leads to the formation of two bridges—one with no poly(A/T) tail and the other with two poly(A/T) tails, one at each end.

However, we also observed three instances of reciprocal translocations where it appears that a single retrotransposition event is employed to simultaneously repair both sides of a DSB located

on separate non-homologous chromosome. This is indicated by the presence of only one poly(A/T) tail and an endonuclease motif within both bridges, as well as the interleaved coordinates between the L1 inserts of each bridge (**Supplementary Fig. 4**). Thus, reciprocal translocations can arise from either one or two independent retrotransposition events, though the latter appears to be the more common mechanism.

As for reciprocal inversions, we also observed that they can be mediated by one or two retrotransposition events. For instance, in the LUSC tumour PD0307a, we identified that a reciprocal inversion involving chromosome 5 is the result of two distinct retrotransposition events (**Fig. 14a**). In contrast, in the HNSC tumour PD0274a, the reciprocal inversion affecting chromosome 2 appears to be caused by a single retrotransposition event (**Fig. 14b**). Significantly, this reciprocal inversion disrupts *LRP1B*, a well-known tumour suppressor gene.

In the case of more complex rearrangements resulting from DSBs, within the LUSC tumour PD0331a, we identified a translocation between chromosomes 12 and 18, coupled with an inversion on chromosome 12 (**Fig. 14c**). Notably, one of the bridges contained a 2,745-bp long L1 insertion, while the other bridge resulted from the integration of a processed pseudogene derived from the *STK3* gene. This discovery offers further evidence that two distinct retrotransposition events are implicated in repairing both ends of a DSB to create a balanced chromosomal rearrangement.

Additionally, we also noted a scenario in which a single L1 retrotransposition event, with an internal inversion, led to the formation of two intrachromosomal junctions (**Fig. 14d**). These junctions resulted in concurrent duplication and deletion events involving two distinct breakpoints, ultimately causing the loss of 31 Mb of genetic material. In this case, the two ends of the inversion junction annealed to two distant genomic regions of the same reference chromosome.

10.3.3 An L1-mediated translocation could have triggered massive chromosomal rearrangements

In the HNSC tumour PD0274a, we identified an unbalanced translocation between chromosomes 10 and 18, mediated by L1 and denoted as t(10;18)(q21.1;p11.32). Significantly, we observed that the breakpoint junction on chromosome 10 was positioned next to a massive chromosomal rearrangement, while the breakpoint junction on chromosome 18 was found to be embedded within this rearrangement (**Fig. 15a,b**). Notably, the massive chromosomal rearrangement closely resembled chromoanasynthesis, a complex rearrangement process dependent on DNA replication.

Chromoanasynthesis is characterized by localized, multiple copy number variations interspaced with copy-neutral chromosomal segments[135,136]. In contrast to chromotripsis, chromoanasynthesis do not stem from chromosomal shattering followed by non-homologous end joining (NHEJ) of the fragments[136]. Instead, two distinct error-prone DNA replication pathways have been suggested to be involved in their formation, including fork stalling and template switching (FoSTeS)[137] and microhomology-mediated break-induced replication (MMBIR)[138,139].

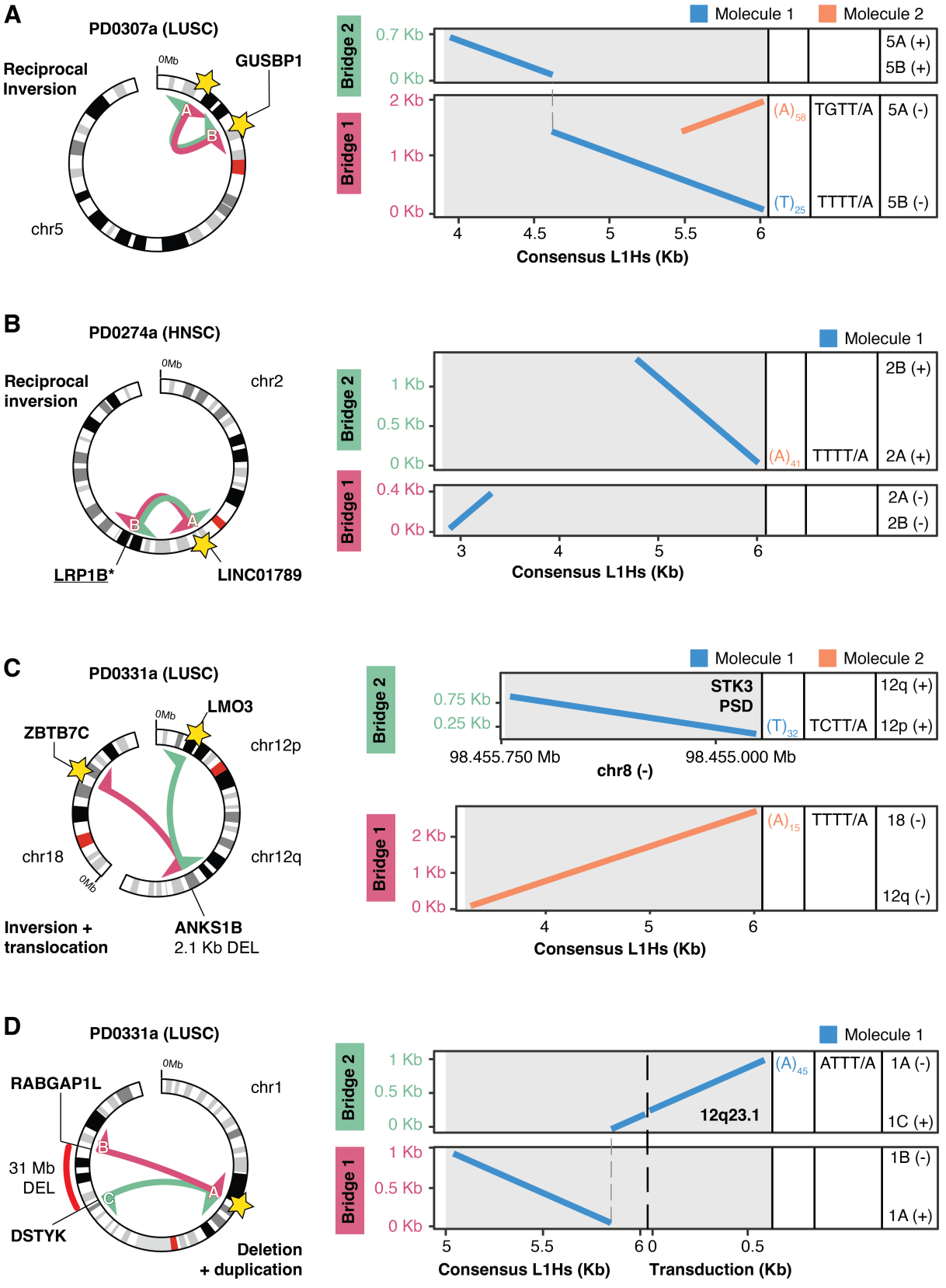


Figure 14. Both bridges of a RT-mediated reciprocal rearrangement can originate from a single retrotransposition event. (a) On the left, a circos plot illustrates a reciprocal inversion within

chromosome 5 in the LUSC tumour PD0307a. Stars mark endonuclease motifs, and gene names at the junction breakpoints are indicated. On the right, sequence dot plots display alignments to the L1Hs consensus sequence for both bridges, detailing aspects such as poly(A/T) tail lengths, endonuclease motifs and breakpoint junction orientations. Alignment colours signify the molecule of origin. Our data collectively shows that this reciprocal inversion results from two separate retrotransposition events. **(b)** A similar analysis depicts a reciprocal inversion involving chromosome 2 in the HNSC tumour PD0274a. This inversion appears to be caused by a single retrotransposition event and notably disrupts *LRP1B*, a known tumour suppressor gene. **(c)** In the LUSC tumour PD0331a, a complex rearrangement involving an inversion on chromosome 12 coupled with a translocation to chromosome 18 is depicted. Here, one bridge features a L1 insertion, while the other results from the integration of a processed pseudogene from the *STK3* gene. This offers further evidence that two distinct retrotransposition events are implicated in repairing both ends of a DSB. **(d)** Furthermore, we observed a case in the same tumour where a single L1 retrotransposition event led to the formation of two intrachromosomal junctions that rescued both sides of a DSB. DSB: Double-Strand Break; HNSC: Head and Neck Squamous Cell Carcinoma; L1: Long Interspersed Nucleotide Element 1; LUSC: Lung Squamous Cell Carcinoma.

The identified pattern prompts the question of whether the L1-mediated translocation was the initiator of the complex chromosomal rearrangement. In this context, it is plausible that the L1-mediated translocation disrupted standard replication processes, leading to replication fork stalling, and consequently set off a series of iterative template switches. Furthermore, given the abundance of L1 elements in our genome, there exists a multitude of potential substrates for homology-mediated template switching.

Further analysis of the microhomologies at the breakpoint junctions is essential to gain a more comprehensive understanding of this intriguing pattern. Additionally, the use of longer reads, which facilitate the reconstruction of the breakpoint graph, would be essential. Notably, understanding the mechanisms behind this event could provide valuable insights into the nature of chromoanasynthesis and reveal a significant role of retrotransposition in triggering extensive genomic instability.

10.1 CHARACTERIZING RETROTRANSPOSITION IN THE CONTEXT OF SHORT READS

To characterize somatic retrotransposition in short-read sequencing data, we took advantage of the code developed for MEIGA and made an implementation for paired-end short-read sequencing data, which we named MEIGA-SR. Although there are numerous methods available for this purpose, we focused on devising an algorithm specifically designed for the precise detection of somatic retrotransposition in sequencing data derived from LCM microdissections of healthy tissues. To this end, our objective was to design a highly sensitive method capable of identifying every sign of somatic retrotransposition within the genomes under investigation, even while relaxing filters commonly applied in methods tailored for cancer genomes. As tumours frequently exhibit high mutation burdens, stringent filters are essential to manage background noise. However, this necessity does not apply to healthy tissues.

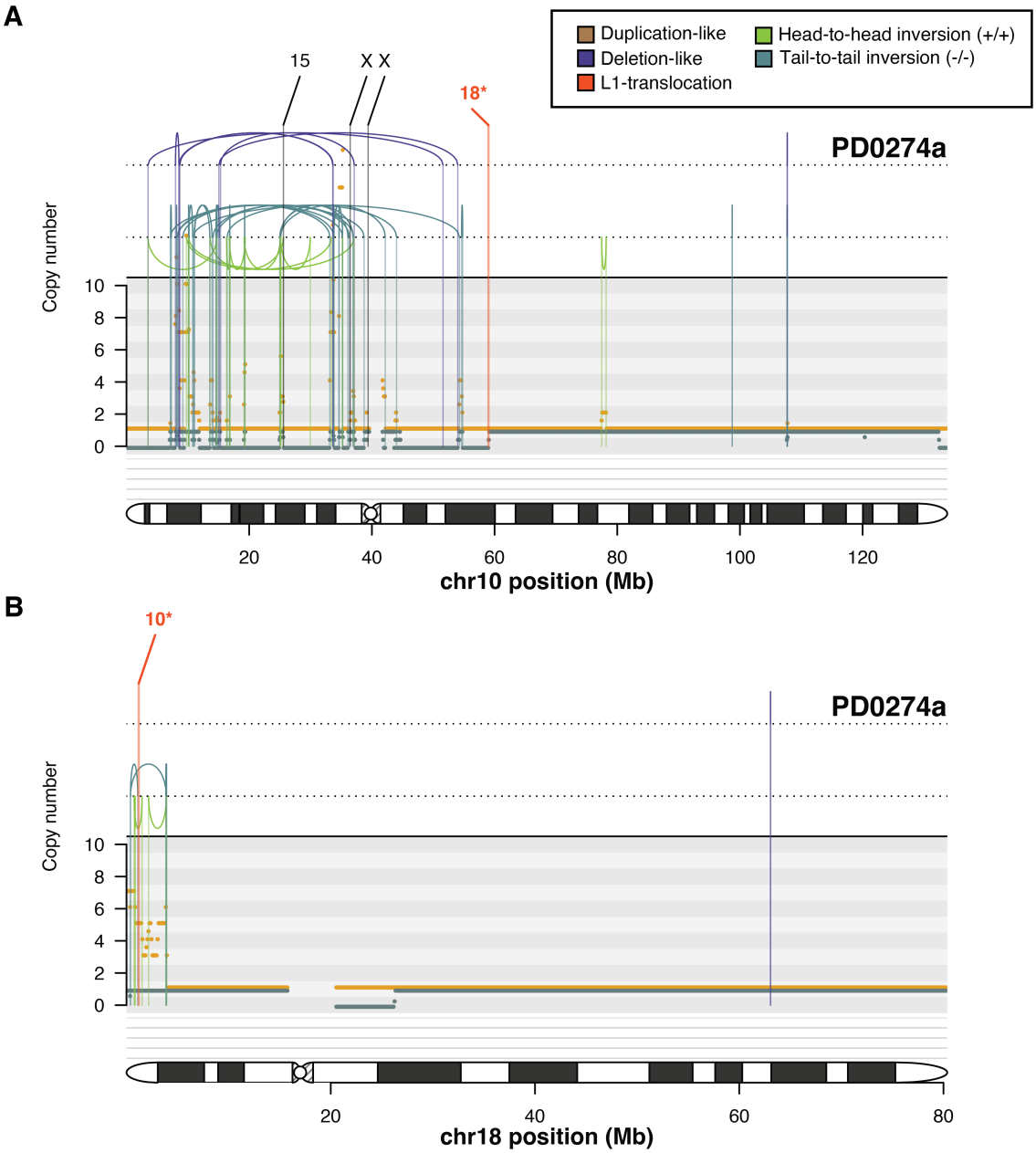


Figure 15. An L1-mediated translocation potentially initiating a massive chromosomal rearrangement in the HNSC tumour PD0274a. This rearrangement involves multiple, localized copy-number alterations affecting chromosomes 10 **(a)** and 18 **(b)**, with noticeable regions of loss of heterozygosity (LOH). In the upper panel, a structural rearrangement graph is displayed. Intrachromosomal rearrangements of all four possible orientations are represented by coloured lines linking DNA segments. Interchromosomal rearrangements are shown with black lines, while those mediated by retrotransposition activity are highlighted in red. The lower panel shows copy-number profiles, as determined by Battenberg analysis of Illumina short-read sequencing data. Notably, an unbalanced translocation t(10;18)(q21.1;p11.32), mediated by L1, has its breakpoint junction on chromosome 10 situated next to the massive chromosomal rearrangement. Conversely, the breakpoint junction on chromosome 18 is embedded within this rearrangement. This pattern raises the possibility that the L1-mediated translocation may have been the trigger for the complex chromosomal rearrangement observed. Further analysis is required to deepen our understanding of this intriguing pattern. L1: Long Interspersed Nucleotide Element 1.

10.1.1 A comprehensive overview of MEIGA-SR

MEIGA-SR accepts as input either a BAM or a CRAM file, specifically derived from Illumina paired-end sequencing with 150-bp reads. While MEIGA-SR is designed to work with inputs aligned onto the human reference genome, assembly GRCh38, it can be easily adapted for other assemblies or species if repeat annotations are available. The MEIGA-SR workflow involves five steps for precisely identifying retrotransposon insertions: (a) Recruitment of supporting reads; (b) Mate annotation; (c) Read clustering; (d) Determining insertion identity; and (e) Applying filtering criteria.

a) Recruitment of supporting reads:

MEIGA-SR initiates the analysis by processing BAM/CRAM alignments from both tumour and matched-normal samples. During this step, MEIGA-SR excludes reads marked as duplicates or with a MAPQ below 20, while retaining discordant reads and clipped reads that may support a retrotransposon insertion. Reads are classified as discordant if their mates do not align to the reference genome with the expected distance or orientation. Clipped reads are identified if they contain a soft or hard-clipped segment larger than 5 bp. Our method defines two alignment segments within clipped reads: the anchor segment, aligning to the region of the reference genome immediately adjacent to the candidate insertion, and the clipped segment, encompassing the inserted sequence.

b) Mate annotation:

In the case of a retrotransposon insertion, discordant reads align near the insertion site, while their paired mates align to regions of the genome containing a retrotransposon from the same family. To recognize this pattern, mate annotation involves assigning an identity to a discordant read based on the retrotransposon repeat to which its paired mate aligns.

To improve the detection of short insertions characterized by clipped reads alone, clipped reads with supplementary alignments undergo a transformation into pseudo-discordant read pairs. This transformation involves configuring the anchor segment and the clipped segment to mimic a pair of discordant mates. Subsequently, both discordant and pseudo-discordant reads receive an identity based on the genomic annotation of their respective mates.

The annotation categories of interest encompass retrotransposon repeats, poly(A/T) repeats, sequences located downstream of source L1 elements (indicative of transductions) and exonic sequences (suggestive of PSD mobilizations). Discordant and pseudo-discordant reads whose mates fall into one of these categories are retained for subsequent analysis, as they hold identities indicative of a potential retrotransposon insertion.

c) Read clustering:

The read clustering process comprises two consecutive steps: primary clustering and meta-clustering. In the primary clustering, by default, a cluster is formed of two reads or more. Discordant reads undergo clustering if the following conditions are met: (i) their mates align onto genomic regions annotated with the same retrotransposon class, (ii) they share the same mapping orientation (i.e., forward or reverse), and (iii) their distance to the nearest read within the cluster is, by default, ≤150 bp. Clipped reads are clustered together if (i) they share the same clipping orientation, and (ii) their distance to the closest breakpoint is, by default, ≤25 bp.

As a result of primary clustering, two types of discordant and clipping clusters are generated, namely forward and reverse, according to their orientation. Then, primary clusters undergo a meta-clustering step to identify pairs of reciprocal forward and reverse clusters closer than 250 bp, by default. Metaclusters undergo a preliminary filtering that requires a minimum of four supporting reads, with none in the normal sample.

d) Determining insertion identity:

In the mate annotation stage, discordant reads were assigned an identity based on the genomic annotation to which its paired mate aligns. Utilizing this information, primary clusters are assigned the predominant identity shared by the reads within each cluster. Subsequently, metaclusters, formed by combining two primary clusters, are designated as solo insertions if both clusters support a solo insertion from the same retrotransposon family. Conversely, those formed by two distinct clusters in which one supports a solo insertion and the other a transduction are designated as partnered transductions. In contrast, metaclusters consisting of two transduction-supporting clusters are categorized as orphan transductions. In addition, the presence of two poly(A/T) clusters leads to the identification of a poly(A/T) insertion. Of note, if a poly(A/T) supporting cluster shares a metacluster with any other identity, the alternative identity prevails. This efficiently captures retrotransposon insertions with long poly(A/T) tails. Metaclusters not fitting in any of these categories are excluded from further analyses.

e) Applying filtering criteria:

A final set of filters is implemented to ensure specificity. Candidate insertions are excluded if they fail to meet any of the following conditions:

i. The number of supporting reads must be between 6 and 500.

ii. The total number of reads in the metacluster region with a MAPQ greater than 20 must exceed 10. Additionally, the median MAPQ of the region must exceed 20.

iii. Conflictive identities assigned to reads within the same metacluster must be below 20%, such as a combination of L1 and *Alu* identities.

iv. If a metacluster is located at a region of the reference genome with a retrotransposon repeat, the repeat identity must differ from that assigned to the metacluster. Alternatively, if the identities match, the repeat must exhibit a milli-divergence relative to its consensus greater than 150, as per the RepeatMasker database[140].

At this stage, candidate calls may undergo additional filters before the method generates a final list of retrotransposon insertions. The configuration of these filters can be finely tuned based on specific objectives of the study. For instance, in the analysis of healthy microbiopsies generated through LCM, two additional filters were applied:

i. The number of discordant or pseudo-discordant reads supporting each metacluster must be greater than 1.

ii. The number of clipped reads supporting a poly(A/T) tail longer than 10 bp must be greater than 1.

These filters were implemented to address sequencing noise specific to LCM libraries, which tend to generate spurious clusters of clipped reads resulting from complex duplication artifacts.

10.1.2 MEIGA-SR exhibits greater sensitivity compared to previous algorithms

Our design incorporates four key features to enhance sensitivity compared to approaches typically employed by previous retrotransposition callers[68,83,91,93]. Firstly, it identifies retrotransposon insertions exclusively supported by clipped reads, often associated with shorter insertions (**Fig. 16a**). Secondly, MEIGA-SR identifies discordant reads supporting poly(A/T) repeats, improving overall read support and enabling recognition of retrotransposon insertions with extended poly(A/T) tails or consisting solely of poly(A/T) tracks. Thirdly, MEIGA-SR requires low support for primary clusters (i.e., two reads), facilitating the identification of events where only one of the two primary clusters on a metacluster receives robust support. Fourth, by relaxing the millidivergence threshold and incorporating additional criteria, such as detecting retrotransposition hallmarks like poly(A/T) and TSD, it better describes nested insertions.

To evaluate the performance of our method, we utilized simulated data. We constructed a mock genome featuring 4,480 retrotransposon insertions representing diverse families and insertion types. This comprised 1000 L1-solo, 500 Alu-solo, 500 SVA-solo, 1240 L1-partnered and 1240 L1-orphan insertions. Subsequently, Illumina paired-end reads, with a length of 150 bp, were simulated to achieve a coverage of 30x and a VAF of 50%. We then applied MEIGA-SR, alongside two reference pipelines, TraFiC and xTea, for the identification of somatic retrotranspositions. All three tools were evaluated using default parameters. Our results demonstrated that MEIGA-SR outperformed the other two tested tools, particularly in terms of recall (**Fig. 16b**), where MEIGA-SR achieved a recall of 95.7%, xTea 83.5% and TraFiC 65%. Precision remained comparable across the three methods, ranging from 99.4% to 99.9%.

Notably, MEIGA-SR demonstrated consistent recall rates across different retrotransposon families and insertion types, ranging from 95.0% to 96.6% (**Fig. 16c**). In addition, the analysis of identity assignment showed that MEIGA-SR performed a precise classification for most insertion types (**Fig. 16c**), with the exception of L1-partnered insertions, which exhibited a tendency to be misclassified as L1-solo insertions (15.9%; 187 out of 1177). We attributed this misclassification to partnered transductions with short transduced regions. Overall, while additional efforts are needed to improve the robustness of MEIGA for broader implementation, MEIGA-SR stands out as a sensitive method for detecting and characterizing retrotransposon insertions, making it a valuable tool for studying somatic retrotransposition in healthy tissues.

10.1.3 MEIGA-SR accurately estimates VAFs of retrotransposition events

To investigate the rate of somatic retrotransposition along different stages of cancer progression, we adapted current timing approaches to the analysis of retrotransposon insertions. Since available timing methods, which were designed for short-read sequencing data, rely on precise VAF estimations, we introduced a genotyper mode within MEIGA-SR specifically designed to assessing allele frequencies of retrotransposition events in tumours.

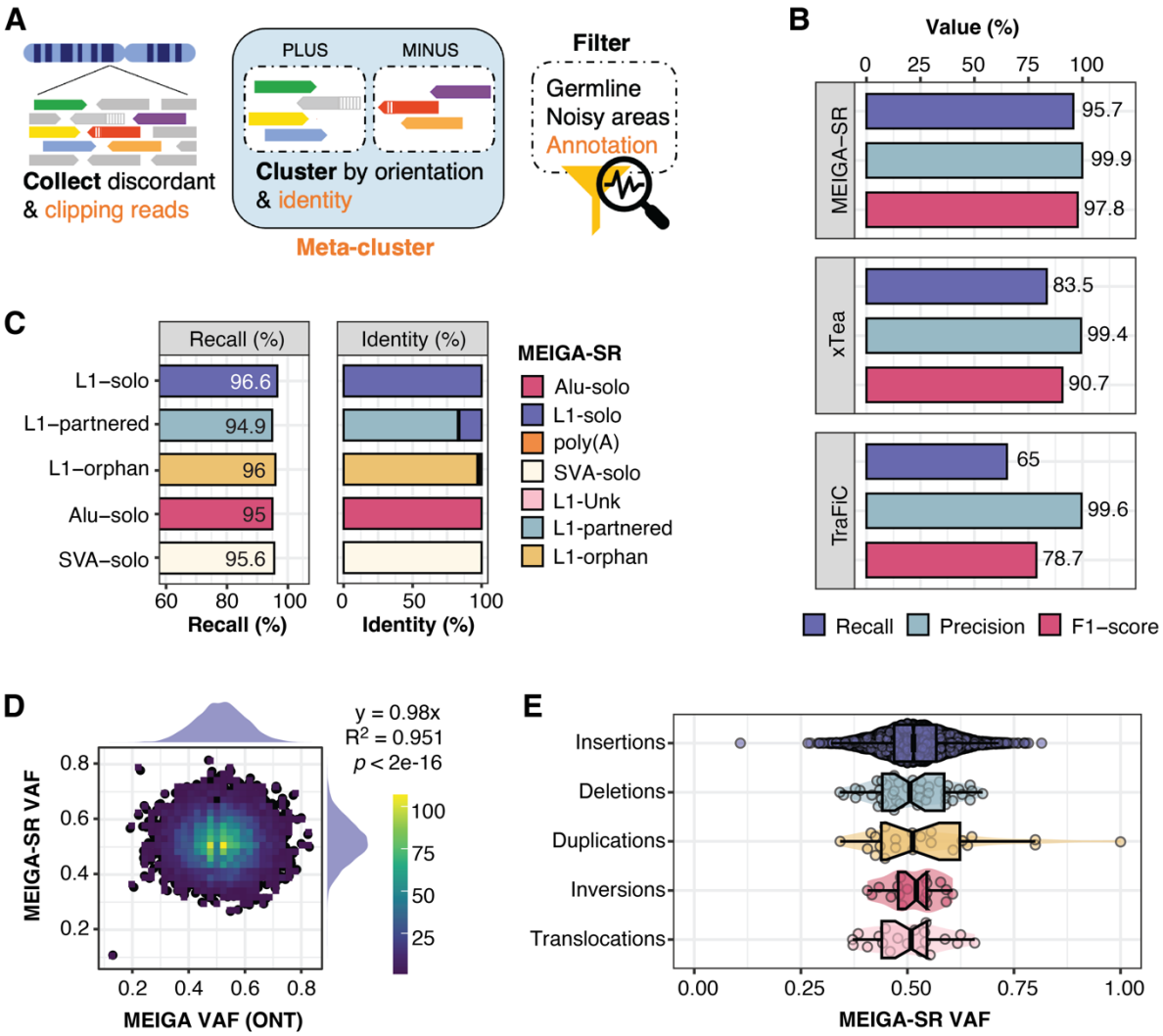


Figure 16. MEIGA-SR demonstrates enhanced sensitivity compared to previous algorithms. (a) MEIGA-SR incorporates four key features, highlighted in orange, to enhance sensitivity over conventional approaches typically used by other retrotransposition callers[68,83,91,93]. Firstly, it identifies RT insertions solely supported by clipped reads, often linked with shorter insertions. Secondly, MEIGA-SR recognises discordant reads that support poly(A/T) repeats, thereby improving overall read support and facilitating detection of RT insertions with extended poly(A/T) tails or those comprised entirely of poly(A/T) tracks. Thirdly, it requires minimal support for primary clusters (i.e., two reads), aiding in identifying events where only one primary cluster in a metacluster receives robust support. Fourthly, relaxing the millidivergence threshold and incorporating criteria such as poly(A/T) and TSD detection aids in the identification of nested insertions. **(b)** In a benchmarking analysis using simulated data, MEIGA-SR was compared against two reference pipelines for detecting somatic RT insertions, TraFiC and xTea. The results showed that MEIGA-SR outperformed the other tools, particularly with regard to recall, where MEIGA-SR achieved a recall of 95.7%, xTea 83.5% and TraFiC 65%. **(c)** Within the same benchmarking analysis, MEIGA-SR maintained consistent recall rates across various retrotransposon families and insertion types. Additionally, the analysis of identity assignment performed by MEIGA-SR showed a precise classification for most insertion types, except for L1-partnered insertions, which were occasionally misclassified as L1-solo insertions (15.9%; 187 out of 1177). **(d)** In an independent benchmarking analysis with simulated data, the accuracy of MEIGA-SR in determining the VAF of somatic RT insertions was evaluated. Complementing this, a long-read sequencing dataset with an

identical set of simulated events, each at a 0.5 VAF, was analysed using MEIGA. The results demonstrated that the estimated VAF distributions for both MEIGA-SR and MEIGA were closely aligned with the expected 0.5 value. Furthermore, the distributions fitted a normal distribution and exhibited a strong linear correlation (y=0.98x; R^2=0.951), indicating no significant bias in either method. **(e)** For simulated RT-mediated rearrangements at 0.5 VAF, MEIGA-SR consistently estimated VAFs centred around 0.5, with no significant differences across the various rearrangement types analysed (Kruskal-Wallis rank sum test, p>0.05). However, the dispersion of VAF estimations was notably higher for duplications (Standard deviations: duplications=0.157, deletions=0.088, inversions=0.061, translocations=0.086, insertions=0.074). L1: Long Interspersed Nucleotide Element 1; RT: Retrotransposition; SVA: SINE-VNTR-*Alu*; TSD: Target Site Duplication; VAF: Variant Allele Frequency.

While tools like SVclone[141] are available for inferring allele frequencies of classical SVs, they encounter limitations when dealing with retrotransposon insertions. Unlike classical SVs, retrotransposon insertions do not exhibit the characteristic formation of two well-defined clusters of discordant pairs—one on each side of the SV junction. Instead, these events are supported by discordant pairs where one read of the pair accurately clusters near the insertion breakpoint, while their mates disperse throughout the genome, mapping to multiple locations where a retrotransposon of the same class is present.

As reference reads precisely match the reference genome, they inherently possess a higher likelihood of alignment compared to alternate reads supporting an SV. Consequently, when computing allele frequencies of retrotransposition events, there is a risk for reference reads to be overrepresented. Using MEIGA-SR, we developed a targeted genotyping strategy to effectively address this challenge.

In our approach, MEIGA-SR thoroughly examines reads within defined intervals both upstream and downstream of the event designated for genotyping. The interval sizes are established based on the median fragment size and the read length of the libraries used in the study (e.g., Median fragment size (m) ~450 bp; Read length (r) = 150 bp; m – r / 2 = 325 bp). The upstream interval extends from the end of the TSD to 325 bp upstream, while the downstream interval spans from the beginning of the TSD to 325 bp downstream. Within these intervals, MEIGA-SR analyses read pairs and categorizes them based on whether they serve as reference or alternate supporting read pairs.

MAPQ filtering is specifically implemented for reference-supporting pairs, setting a minimum requirement of 30, considering their tendency to yield higher MAPQ values. Notably, for both reference and alternate categories, forward reads are excluded if they align after the beginning of the TSD, and reverse reads are constrained from ending before the end of the TSD. In such instances, discerning whether these reads represent reference or alternate supporting reads is challenging. Following the classification of read pairs, MEIGA-SR utilizes the maximum count of observed alternate read pairs within the upstream and downstream clusters to estimate allelic frequency. This approach is adopted due to the tendency of the 3' end of retrotransposition events to exhibit clusters with lower support.

We next evaluated the precision of MEIGA-SR in estimating the VAF of somatic retrotransposition events using a simulation. We simulated short-read sequencing data comprising 6,420 retrotransposon insertions at a 0.5 VAF with a sequencing depth of 30x. In

this simulation, we precisely replicated the set of retrotransposon insertions previously used to evaluate MEIGA on long-reads (Section: '10.1.2. MEIGA outperforms previous retrotransposition callers').

To offer a more comprehensive perspective, we included both short and long-read simulated data in this benchmarking analysis, each analysed with MEIGA-SR and MEIGA, respectively. Utilizing their genotyping modes, both methods were employed to assess the VAFs of the simulated events. Notably, genotyping was successfully accomplished for 98.57% of the events (6,328 out of 6,420) across both sequencing technologies. Subsequently, we applied a set of rigorous filtering criteria to each detected insertion:

i. Minimum MAPQ: 50

ii. TSD length: 0 to 150 bp

iii. Minimum coverage at insertion: 20x

We deliberately implemented these rigorous filters to reduce potential noise in the subsequent timing analysis. As a result, 89.63% (5,754 out of 6,420) of the events were retained for subsequent analysis. Our results showed that the distributions of estimated VAFs, for both Illumina and ONT datasets, were centred around the expected value of 0.5 (**Fig. 16d**). Additionally, the observed dispersion fitted a normal distribution in both cases, consistently falling within the 97.5% margin of the expected VAF distribution. Overall, these results confirmed that both MEIGA and MEIGA-SR precisely estimate the VAF of retrotransposon insertions in long-read and short-read sequencing data, respectively.

Estimating allele frequencies for retrotransposon-mediated rearrangements is notably more challenging than for canonical insertions, with complexity varying depending on the type and size of the genomic rearrangement. While canonical insertions typically involve estimating the reference allele counts at the insertion site, in rearrangements, the two breakpoints of the event may contain varying copy numbers of the reference allele. Duplications, in particular, introduce added complexity to the analysis since the rearrangement process duplicates both reference and alternate alleles. Consequently, the precision of allele frequency estimations is expected to be lower for rearrangements compared to insertions of retrotransposons.

To evaluate the performance of our approach, we applied our method to a simulated dataset that included 20 deletions, 20 duplications, 20 inversions and 10 translocations. All these events were mediated by L1, simulated at 30x coverage and with a VAF of 0.5. The results demonstrated that our method consistently estimated VAFs centred around 0.5, with no significant differences observed across the different rearrangement types analysed (Kruskal-Wallis rank sum test, $p>0.05$; **Fig. 16e**). However, the dispersion of VAF estimations appeared to be notably higher for duplications (Standard deviations: duplications=0.157, deletions=0.088, inversions=0.061, translocations=0.086, insertions=0.074). This can be attributed to the fact that duplications double the number of reference reads, introducing greater variability in the inferences. As rearrangements involve multiple breakpoints, we restricted our analysis to events where all breakpoints shared consistent relative timing categories within the subsequent timing analysis, effectively mitigating the impact of VAF estimates with higher levels of dispersion.

10.2 INSIGHTS INTO THE TIMING OF SOMATIC RETROTRANSPOSITION DURING TUMOUR EVOLUTION

The genome of a cancer cell represents a record of the somatic mutations accumulated over time in the clonal lineage from which it originates. In principle, each mutation arises stochastically on a single chromosome within a single cell, giving rise to a lineage of cells sharing that same mutation. If, at a later stage, the chromosomal region containing that mutation is duplicated, any mutation present on the allele before the duplication will subsequently exist on both resulting allelic copies. Conversely, mutations occurring after the duplication event or on the alternate allele will not exhibit such a duplicated pattern.

Using sequencing data, we can precisely determine the allelic frequency of each specific mutation, allowing us to classify mutations as either early or late clonal variants. Early variants precede copy number gains, while late variants succeed them. Additionally, there is a subset of mutations referred to as unspecified clonal variants, denoted as clonal NA (Not Assigned), which are present in all cancer cells but cannot be further classified as either early or late variants. Lastly, subclonal mutations emerged at later stages in the tumour development and are exclusively present within a subset of cells in the tumour sample, revealing intratumour heterogeneity.

10.2.1 Consistent relative timing estimates of retrotransposition events

Our timing approach employed the estimated allele frequencies from MEIGA-SR, along with information on tumour purity and copy number states from Battenberg, as inputs for mutationTime.R. This enabled us to obtain relative timing estimates for insertions and other rearrangements mediated by retrotransposition. To validate our strategy, we used patient sequencing data. Specifically, we conducted short-read sequencing across six distinct tumour regions from patient PD0270: PD0270a, PD0270a_6, PD0270a_7, PD0270a_10, PD0270a_12 and PD0270a_14.

In our multi-region analysis, we assessed the consistency of relative timing estimations among these different regions. We identified a conflict when the same retrotransposon insertion was labelled as clonal early and clonal late, or as clonal early and subclonal in different regions. Our results revealed that conflicting timing labels were found in 12.39% of cases across at least one of the six sequenced regions (**Fig. 17**). Nevertheless, we achieved an accuracy rate of 97.94% when evaluating our ability to correctly assign the most probable timing label to an individual region (**Fig. 17**), thus demonstrating the robustness of our methodology.

10.2.2 Retrotransposition is active early in tumorigenesis

We proceeded by examining the entire set of 6,266 somatic insertions and 152 rearrangements mediated by retrotransposition identified in the 10 donors of our ONT cohort. These events underwent allele frequency inference and were subsequently included in the timing pipeline, as outlined in the Methodology section. This approach allowed us to unequivocally characterize the relative timing of 4,644 out of 6,266 events (74.12%). Although the analysis showed remarkable interindividual variability, we observed that *clonal early* was the most frequent

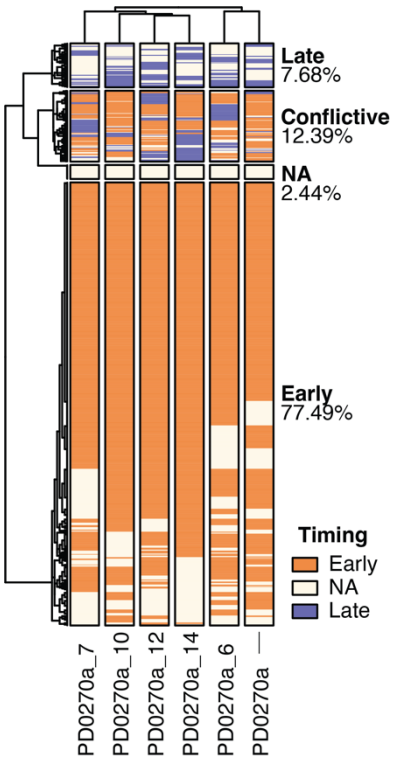


Figure 17. Consistent relative timing estimations across multiple samples. Heatmap showing the relative timing estimates across multiple samples from donor PD0270a. Gower's distance was used to cluster the samples and the timing labels. Timing labels 'Clonal Late' and 'Subclonal' were grouped as 'Late,' while 'Clonal Early' events were categorized as 'Early.' Any instances where 'Early' and 'Late' labels were assigned to the same event in different samples were considered as 'Conflictive'. NA refers to clonal *Not Assigned* events.

category for all tumours except for donor PD0287a, where most events were assigned to a clonal late stage (Ranges (%): *clonal early*=[27.4, 77.9], *clonal late*=[0.7, 34.8], *clonal NA*=[0, 35.1], *subclonal*=[0.8, 14.6]; **Fig. 18a**). It is important to note that our approach is biased towards detecting clonal events, as mutations are more easily detectable when they display higher clonality. Consequently, the percentages of timing categories are not directly comparable within a single sample. Nevertheless, they can, theoretically, be compared across different samples.

Overall, our analysis notably unveiled that 2,975 retrotransposition insertions (64.1% of 4,644) occurred early in the initial stages of tumour development, while 692 (14.9%) were clonal late events, 674 (14.5%) fell into the category of clonal unassigned, and 303 (6.5%) were designated as subclonal events occurring in the later stages of tumorigenesis (**Fig. 18b,c**). This suggests that somatic retrotransposition is not a consequence of the chaotic genomic environment characteristic of later stages but instead a mutational process that can be highly active during the early stages of tumour development. Moreover, despite the decreased sensitivity of our method in detecting late, subclonal events, it also appears that somatic retrotransposition remains an ongoing mutational mechanism throughout the course of cancer progression.

10.2.3 Somatic retrotransposition is active years before tumour diagnosis

Whole-genome doubling (WGD) is a genetic event where a cell acquires an additional set of chromosomes, resulting in the duplication of the entire genome. This phenomenon is commonly observed in cancer genomes[142,143] and arises due to errors in cell division. Importantly, WGD promotes chromosomal instability and tumour heterogeneity, often playing a role in cancer progression and therapy resistance[142,144–147]. WGD events were observed to typically occur during an intermediate stage of tumour development across cancer types, with TP53 mutations

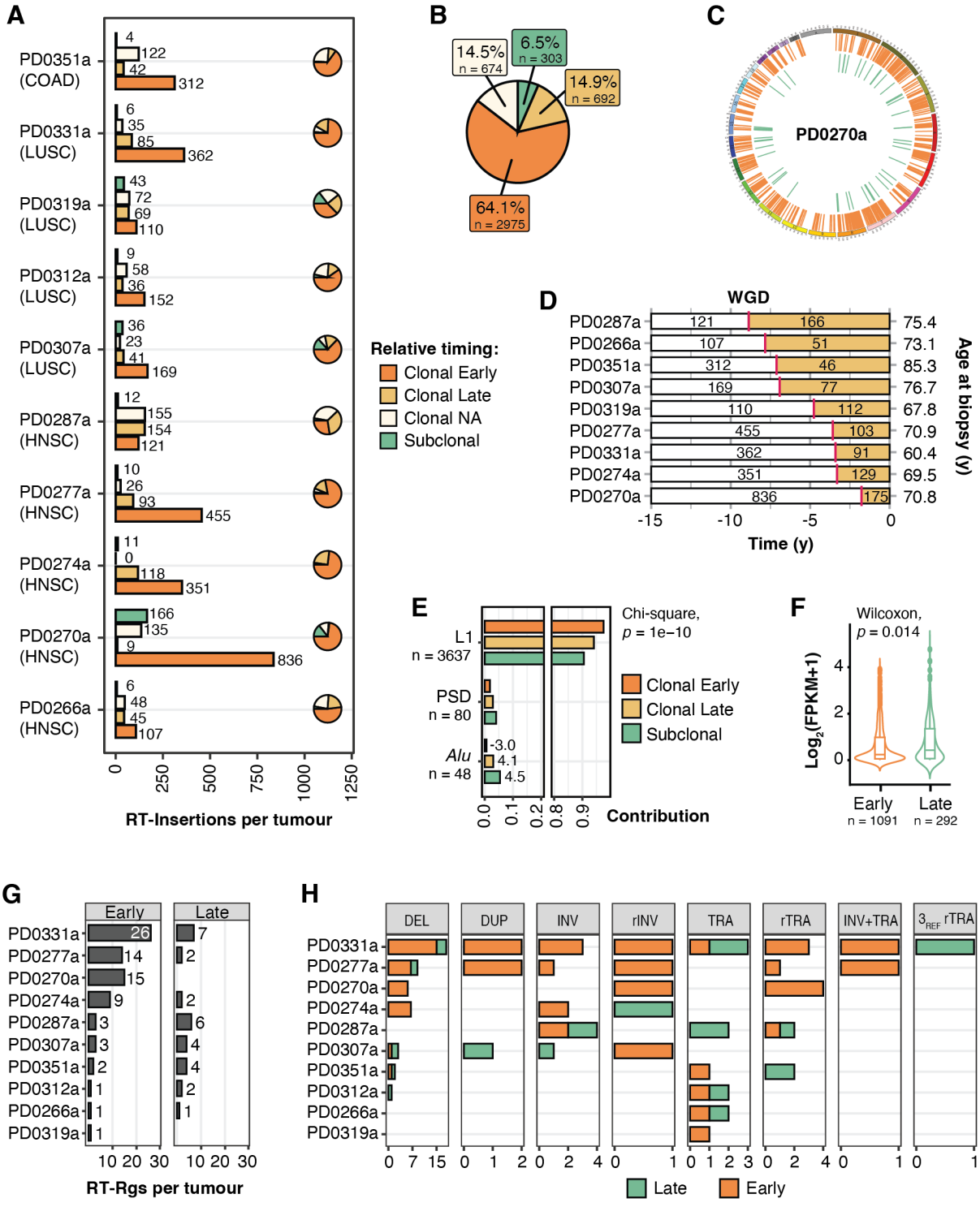


Figure 18. The tempo of somatic retrotransposition in human cancer. (a) Number of somatic retrotransposition insertions assigned to each relative timing category, including clonal early, clonal late, clonal NA (Not Assigned) and subclonal. Clonal early was the most frequent category for all tumours except for donor PD0287. **(b)** Overall percentage of somatic retrotransposition events attributed to each of the relative timing categories mentioned in (a) across the same set of 10 tumours. Notably, up to 2,975 insertions were identified as occurring early, underscoring that somatic retrotransposition can be highly active during the early stages of tumour development. **(c)** circos plot

illustrating the genomic distribution of events categorized by their relative timing in the HNSC tumour PD0270a. Timing labels 'Clonal Late' and 'Subclonal' were grouped as 'Late,' while 'Clonal Early' events were categorized as 'Early' for simplicity. **(d)** Real-time timing estimation of retrotransposition insertions over patients' lifetime in relation to whole-genome doubling (WGD) events. The X axis illustrates the time intervals when somatic retrotransposon insertions occurred relative to the WGD event – before (white) and after (yellow). Pink arrows indicate the occurrence of WGD events in years before the biopsy, and the numbers within the time intervals represent the count of insertion events. The numbers at the end of the timeline denote the age of the patient at biopsy. Remarkably, all tumours displayed over 100 somatic retrotransposon insertions occurring prior to the WGD event. **(e)** Contribution of different retrotransposon families throughout tumour development. This analysis revealed a tendency for the mobilization of *Alu* elements during the later stages of tumorigenesis (Pearson's Chi-squared test, $p<0.0001$, significant residuals: *Alu*-Clonal early=-3, *Alu*-Clonal late=4.1, *Alu*-Subclonal=4.5). **(f)** The frequency of somatic retrotransposon insertions within active genes (FPKM > 0) was examined based on the timing categories stated in (b). This analysis utilized the mean gene expression specific to each cancer type from the TCGA expression dataset. Notably, insertions were found to be less frequent in active genes within the 'Early' category (Wilcoxon rank-sum test, $p < 0.014$), suggesting a signal of negative selection against retrotransposon insertions within active genes. **(g)** Number of retrotransposon-mediated rearrangements per tumour classified into the categories 'Early' and 'Late' as described in (b). **(h)** Same as in (g), but the number of retrotransposon-mediated rearrangements is depicted across different types of rearrangements. COAD: Colorectal Adenocarcinoma; DEL: Deletion; DUP: Duplication; FPKM: Fragments Per Kilobase of sequence per Million mapped reads; HNSC: Head and Neck Squamous Cell Carcinoma; INV: Inversion; L1: Long Interspersed Nucleotide Element 1; LUSC: Lung Squamous Cell Carcinoma; rINV: Reciprocal Inversion; rTRA: Reciprocal Translocation; WGD: Whole-Genome Doubling.

occurring earlier and most copy number alterations happening later[143].

Real timing, quantified in years, of WGDs can be deduced by assessing the copy number of clock-like mutations* (CpG>TpG). Mutations that originated before the gain event become duplicated as part of the genome, while mutations that occur after it remain in single copy. Consequently, the ratio of single-copy to double-copy mutations provides an estimation of when the WGD was acquired in terms of mutational time. By assuming a linear acceleration in the CpG>TpG mutation rate before diagnosis, we can obtain real-time estimations spanning the patients' lifetimes and identify early clonal events occurring prior to WGDs.

To gain deeper insights into the timing of somatic retrotransposition during cancer progression, we conducted real-time estimations of WGDs in our 10 tumour samples. It is worth noting that our timing analysis may introduce a bias towards later WGD events due to our limited ability to detect subclonal mutations. Nevertheless, this analysis provides us with the confidence to assert that WGD occurred at or before the indicated time. While all tumours exhibited WGDs, the timing of the WGD event in sample PD0312a could not be accurately determined. This sample showed a modal major copy number state greater than two, suggesting the occurrence of multiple WGDs that could not be dated using this method.

Our analysis determined that WGDs occurred at a median time of 4.77 years before the biopsy in the studied cohort, with a range spanning from 1.77 to 8.87 years (**Fig. 18d**). Notably, all

* Clock-like mutations refer to genetic alterations that accumulate at a relatively constant and predictable rate over time.

tumours exhibited more than 100 somatic retrotransposon insertions occurring before the WGD event. In a relevant tumour, PD0270a, we identified up to 836 retrotransposon insertions occurring prior to the WGD, dating back to at least 1.77 years before the biopsy, and 175 after the WGD. On the other hand, tumour PD0287a exhibited 121 retrotransposon insertions occurring at least 8.87 years prior to the biopsy and 166 after that time point. Hence, somatic retrotransposition constitutes an ongoing mutational process occurring years prior to diagnosis.

10.2.4 Retrotransposition patterns change along tumour evolution

Next, we proceeded to examine the contribution of different retrotransposon families throughout tumour development. This analysis revealed a tendency for the mobilization of *Alu* elements during the later stages of tumorigenesis (Pearson's Chi-squared test, $p<0.0001$, residuals for *Alu*: clonal early=-3, clonal late=4.1, subclonal=4.5; **Fig. 18e**). A similar pattern was observed for the category of processed pseudogenes, albeit without statistical significance. These findings collectively suggest that in the early stages, the L1 retrotransposition machinery exhibits a strong preference for cis-insertions involving their own RNA. However, this preference shifts during tumour development, becoming more permissive and facilitating the trans-mobilization of *Alu* elements and nuclear mRNAs.

To gain insights into how evolutionary pressures shape the retrotransposition landscape, we investigated potential correlations between retrotransposition events and various genomic features at different tumour stages. Our analysis found no significant differences in the distribution of somatic retrotransposon insertions within genic regions between early and late stages. (Fisher's Exact Test, $p=0.8791$; **Supplementary Fig. 5**). However, when we exclusively examined the frequency of somatic retrotransposon insertions within active genes (FPKM>0), we observed an apparent signal of negative selection against retrotransposon insertions within active genes (Wilcoxon rank-sum test, $p<0.05$; **Fig. 18f**). For this analysis, we utilized mean gene expression values specific to each cancer type from the TCGA expression dataset. In summary, this data suggests that previously observed negative associations between L1 retrotransposition rates and features of active transcription in tumours do not solely result from integration preferences but are also influenced by purifying selection processes.

Considering that a majority of disease-causing L1 insertions occur in the sense orientation relative to the disrupted gene[25,148], we also examined the orientation of L1 retrotransposons somatically inserted within genes. However, this analysis did not reveal any significant difference between sense and antisense preferences during early and late stages of tumour development (Fisher's Exact Test, $p=0.5054$). As noted by Sultana et al.[149], it appears that the bias in L1 orientation within genes arises from an uneven distribution of L1 target site motifs between the transcribed and non-transcribed strands. This imbalance significantly skews the orientation during integration, and this effect may be further reinforced by purifying selection after integration. In our analysis, we cannot definitively conclude whether the orientation effect is too subtle to be detected within our dataset, or if antisense L1 insertions may also be subject to negative selection, given their potential to disrupt gene transcription or translation processes.

10.2.5 Retrotransposons mediate large-scale genomic rearrangements early in tumorigenesis

The timing of retrotransposon-mediated rearrangements revealed that as many as 75 rearrangements corresponded to early events, whereas 28 were identified as late events, out of a total of 103 successfully timed instances (**Fig. 18g**). Notably, we observed that large-scale chromosomal rearrangements, like nine reciprocal translocations, occurred early in tumorigenesis (**Fig. 18h**). Remarkably, in tumour PD0827a, a reciprocal translocation was identified to have occurred at least 8.8 years prior to the tumour biopsy (CI range=[7.1,22.2]). Additionally, we also identified eight unbalanced translocations, eight unbalanced inversions, four reciprocal inversions and two translocations coupled to an inversion as early events mediated by retrotransposon activity.

As for events in later stages, reciprocal translocations can also arise late in tumour development, as evidenced in tumours such as PD0351a (**Fig. 18g,h**). Additionally, we observed late events with potential impact on the cancer genome, including seven unbalanced translocations, three unbalanced inversions, three reciprocal translocations, one reciprocal inversion and one complex translocation affecting three chromosomes. Overall, this indicates that large-scale chromosomal rearrangements mediated by retrotransposition initiate early in tumours and persist throughout cancer evolution.

For example, in a relevant LUSC tumour, PD0331a, we found up to 26 rearrangements occurring early in tumour development (**Fig. 18g,h**), including 15 deletions, five translocations, four inversions and two duplications, all of which have the potential to significantly impact the cancer genome. For instance, a reciprocal translocation within this set affected the tumour suppressor *EP300*, potentially resulting in a non-functional transcript. Notably, the real-time estimation of the WGD event for this patient indicated that all these rearrangements occurred at least 3.4 years before the tumour biopsy (CI range=[2.5,5.8]).

In another relevant HNSC tumour, PD0274a, we identified the unbalanced translocation that could have triggered the catastrophic cellular event of chromoanasynthesis was classified as an early event, occurring at least 3.3 years prior tumour biopsy (CI range=[2.5,6.3]). In addition, also in tumour PD0274a, we timed the reciprocal inversion affecting the tumour suppressor *LRP1B* as a late event, presumably providing a selective advantage later in tumorigenesis. Remarkably, in the case of HNSC tumour PD0270a, we noted the occurrence of four reciprocal translocations mediated by retrotransposons in the early stages of tumour development. Although the impact remains uncertain, these events have the potential to significantly disrupt functional domains within the genome.

All in all, these findings further support the notion that somatic retrotransposition, through aberrant integration events resulting in genomic rearrangements, can actively contribute to tumour instability and promote tumour progression, rather than being a mere consequence of the later chaotic tumour environment.

10.2.6 Dynamic evolution of source elements activity throughout tumour development

The landscape of active source L1 elements contributing to the total burden of somatic retrotransposition throughout tumour development remained largely uncharacterized. Various models could elucidate their behaviour (**Fig. 19a**). One possibility is that the number of active source elements escaping host control increases over time. Conversely, it is plausible that host

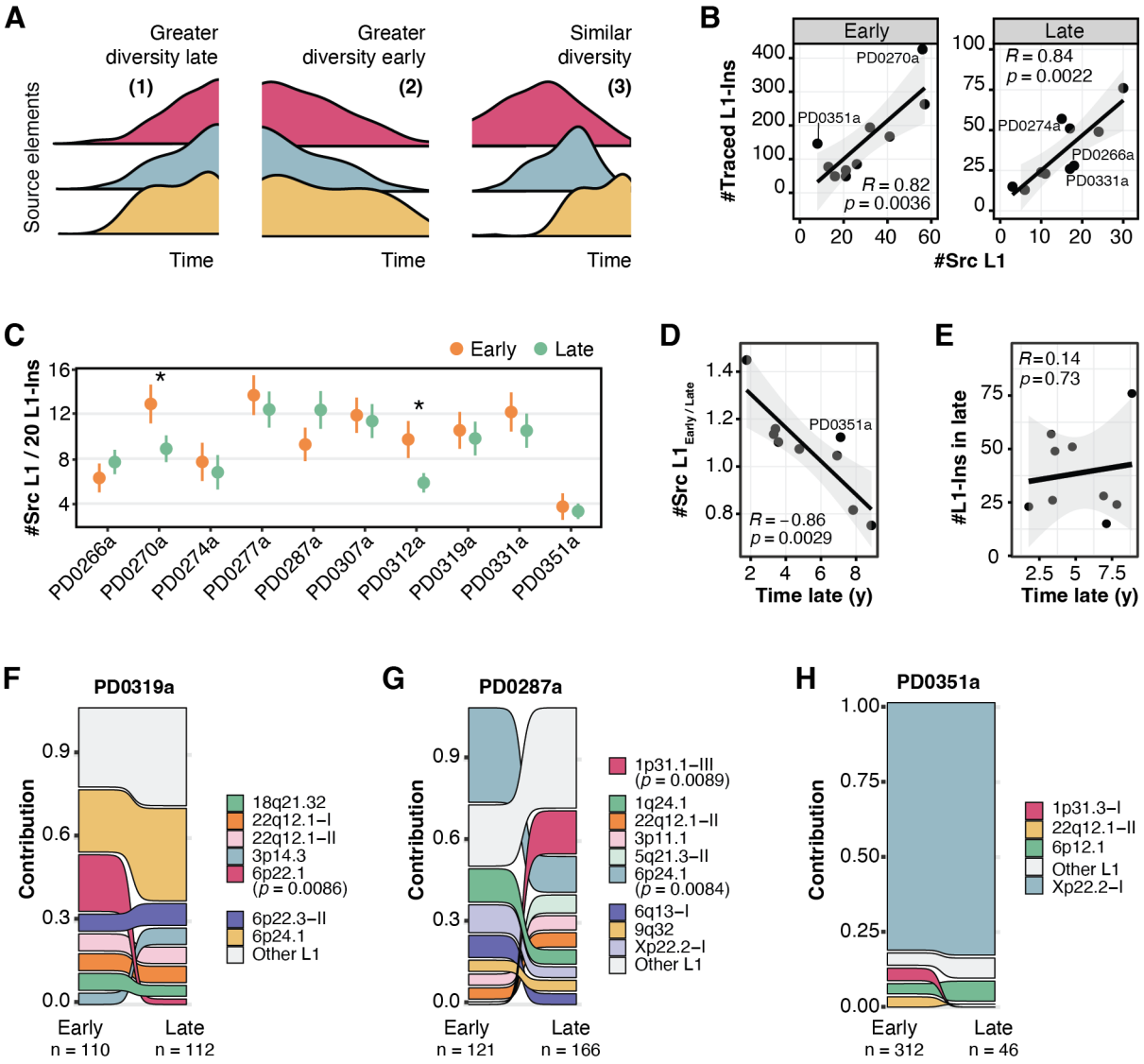


Figure 19. Dynamic evolution of source L1 elements activity throughout tumour development. (a) Models that may elucidate the activity dynamics of source L1 elements throughout tumour development. **(b)** Correlation between the number of active source elements and the total number of source-traced events for both early (left; Pearson correlation test, R=0.82, *p*=0.0036) and late stages (right; Pearson correlation test, R=0.84, *p*=0.0022). The number of active source elements was calculated using both transductions and solo-traced source information. **(c)** Comparative analysis of the number of distinct source elements active per tumour between early and late stages, using simple random sampling. This analysis revealed that tumours PD0270a and PD0312a harboured a higher diversity of active source elements at early stages (Probability of distinct source diversity, PD0270a: *p*=0.0161, PD0312a: *p*=0.0062). **(d)** Correlation between the years since the WGD events occurred and the ratio of distinct source elements active during early and late stages (Pearson correlation test, R=-0.86, p=0.0029). Our analysis unveiled a significant negative correlation, indicating that the differences observed in (c) are not a result of increased source diversity in either early or late stages. Instead, these differences stem from the duration of the time periods defined as early or late. **(e)** Correlation between the years since WGD events and the total number of source-traced retrotransposon insertions in late stages. No apparent correlation is observed (Pearson correlation test, R=0.14, *p*=0.73), which challenges the presumption of a consistent insertion rate over time. The relative contribution of each source element to the total

retrotransposition burden is illustrated for three notable cases: **(f)** PD0319a, **(g)** PD0287a and **(h)** PD0351a. Significant instances of both activation and deactivation of distinct source elements are observed. P-values were calculated using a two-proportions Z-test with continuity correction. Source elements with fewer than five assigned insertions are collectively grouped into 'Other L1'. Ins: Insertion; L1: Long Interspersed Nucleotide Element 1; Src L1: Source L1 element; WGD: Whole-Genome Doubling.

factors gradually suppress the activity of source elements as tumour development progresses. Lastly, it is possible that active source elements display a pattern reminiscent of the alternating activity observed in the germline for retrotransposon families. Hence, the activity rates of source elements would reflect their varying evolutionary success in invading and persisting within their host, leading to the prevalence of specific elements at particular times.

By utilizing both transductions and solo-traced source information, we thoroughly investigated the number of active source L1 elements per tumour at both early and late tumour stages. First, our analysis revealed a robust positive correlation between the number of active source elements and the total number of events for both early and late events (Pearson correlation test, R=0.82, $p<0.05$; **Fig. 19b**). However, there were notable exceptions. For example, in COAD sample PD0351a, the activity of just eight different source elements accounted for 146 somatic insertions during early stages. In contrast, as many as 56 different source elements were already active before the WGD event in the HNSC tumour PD0270a, contributing to a total of 426 somatic L1 insertions.

To ensure the comparability of source element diversity between samples and among early and late groups, given their differences in sample sizes, we conducted a resampling approach. This involved 10,000 iterations of simple random sampling with replacement, using a fixed sample size of 20 insertions per group. Subsequently, we compared the number of distinct source elements per tumour between early and late stages for each iteration. This analysis revealed that tumours PD0270a and PD0312a harboured a higher diversity of active source elements at early stages (PD0270a: $p=0.0161$, PD0312a: $p=0.0062$). In contrast, in sample PD0287a, we observed the opposite trend, although it did not reach statistical significance ($p=0.0523$). For the remaining samples, no significant differences were found, indicating similar number of active source elements between early and late stages (**Fig. 19c**).

Given the distinct trends observed in samples with varying time spans since the WGD events, we investigated the relationship between the years since WGDs and the ratio of active source L1s during early and late stages. Our analysis unveiled a significant negative correlation (Pearson correlation test, R=-0.86, $p=0.0029$; **Fig. 19d**), indicating that the previously observed differences are not due to increased source diversity in either early or late stages; instead, they stem from the duration of the time periods defined as early or late. Consistently, we found no correlation between the years since the WGD and the number of active source elements in late stages (Pearson correlation test, R=0.12, $p=0.77$), reinforcing the model of similar source diversity throughout tumour development. Additionally, when we studied the correlation between the years since WGDs and the total number of retrotransposon insertions in late stages, no apparent correlation was observed (Pearson correlation test, R=0.14, $p=0.73$; **Fig. 19e**), which challenges the presumption of a consistent insertion rate over time.

To further clarify this phenomenon, we investigated the dynamics of source L1s activity within specific tumours. Three notable instances are illustrated in **Figure 19f-h**. In the case of tumour PD0319a, we observed that some source elements, such as 6p22.1, became nearly inactive in the later stages of tumour development, despite their initial contribution of 20.41% (10 out of 49) to the retrotransposition burden with traced source. Conversely, in the context of tumour PD0287a, we identified the late activation of the source element 1p31.1-III, eventually becoming the most active source L1 element at that stage (15.8%, 12 out of 76). Lastly, within tumour PD0351a, we noted that source element Xp22.2-I accounted for a substantial 84.93% of the overall early activity with traced source (124 out of 146), and this high level of activity persisted into later stages (86.87%, 13 out of 15).

Overall, our findings suggest that the activity rates of source L1 elements fluctuate throughout tumour development, with instances of both activation and deactivation, while the overall number of active source elements remains limited. Notably, these results align with a model of consistent source diversity along tumorigenesis, resembling the dynamics observed in germline processes (**Fig. 19a, model 3**). Additionally, we observe that the relative contribution of source elements varies not only over time but also among different tumours, suggesting a dynamic scenario of regulation and deregulation of these potential mutators within the cancer genomes.

10.3 UNCOVERING THE DYNAMICS OF SOMATIC RETROTRANSPOSITION IN HEALTHY TISSUES

While somatic retrotransposition plays a significant role in tumorigenesis[68,79,83,92,125], there is still a lack of compelling evidence relative to its relevance and extent in normal tissues[101,106,108,111,113,116,124]. Unravelling somatic mutations in normal, non-clonally expanded tissues has posed a significant challenge; however, recent technological advancements have started to enable their detection[113,116,119,150,151]. Among these advancements is the laser capture microdissection (LCM) technique, which allows the isolation of small tissue sections, typically encompassing a few hundred cells of interest, that can then be subjected to comprehensive whole-genome sequencing. As a result, the mutational landscape of the isolated cells can be characterized.

To gain insights into the dynamics of somatic retrotransposition in normal tissues, we utilized Illumina whole-genome sequencing data obtained from LCM microbiopsies derived from healthy individuals. In total, our study involved the analysis of 504 microbiopsies obtained from 28 deceased healthy donors (**Supplementary Table 2**). These samples covered eight tissue types, including placenta (n=117), stomach (n=145), appendix (n=34), colon (n=59), pancreas (n=19), skin (n=12), small bowel (n=59) and testis (n=59) (**Fig. 20a**). For this analysis, we utilized our method, MEIGA-SR (v1.1.0), specifically designed for analysing retrotransposition in short-reads datasets. We conducted paired analyses using matched control samples typically obtained from muscle or blood.

Our approach identified a total of 220 somatically acquired retrotransposon insertions, all confirmed by IGV inspection. L1-solo elements represented the most common type of insertion, accounting for 62.27% (n=137) of the total retrotransposition events, followed by L1 transductions (20%, n=44) and poly(A/T) tracts (6.82%; n=15) (**Fig. 20b**). Additionally, we

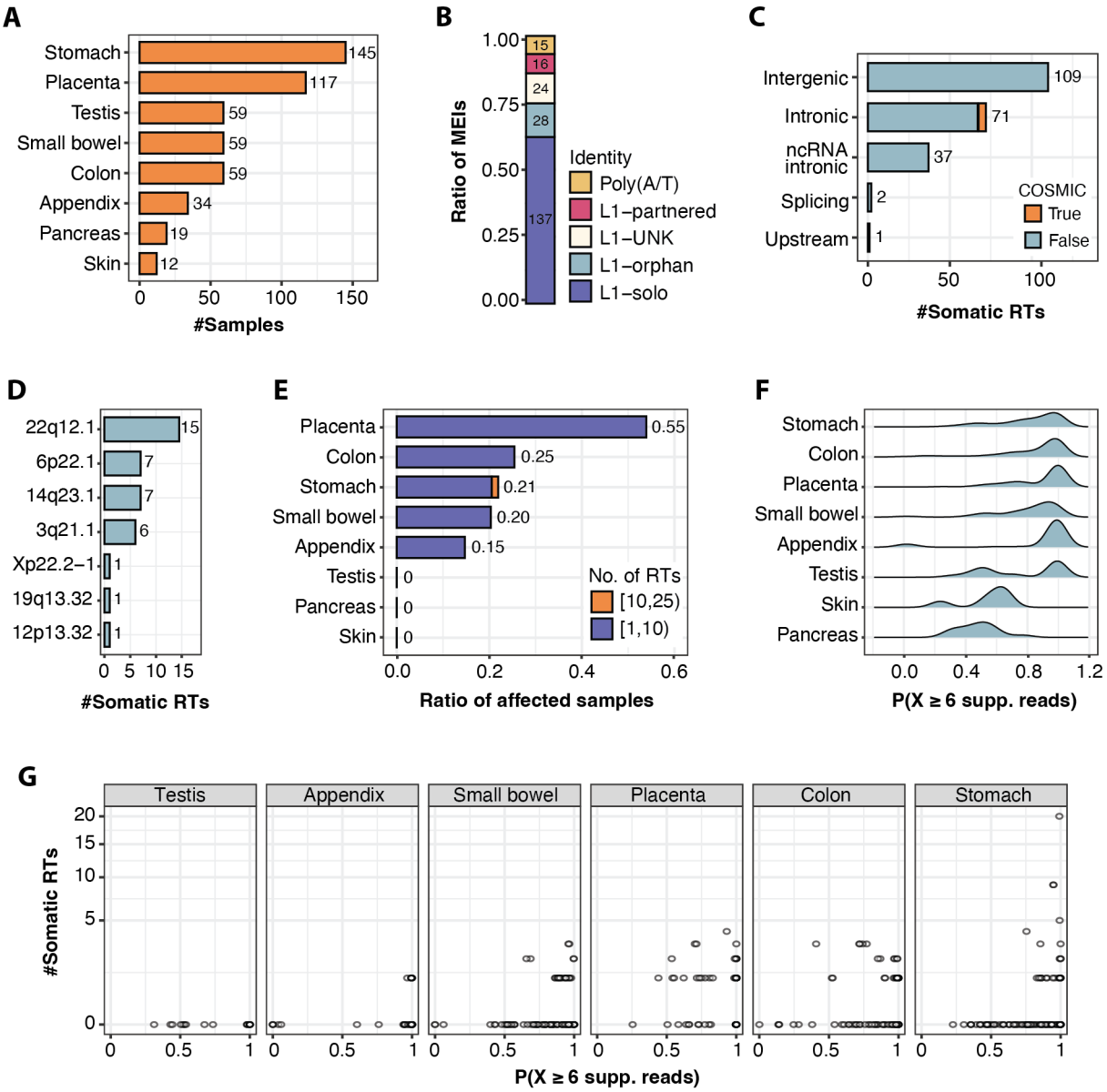


Figure 20. Somatic retrotransposition in active in multiple healthy tissues. (a) Number of LCM microbiopsies studied by tissue type. **(b)** Overall percentage of somatic retrotransposition events (n=220) attributed to each retrotransposon identity: L1-solo insertions (62.27%), L1-orphan transductions (12.7%), L1-partenered transductions (8%), poly(A/T) tracts (6.82%) and L1-UNK insertions (10.91%). L1-UNK signifies cases where resolution was insufficient to determine whether the inserted element was a solo-L1 or an L1-partnered transduction. **(c)** Overall number of events classified per insertion site, with the COSMIC database employed to determine whether the insertion occurred within a cancer-related gene. **(d)** Overall number of events attributed to each source element using a transduction-based strategy. **(e)** Proportion of microbiopsies displaying one or more somatic insertions across tissue types, with a purple shade representing a retrotransposition rate between one to nine, while values exceeding this range are depicted in orange. **(f)** Comparative analysis of the detection power across microbiopsies and tissues. A binomial cumulative distribution function was used to estimate the probability of obtaining six or more supporting reads (MEIGA-SR requirement to make a call) based on the sequencing depth and clonality values of each microbiopsy. This probability, taken as a proxy for the detection power, is consistent across all

tissues except for skin and pancreas, where it was notably lower. Median probability (X≥6 supporting reads) for each tissue type: pancreas=0.495, skin=0.618, testis=0.857, stomach=0.862, small bowel=0.878, colon=0.954, appendix=0.996, placenta=0.998. **(g)** Correlation between the total number of somatic retrotransposon insertions per microbiopsy and the proxy for the detection power. Somatic retrotransposition events are identified in the small bowel, stomach, placenta and colon, while no apparent activity is detected in the testis. ncRNA: non-coding RNA; L1: Long Interspersed Nucleotide Element 1; RT: Retrotransposition; Supp: Supporting; UNK: Unknown.

encountered 10.91% (n=24) of L1-UNK events, signifying cases where resolution was insufficient to determine whether the inserted element was a solo-L1 or a L1-transduction. Notably, we did not find evidence for *Alu* or SVA activity in the samples analysed.

Half of the somatic retrotransposon insertions were identified within genes (50.45%; 111 out of 220), preferentially at intronic regions. Of these insertions, 74 had the potential to impact mRNA processing—71 were located intronically, two affected splicing and one was positioned upstream within the promoter region (**Fig. 20c**). Notably, our analysis revealed three intronic insertions within cancer-associated genes, specifically *ALK, CBLB* and *PTPRD*, as per the COSMIC database[129].

We examined the contribution of specific source L1 elements to the overall retrotransposition burden in healthy tissues (**Fig. 20d**). Notably, we observed that source elements associated with multiple transductions in our study, including 22q12.1, 6p22.1, 14q23.1 and 3q21.1, were previously recognized as among the top ten most active source elements in cancer[68,92]. This observation indicates a consistent pattern with tumours, suggesting that highly active source elements in cancer are already unrestrained within healthy tissues years before the onset of the disease.

The comparison of retrotransposition rates across different tissue types revealed the placenta as the tissue with the highest proportion of biopsies affected by one or more somatic retrotransposition events (54.70%, 64 out of 117), underscoring that retrotransposition was active in over half of the biopsies at some point during placental development (**Fig. 20e**). Notably, the colon followed with a rate of 25.42% (15 out of 59), along with the stomach at 21.38% (31 out of 145), small bowel at 20.34% (12 out of 59) and appendix at 14.71% (5 out of 34). Remarkably, the stomach displayed the highest number of somatic retrotransposon insertions per biopsy, revealing up to 20 distinct somatic insertions within a single microbiopsy. Overall, these findings indicate that retrotransposition is a mutational mechanism actively occurring in various healthy tissues.

However, the ability to detect somatic mutations is directly linked to both the sequencing depth and the clonality of the tissue under investigation. To assess whether the detection power was consistent across biopsies and tissues, we employed a binomial cumulative distribution function (**Fig. 20f,g**). We used the following formula:

$$P(X > x) = 1 - \sum_{i=0}^{x-1} \binom{n}{i} p^i (1-p)^{(n-i)}$$

where n corresponds to the sequencing depth, p represents the probability of success, which corresponds to the VAF of the main cluster, and x was set to the minimum number of supporting

reads required by MEIGA-SR to call an insertion, specifically six. This binomial cumulative distribution enabled us to estimate the probability of obtaining six or more supporting reads given the sequencing depth and clonality values of each microbiopsy. We then used the estimated probability as a proxy for the likelihood of MEIGA-SR to detect an insertion.

Our findings showed that the detection power is consistent across all tissues except for skin and pancreas, where it was notably lower (Median probability ($X \geq 6$ reads): pancreas=0.495, skin=0.618, testis=0.857, stomach=0.862, small bowel=0.878, colon=0.954, appendix=0.996, placenta=0.998). Consequently, our results suggest that the observed differences in activity rates for placenta, appendix, colon, stomach and testis are genuine, but the same is not true for skin and pancreas biopsies (**Fig. 20f**).

In summary, in line with previous studies demonstrating extensive mutagenesis in placental tissues[117], our observations indicated that over half of the placental microbiopsies were impacted by somatic retrotransposition, making it the most affected tissue type. This suggests a scenario where there might not be significant selective pressure to preserve the genetic material of this fast-growing, short-lived, disposable organ. In contrast, there was no apparent retrotransposon activity in the testis, supporting the low mutational burden already reported in this tissue[118]. Notably, the number of somatic retrotranspositions observed in certain healthy stomach microbiopsies was higher than those reported for most stomach adenocarcinomas[68]. Overall, our results suggest that somatic retrotransposition is active years before the onset of diseases like cancer, presumably contributing to the generation of genomic variability that can ultimately contribute to tumour progression.

CONCLUSIONS

11 CONCLUSIONS

i. **Our bioinformatic pipeline, MEIGA, stands out as a robust and reliable tool for detecting somatic retrotransposition using long-read sequencing.** MEIGA significantly outperforms established methods like xTea, PALMER and rMETL in detecting retrotransposition events. It particularly excels in the accurate detection of transductions and is the only method suitable for identifying retrotransposition-mediated rearrangements. Moreover, MEIGA is the only tool that enables paired tumour-normal analysis to identify tumour-specific events. Overall, MEIGA is a valuable asset in the field of retrotransposition, offering comprehensive characterization of a wide range of mutational patterns driven by retrotransposons.

ii. **Long-read sequencing unveiled a rich landscape of cryptic somatic retrotransposition events in primary tumours.** MEIGA analysis of five HNSC, four LUSC and a COAD identified as many as 6,266 somatic insertions and 152 genomic rearrangements, all mediated by retrotransposition. Notably, in one exceptional HNSC tumour, we identified up to 1,311 somatic retrotranspositions. In addition to previously undisclosed levels of somatic retrotransposition, this analysis revealed the hallmark features of somatic retrotransposition with an unprecedented level of resolution.

iii. **Somatic retrotransposition plays a significant role in promoting large-scale genomic instability in certain tumours.** Our long-read sequencing analysis of a cancer cohort unveiled a hidden landscape of balanced chromosomal rearrangements mediated by retrotransposition, including reciprocal translocations and inversions. We consistently observed this pattern in nine out of the ten analysed tumours, highlighting the substantial role that retrotransposons play in reshaping the cancer genome. Notably, our findings shed light on the mechanism behind these events: they can arise from either one or two independent somatic retrotransposition events, though the latter appears to be the more common mechanism.

iv. **Somatic retrotransposition is an early mutational process active throughout tumorigenesis.** Our timing analysis facilitated the identification of retrotransposition events along different stages of cancer progression, offering valuable insights into the temporal dynamics of this mutational process. This unveiled that insertions, but also large-scale chromosomal rearrangements mediated by retrotransposition, frequently occurred early in tumorigenesis. Overall, our findings indicated that somatic retrotransposition can be active years before tumour diagnosis and persist throughout tumour development.

v. **Somatic retrotransposition events are relatively common in multiple healthy tissues, but retrotransposition rates vary across tissue types.** We identified somatic retrotransposition events in the placenta, stomach, colon, small bowel and appendix, while the testis displayed no apparent retrotransposition activity. Overall, our results suggested that somatic retrotransposition is active years before the onset of diseases like cancer, presumably contributing to the generation of genomic variability that can ultimately play a role in tumour progression.

BIBLIOGRAPHY

12 BIBLIOGRAPHY

1. Chandler, M., Gellert, M., Lambowitz, A. M., Rice, P. A. & Sandmeyer, S. B. Mobile DNA III. *Mobile DNA III* **1**, 1–1321 (2015).
2. Hartl, D. L., Lozovskaya, E. R. & Lawrence, J. G. Nonautonomous transposable elements in prokaryotes and eukaryotes. *Genetica* **86**, 47–53 (1992).
3. Biémont, C. & Vieira, C. Genetics: Junk DNA as an evolutionary force. *Nature* **443**, 521–524 (2006).
4. McClintock, B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36**, 344–355 (1950).
5. Tenaillon, M. I., Hufford, M. B., Gaut, B. S. & Ross-Ibarra, J. Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. *Genome Biol Evol* **3**, 219–229 (2011).
6. Buisine, N., Quesneville, H. & Colot, V. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics* **91**, 467–475 (2008).
7. Kaul, S. *et al.* Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).
8. Filée, J., Siguier, P. & Chandler, M. Insertion Sequence Diversity in Archaea. *Microbiology and Molecular Biology Reviews* **71**, 121–157 (2007).
9. Siguier, P., Filée, J. & Chandler, M. Insertion sequences in prokaryotic genomes. *Curr Opin Microbiol* **9**, 526–531 (2006).
10. Touchon, M. & Rocha, E. P. C. Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* **24**, 969–981 (2007).
11. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
12. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
13. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691–703 (2009).
14. Feschotte, C. & Pritham, E. J. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41**, 331–368 (2007).
15. Mills, R. E., Bennett, E. A., Iskow, R. C. & Devine, S. E. Which transposable elements are active in the human genome? *Trends in Genetics* **23**, 183–191 (2007).

16. Pace, J. K. & Feschotte, C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res* **17**, 422–432 (2007).
17. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**, 657–663 (1999).
18. Sinzelle, L., Izsvák, Z. & Ivics, Z. Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cellular and Molecular Life Sciences* **66**, 1073–1093 (2009).
19. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* **12**, 615–627 (2011).
20. Thompson, C. B. New insights into V(D)J recombination and its role in the evolution of the immune system. *Immunity* **3**, 531–539 (1995).
21. Zhang, Y. *et al.* Transposon molecular domestication and the evolution of the RAG recombinase. *Nature* **569**, 79–84 (2019).
22. Boeke, J. D. The unusual phylogenetic distribution of retrotransposons: A hypothesis. *Genome Res* **13**, 1975–1983 (2003).
23. Galindo-González, L., Mhiri, C., Deyholos, M. K. & Grandbastien, M. A. LTR-retrotransposons in plants: Engines of evolution. *Gene* **626**, 14–25 (2017).
24. Wessler, S. R., Bureau, T. E. & White, S. E. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr Opin Genet Dev* **5**, 814–821 (1995).
25. Zhang, Y., Romanish, M. T. & Mager, D. L. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput Biol* **7**, e1002046 (2011).
26. Wilhelm, M. & Wilhelm, F. X. Reverse transcription of retroviruses and LTR retrotransposons. *Cellular and Molecular Life Sciences* **58**, 1246–1262 (2001).
27. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell* **72**, 595–605 (1993).
28. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**, 41–48 (2003).
29. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H. SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. *Am J Hum Genet* **73**, 1444–1451 (2003).
30. Wang, H. *et al.* SVA elements: A hominid-specific retroposon family. *J Mol Biol* **354**, 994–1007 (2005).
31. Cordaux, R., Hedges, D. J., Herke, S. W. & Batzer, M. A. Estimating the retrotransposition rate of human Alu elements. *Gene* **373**, 134–137 (2006).
32. Xing, J. *et al.* Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res* **19**, 1516–1526 (2009).
33. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**, e1002236 (2011).
34. Huang, C. R. L. *et al.* Mobile interspersed repeats are major structural variants in the human genome. *Cell* **141**, 1171–1182 (2010).
35. Ewing, A. D. & Kazazian, H. H. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* **20**, 1262–1270 (2010).
36. Babushok, D. V. & Kazazian, H. H. Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat* **28**, 527–539 (2007).

37. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**, 370–379 (2002).
38. Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J. & Schmitz, J. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends in Genetics* **23**, 158–161 (2007).
39. Swergold, G. D. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol* **10**, 6718–6729 (1990).
40. Goodier, J. L. & Kazazian, H. H. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. *Cell* **135**, 23–35 (2008).
41. Ostertag, E. M. & Kazazian, J. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* **11**, 2059–2065 (2001).
42. Morrish, T. A. *et al.* DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31**, 159–165 (2002).
43. Callinan, P. A. & Batzer, M. A. Retrotransposable elements and human disease. *Genome Dyn* **1**, 104–115 (2006).
44. Deininger, P. L. & Batzer, M. A. Alu repeats and human disease. *Mol Genet Metab* **67**, 183–193 (1999).
45. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet* **12**, 187–215 (2011).
46. Kazazian, H. H. *et al.* Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).
47. Chen, J. M., Stenson, P. D., Cooper, D. N. & Férec, C. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet* **117**, 411–427 (2005).
48. Belancio, V. P., Deininger, P. L. & Roy-Engel, A. M. LINE dancing in the human genome: Transposable elements and disease. *Genome Med* **1**, (2009).
49. Kazazian, H. H. & Moran, J. V. Mobile DNA in Health and Disease. *New England Journal of Medicine* **377**, 361–370 (2017).
50. Belancio, V. P., Hedges, D. J. & Deininger, P. Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Res* **18**, 343–358 (2008).
51. Belancio, V. P., Roy-Engel, A. M. & Deininger, P. The impact of multiple splice sites in human L1 elements. *Gene* **411**, 38–45 (2008).
52. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science* **351**, (2016).
53. Han, J. S., Szak, S. T. & Boeke, J. D. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**, 268–274 (2004).
54. Chen, J., Rattner, A. & Nathans, J. Effects of L1 retrotransposon insertion on transcript processing, localization and accumulation: Lessons from the retinal degeneration 7 mouse and implications for the genomic ecology of L1 elements. *Hum Mol Genet* **15**, 2146–2156 (2006).
55. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**, 563–571 (2009).
56. Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).
57. Garcia-Perez, J. L. *et al.* Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. *Nature* **466**, 769–773 (2010).
58. Estécio, M. R. H. *et al.* SINE retrotransposons cause epigenetic reprogramming of adjacent gene promoters. *Molecular Cancer Research* **10**, 1332–1342 (2012).

59. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).
60. Pickeral, O. K., Makałowski, W., Boguski, M. S. & Boeke, J. D. Frequent human genomic DNA transduction driven by line-1 retrotransposition. *Genome Res* **10**, 411–415 (2000).
61. Goodier, J. L., Ostertag, E. M. & Kazazian, H. H. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet* **9**, 653–657 (2000).
62. Xing, J. *et al.* Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc Natl Acad Sci U S A* **103**, 17608–17613 (2006).
63. Gilbert, N., Lutz, S., Morrish, T. A. & Moran, J. V. Multiple Fates of L1 Retrotransposition Intermediates in Cultured Human Cells. *Mol Cell Biol* **25**, 7780–7795 (2005).
64. Gilbert, N., Lutz-Prigge, S. & Moran, J. V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**, 315–325 (2002).
65. Symer, D. E. *et al.* Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* **110**, 327–338 (2002).
66. Callinan, P. A. *et al.* Alu retrotransposition-mediated deletion. *J Mol Biol* **348**, 791–800 (2005).
67. Miné, M. *et al.* A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. *Hum Mutat* **28**, 137–142 (2007).
68. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet* **52**, 306–319 (2020).
69. Burwinkel, B. & Kilimann, M. W. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol* **277**, 513–517 (1998).
70. Lehrman, M. A. *et al.* Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science* **227**, 140–146 (1985).
71. Lee, J., Han, K., Meyer, T. J., Kim, H. S. & Batzer, M. A. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS One* **3**, e4047 (2008).
72. Bailey, J. A., Liu, G. & Eichler, E. E. An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. *Am J Hum Genet* **73**, 823–834 (2003).
73. Song, M. & Boissinot, S. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. *Gene* **390**, 206–213 (2007).
74. Temtamy, S. A. *et al.* Long interspersed nuclear element-1 (LINE1)-mediated deletion of EVC, EVC2, C4orf6, and STK32B in ellis-van Creveld syndrome with borderline intelligence. *Hum Mutat* **29**, 931–938 (2008).
75. Segal, Y. *et al.* LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. *Am J Hum Genet* **64**, 62–69 (1999).
76. Miki, Y. *et al.* Disruption of the APC Gene by a Retrotransposal Insertion of LI Sequence in a Colon Cancer. *Cancer Res* **52**, 643–645 (1992).
77. Morse, B., Rotherg, P. G., South, V. J., Spandorfer, J. M. & Astrin, S. M. Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. *Nature* **333**, 87–90 (1988).
78. Iskow, R. C. *et al.* Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**, 1253–1261 (2010).

79. Solyom, S. *et al.* Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res* **22**, 2328–2338 (2012).
80. Shukla, R. *et al.* Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111 (2013).
81. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**, 685–696 (2010).
82. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
83. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).
84. Helman, E. *et al.* Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res* **24**, 1053–1063 (2014).
85. Erwin, J. A. *et al.* L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci* **19**, 1583–1591 (2016).
86. Maciejowski, J. & De Lange, T. Telomeres in cancer: Tumour suppression and genome instability. *Nat Rev Mol Cell Biol* **18**, 175–186 (2017).
87. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* **12**, 780–786 (2015).
88. Witherspoon, D. J. *et al.* Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics* **11**, 1–15 (2010).
89. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727–732 (2005).
90. Xing, J., Witherspoon, D. J. & Jorde, L. B. Mobile element biology: New possibilities with high-throughput sequencing. *Trends in Genetics* **29**, 280–289 (2013).
91. Thung, D. T. *et al.* Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol* **15**, 488 (2014).
92. Tubio, J. M. C. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, (2014).
93. Gardner, E. J. *et al.* The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res* **27**, 1916–1929 (2017).
94. Ewing, A. D. Transposable element detection from whole genome sequence data. *Mob DNA* **6**, 1–9 (2015).
95. Jiang, T., Liu, B., Li, J. & Wang, Y. RMETL: Sensitive mobile element insertion detection with long read realignment. *Bioinformatics* **35**, 3484–3486 (2019).
96. Zhou, W. *et al.* Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res* **48**, 1146–1163 (2020).
97. Ewing, A. D. *et al.* Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol Cell* **80**, 915-928.e5 (2020).
98. Chu, C. *et al.* Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun* **12**, 1–12 (2021).
99. van den Hurk, J. A. J. M. *et al.* L1 retrotransposition can occur early in human embryonic development. *Hum Mol Genet* **16**, 1587–1592 (2007).
100. Garcia-Perez, J. L. *et al.* LINE-1 retrotransposition in human embryonic stem cells. *Hum Mol Genet* **16**, 1569–1577 (2007).
101. Kano, H. *et al.* L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev* **23**, 1303–1312 (2009).

102. Burton, A. & Torres-Padilla, M. E. Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. *Nat Rev Mol Cell Biol* **15**, 722–734 (2014).
103. Fadloun, A. *et al.* Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nat Struct Mol Biol* **20**, 332–338 (2013).
104. Jachowicz, J. W. *et al.* LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet* **49**, 1502–1510 (2017).
105. Chow, J. C. *et al.* LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**, 956–969 (2010).
106. Evrony, G. D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
107. Upton, K. R. *et al.* Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228–239 (2015).
108. Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534–537 (2011).
109. Muotri, A. R. *et al.* L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**, 443–446 (2010).
110. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–910 (2005).
111. Yamaguchi, K. *et al.* Striking heterogeneity of somatic L1 retrotransposition in single normal and cancerous gastrointestinal cells. *Proc Natl Acad Sci U S A* **117**, 32215–32222 (2020).
112. Doucet-O'Hare, T. T. *et al.* Somatically Acquired LINE-1 Insertions in Normal Esophagus Undergo Clonal Expansion in Esophageal Squamous Cell Carcinoma. *Hum Mutat* **37**, 942–954 (2016).
113. Nam, C. H. *et al.* Widespread somatic L1 retrotransposition in normal colorectal epithelium. *Nature 2023 617:7961* **617**, 540–547 (2023).
114. Nawy, T. Single-cell sequencing. *Nat Methods* **11**, 18 (2014).
115. Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp Mol Med* **52**, 1419–1427 (2020).
116. Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, (2020).
117. Coorens, T. H. H. *et al.* Inherent mosaicism and extensive mutation of human placentas. *Nature* **592**, 80–85 (2021).
118. Coorens, T. H. H. *et al.* Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387–392 (2021).
119. Moore, L. *et al.* The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
120. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**, 155–158 (2020).
121. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737–746 (2017).
122. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
123. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
124. Ewing, A. D. *et al.* Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome Res* **25**, 1536–1545 (2015).

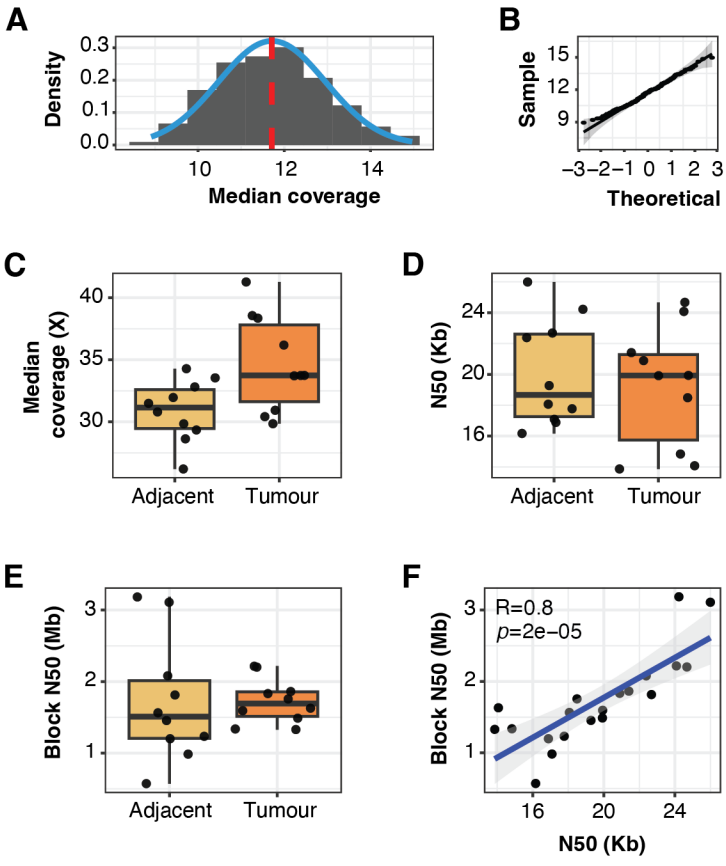
125. Scott, E. C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res* **26**, 745–755 (2016).
126. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, (2021).
127. Beck, C. R. *et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010).
128. Shiraishi, Y. *et al.* Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. *Nucleic Acids Res* **51**, E74 (2023).
129. Tate, J. G. *et al.* COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941–D947 (2019).
130. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**, 539–543 (2008).
131. Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534–538 (2008).
132. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* **178**, 91-106.e23 (2019).
133. Bairoch, A. The cellosaurus, a cell-line knowledge resource. *Journal of Biomolecular Techniques* **29**, 25–38 (2018).
134. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461–468 (2018).
135. Pellestor, F. & Gatinois, V. Chromoanagenesis: A piece of the macroevolution scenario. *Mol Cytogenet* **13**, (2020).
136. Venkatesan, S., Natarajan, A. T. & Hande, M. P. Chromosomal instability--mechanisms and consequences. *Mutat Res Genet Toxicol Environ Mutagen* **793**, 176–184 (2015).
137. Lee, J. A., Carvalho, C. M. B. & Lupski, J. R. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell* **131**, 1235–1247 (2007).
138. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet* **5**, e1000327 (2009).
139. Koumbaris, G. *et al.* FoSTeS, MMBIR and NAHR at the human proximal Xp region and the mechanisms of human Xq isochromosome formation. *Hum Mol Genet* **20**, 1925–1936 (2011).
140. Smit, AFA, Hubley, R & Green, P. RepeatMasker Home Page. *RepeatMasker Open-4.0.5*. <http://www.repeatmasker.org>
141. Cmero, M. *et al.* Inferring structural variant cancer cell fraction. *Nat Commun* **11**, 1–15 (2020).
142. Bielski, C. M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet* **50**, 1189–1195 (2018).
143. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature 2020 578:7793* **578**, 122–128 (2020).
144. Dewhurst, S. M. *et al.* Tolerance of whole- genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov* **4**, 175–185 (2014).
145. Kuznetsova, A. Y. *et al.* Chromosomal instability, tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells. *Cell Cycle* **14**, 2810–2820 (2015).

146. Gemble, S. *et al.* Genetic instability from a single S phase after whole-genome duplication. *Nature* **604**, 146–151 (2022).
147. Fujiwara, T. *et al.* Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. *Nature* **437**, 1043–1047 (2005).
148. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**, (2016).
149. Sultana, T. *et al.* The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Mol Cell* **74**, 555-570.e7 (2019).
150. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
151. Luquette, L. J. & Park, P. J. Somatic mutation accumulation seen through a single-molecule lens. *Cell Res* **31**, 949–950 (2021).
152. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
153. Tischler, G. & Leonard, S. Biobambam: Tools for read pair collation based algorithms on BAM files. *Source Code Biol Med* **9**, 13 (2014).
154. Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. doi:10.1101/085050.
155. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).
156. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213–219 (2013).
157. Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and indels by local assembly. *Genome Res* **28**, 581–591 (2018).
158. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
159. Jolly, C. & Van Loo, P. Timing somatic events in the evolution of cancer. *Genome Biol* **19**, 1–9 (2018).
160. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: A next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
161. Ono, Y., Asai, K. & Hamada, M. PBSIM: PacBio reads simulator - Toward accurate genome assembly. *Bioinformatics* **29**, 119–121 (2013).
162. Kent, W. J. BLAT —The BLAST-Like Alignment Tool. *Genome Res* **12**, 656–664 (2002).

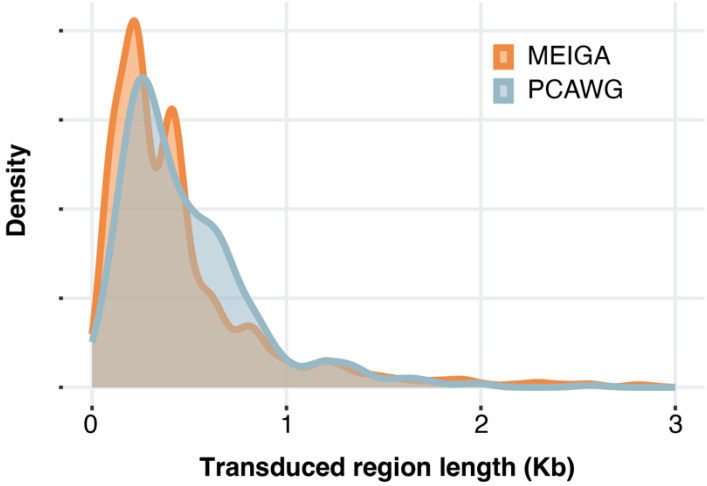
APPENDICES

13 APPENDICES

13.1 SUPPLEMENTARY FIGURES

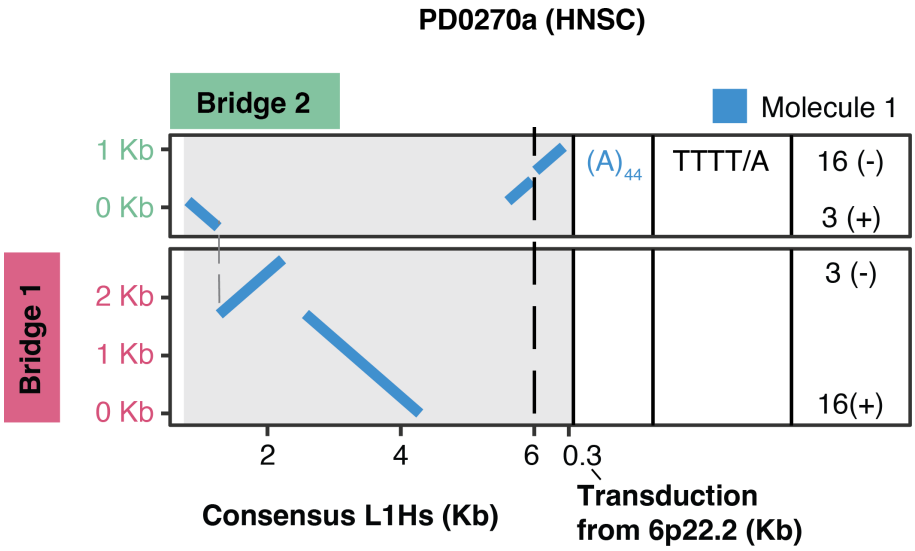


Supplementary figure 1. Sequencing statistics. (a) Distribution of the median coverage in a cohort of 150 tumours sequenced with Illumina shallow whole-genome sequencing, as part of a prospective screening. The resulting median value was 11.7x. **(b)** Quantile-quantile plot displaying the quantiles of the residual distribution from (a) on the y-axis, and the theoretical quantiles from a normal distribution (x-axis). This analysis indicates that the distribution of sequencing coverage conforms to a normal distribution, suggesting an unbiased detection power across the screening cohort. **(c)** Boxplots depicting the median coverage of 10 tumours with high retrotransposition rates and their respective adjacent tissues, sequenced using Oxford Nanopore Technologies (ONT). This sequencing yielded a median coverage of 33.7x for tumours and 31.1x for the adjacent tissues. **(d)** For the ONT dataset introduced in (c), boxplots displaying the N50 read length metric. The observed N50 values were 19.9 kb for the tumours, and 18.7 kb for the adjacent tissues. **(e)** For the ONT dataset introduced in (c), boxplots show the N50 values of haplotype blocks derived from Whatshap phase sets. **(f)** A significant correlation was observed between the N50 read length and the N50 phasing-blocks (Pearson correlation test, R=0.8, $p=2{\cdot}10^{-5}$).

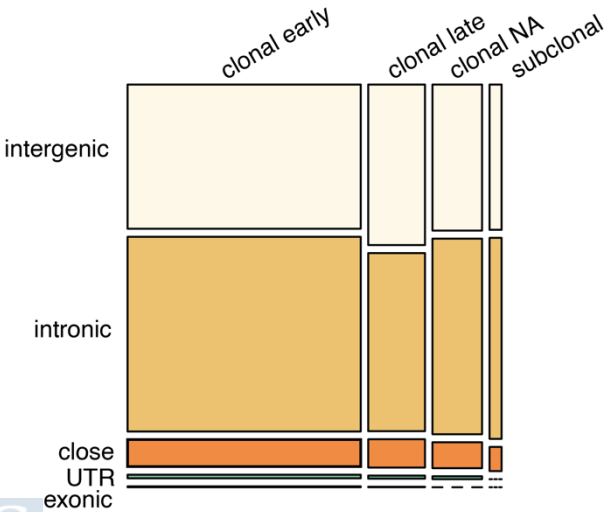


Supplementary figure 2. MEIGA identifies L1-partnered transductions with exceptionally short transduced regions. Comparative analysis of the length distribution of L1-partnered transductions detected by MEIGA in 10 cancer whole genomes sequenced with ONT, and TraFiC in the PCAWG dataset sequenced with Illumina. L1: Long Interspersed Nucleotide Element 1; ONT: Oxford Nanopore Technologies.

Supplementary figure 3. IGV visualization of a reciprocal translocation mediated by L1 retrotransposition identified by MEIGA. The pattern formed at each breakpoint junction is indistinguishable from a canonical L1 insertion when analysed with Illumina short-read sequencing. L1: Long Interspersed Nucleotide Element 1.



Supplementary figure 4. Both bridges of a L1-mediated reciprocal translocation can stem from a single retrotransposition event. The sequence dot plots demonstrate alignments to the L1Hs consensus sequence and the downstream sequence of the source L1 element at 6p22.2 for both bridges. Features such as poly(A/T) tail lengths, endonuclease motifs and the orientations of the breakpoint junctions are detailed. Alignment colours indicate the molecule of origin. Overall, our data suggests that this reciprocal translocation arises from a single retrotransposition event. L1: Long Interspersed Nucleotide Element 1; L1Hs: Long Interspersed Nucleotide Element 1 Human-Specific.



Supplementary figure 5. There are no significant differences in the distribution of somatic retrotransposon insertions within genic regions along tumour evolution. Comparative analysis of the retrotransposition frequency along different relative timing categories, including clonal early, clonal late, clonal NA and subclonal. NA: Not assigned; UTR: Untranslated region.

13.2 SUPPLEMENTARY TABLES

Supplementary Table 1. Description of the tumour samples included in the primary tumours cohort. For each tumour, the following metadata is reported: tumour_id=unique tumour identifier; tumour_type=histological tumour type; source=source it provides from. BB: Basque Biobank; GFGM: Galician Foundation of Genomic Medicine; selected_ONT=(1) if selected for ONT sequencing; blood=(1) if blood available; adj_tissue=(1) if adjacent tissue available; sex=patient sex; stage=tumour stage; relapse; metastasis; death; tobacco; alcohol. COAD: Colorectal Adenocarcinoma; HNSC: Head and Neck Squamous Cell Carcinoma; LUSC: Lung Squamous Cell Carcinoma.

tumour_id	tumour_type	source	selected_ONT	blood	adj_tissue	sex	stage	relapse	metastasis	death	tobacco	alcohol
PD0266a	HNSC	BB	1	1	1	male	3	no	no	no	yes	yes
PD0270a	HNSC	BB	1	1	1	male	3	no	no	no	yes	no
PD0274a	HNSC	BB	1	1	1	female	4a	no	no	no	yes	yes
PD0277a	HNSC	BB	1	1	1	male	3	no	no	no	yes	yes
PD0287a	HNSC	BB	1	1	1	male	4	no	no	no	yes	no
PD0299a	LUSC	BB	1	1	1	male	2a	no	no	no	yes	no
PD0307a	LUSC	BB	1	1	1	male	3	no	no	no	no	no
PD0319a	LUSC	BB	1	1	1	male	2a	no	no	no	no	yes
PD0331a	LUSC	BB	1	1	1	male	2a	no	no	no	no	no
PD0351a	COAD	BB	1	0	1	male	NA	no	yes	yes	no	no
PD0002a	HNSC	GFGM	0	0	1	male	4	no	no	no	yes	yes
PD0003a	HNSC	GFGM	0	0	1	male	2	no	no	no	yes	yes
PD0004a	HNSC	GFGM	0	0	1	male	4	NA	no	NA	no	yes
PD0005a	HNSC	GFGM	0	0	1	male	3	NA	no	NA	yes	yes
PD0006a	HNSC	GFGM	0	0	1	male	3	NA	no	NA	no	yes
PD0008a	HNSC	GFGM	0	0	1	male	4	no	yes	yes	no	no
PD0011a	HNSC	GFGM	0	0	1	male	4	yes	no	no	yes	yes
PD0012a	HNSC	GFGM	0	1	1	male	3	no	no	no	no	yes
PD0015a	HNSC	GFGM	0	0	1	male	4	no	no	no	NA	NA
PD0030a	HNSC	GFGM	0	0	1	female	4	NA	yes	NA	no	yes
PD0077a	HNSC	GFGM	0	0	1	male	NA	no	no	no	NA	NA
PD0094a	HNSC	GFGM	0	0	1	male	4	no	yes	yes	yes	yes
PD0106a	HNSC	GFGM	0	1	1	male	2	no	no	no	yes	yes
PD0251a	HNSC	BB	0	1	1	male	3	no	yes	yes	yes	yes
PD0252a	HNSC	BB	0	1	1	male	4	no	yes	yes	yes	yes
PD0253a	HNSC	BB	0	1	1	male	2	no	NA	yes	yes	yes
PD0254a	HNSC	BB	0	1	1	male	3	yes	no	yes	yes	yes
PD0255a	HNSC	BB	0	1	1	male	2	yes	no	yes	yes	yes
PD0256a	HNSC	BB	0	1	1	male	4	no	no	yes	yes	no
PD0257a	HNSC	BB	0	1	1	male	2	yes	no	yes	yes	yes
PD0258a	HNSC	BB	0	1	1	male	3	no	no	yes	yes	yes
PD0259a	HNSC	BB	0	1	1	male	NA	no	no	no	yes	yes
PD0260a	HNSC	BB	0	1	1	male	NA	no	no	no	yes	yes
PD0261a	HNSC	BB	0	1	1	male	4	no	yes	yes	yes	yes
PD0262a	HNSC	BB	0	1	1	male	4	no	no	no	yes	yes
PD0263a	HNSC	BB	0	1	1	male	4	no	yes	yes	NA	no
PD0264a	HNSC	BB	0	1	1	male	2	yes	no	yes	yes	yes
PD0265a	HNSC	BB	0	1	1	male	4a	no	no	no	yes	yes
PD0267a	HNSC	BB	0	1	1	male	4	no	no	no	yes	yes
PD0268a	HNSC	BB	0	1	1	male	4	yes	no	yes	yes	yes
PD0269a	HNSC	BB	0	1	1	male	4	no	no	yes	yes	NA
PD0271a	HNSC	BB	0	1	1	male	3	yes	no	yes	NA	yes
PD0273a	HNSC	BB	0	1	1	male	2	no	no	no	yes	yes

PD0275a	HNSC	BB	0	1	1	male	4	NA	no	no	yes	yes
PD0276a	HNSC	BB	0	1	1	male	2c	no	yes	yes	yes	yes
PD0278a	HNSC	BB	0	1	1	male	3	no	no	no	yes	yes
PD0279a	HNSC	BB	0	1	1	male	NA	yes	no	yes	yes	yes
PD0280a	HNSC	BB	0	1	1	male	4	no	no	no	yes	yes
PD0281a	HNSC	BB	0	1	1	male	4a	yes	no	yes	yes	yes
PD0282a	HNSC	BB	0	1	1	male	4	no	no	no	yes	yes
PD0283a	HNSC	BB	0	1	1	male	4	no	no	no	no	yes
PD0284a	HNSC	BB	0	1	1	male	4	no	no	no	yes	yes
PD0285a	HNSC	BB	0	1	1	male	3	no	no	yes	yes	yes
PD0286a	HNSC	BB	0	1	1	male	4	yes	no	no	yes	yes
PD0288a	HNSC	BB	0	1	1	male	4a	no	no	no	yes	yes
PD0289a	LUSC	BB	0	1	1	male	1a	no	yes	no	NA	no
PD0290a	LUSC	BB	0	1	1	male	3	no	no	no	yes	yes
PD0291a	LUSC	BB	0	1	1	male	3	no	yes	yes	no	no
PD0292a	LUSC	BB	0	1	1	male	NA	no	yes	yes	no	no
PD0293a	LUSC	BB	0	1	1	male	NA	no	no	no	no	yes
PD0294a	LUSC	BB	0	1	1	male	2b	no	yes	yes	no	no
PD0295a	LUSC	BB	0	1	1	male	2b	yes	no	yes	yes	yes
PD0296a	LUSC	BB	0	1	1	male	2b	no	no	no	yes	yes
PD0297a	LUSC	BB	0	1	1	male	2a	yes	yes	yes	NA	no
PD0298a	LUSC	BB	0	1	1	male	1b	no	no	no	yes	yes
PD0300a	LUSC	BB	0	1	1	male	2b	no	no	no	yes	yes
PD0301a	LUSC	BB	0	1	1	male	2a	yes	no	no	no	no
PD0302a	LUSC	BB	0	1	1	male	1b	no	no	no	yes	no
PD0303a	LUSC	BB	0	1	1	female	1b	no	no	no	yes	yes
PD0304a	LUSC	BB	0	1	1	female	3	no	no	no	no	no
PD0305a	LUSC	BB	0	1	1	male	1b	no	no	no	no	yes
PD0306a	LUSC	BB	0	1	1	male	4	no	no	no	yes	no
PD0308a	LUSC	BB	0	1	1	male	1a	no	no	no	yes	no
PD0309a	LUSC	BB	0	1	1	male	2b	no	no	yes	no	no
PD0310a	LUSC	BB	0	1	1	male	2a	no	no	yes	no	no
PD0311a	LUSC	BB	0	1	1	male	2a	yes	yes	yes	no	no
PD0312a	LUSC	BB	0	1	1	female	2a	no	no	no	yes	yes
PD0313a	LUSC	BB	0	1	1	male	2a	no	no	no	no	no
PD0314a	LUSC	BB	0	1	1	male	1b	no	no	no	no	no
PD0315a	LUSC	BB	0	1	1	male	1a	no	no	no	no	no
PD0316a	LUSC	BB	0	1	1	male	1a	no	no	no	no	yes
PD0317a	LUSC	BB	0	1	1	male	2a	no	no	no	no	no
PD0318a	LUSC	BB	0	1	1	male	2b	no	yes	no	yes	yes
PD0320a	LUSC	BB	0	1	1	male	2b	no	no	no	no	no
PD0321a	LUSC	BB	0	1	1	female	2b	no	no	no	yes	yes
PD0322a	LUSC	BB	0	1	1	male	2a	no	no	no	no	no
PD0323a	LUSC	BB	0	1	1	male	2a	no	no	no	yes	yes
PD0324a	LUSC	BB	0	1	1	male	2a	no	yes	no	yes	no
PD0325a	LUSC	BB	0	1	1	male	3	no	yes	yes	yes	no
PD0326a	LUSC	BB	0	1	1	male	3	no	yes	yes	no	no
PD0327a	LUSC	BB	0	1	1	male	1c	no	no	no	no	yes

PD0328a	LUSC	BB	0	1	1	male	2a	no	no	no	no	yes
PD0329a	LUSC	BB	0	1	1	male	2a	no	no	no	yes	yes
PD0330a	LUSC	BB	0	1	1	male	2a	no	no	no	no	no
PD0332a	LUSC	BB	0	1	1	male	1c	no	no	no	no	no
PD0333a	LUSC	BB	0	1	1	male	2b	no	no	no	yes	yes
PD0334a	LUSC	BB	0	1	1	male	3	no	no	no	no	yes
PD0335a	LUSC	BB	0	1	1	male	4	yes	yes	no	no	yes
PD0336a	LUSC	BB	0	1	1	male	2b	no	no	no	no	no
PD0337a	LUSC	BB	0	1	1	male	2a	no	no	no	no	no
PD0338a	LUSC	BB	0	1	1	male	2a	no	no	no	no	no
PD0339a	COAD	BB	0	0	1	male	NA	no	yes	yes	yes	yes
PD0340a	COAD	BB	0	0	1	female	NA	no	no	no	no	no
PD0341a	COAD	BB	0	0	1	female	NA	no	no	no	no	no
PD0342a	COAD	BB	0	0	1	male	NA	no	yes	no	yes	yes
PD0343a	COAD	BB	0	0	1	male	NA	no	yes	yes	yes	yes
PD0344a	COAD	BB	0	0	1	male	NA	no	yes	no	yes	no
PD0345a	COAD	BB	0	0	1	male	NA	no	yes	no	no	no
PD0346a	COAD	BB	0	0	1	female	NA	no	no	no	yes	no
PD0347a	COAD	BB	0	0	1	male	NA	yes	yes	yes	no	no
PD0348a	COAD	BB	0	0	1	male	NA	no	no	no	no	no
PD0349a	COAD	BB	0	0	1	male	NA	no	yes	no	no	no
PD0350a	COAD	BB	0	0	1	male	NA	no	no	yes	yes	yes
PD0352a	COAD	BB	0	0	1	female	NA	no	yes	no	no	no
PD0353a	COAD	BB	0	0	1	female	NA	no	no	no	no	no
PD0354a	COAD	BB	0	0	1	male	NA	no	no	no	yes	yes
PD0355a	COAD	BB	0	0	1	male	NA	no	no	no	no	no
PD0356a	COAD	BB	0	0	1	female	NA	no	no	no	no	no
PD0357a	COAD	BB	0	0	1	female	NA	no	no	no	no	no
PD0358a	COAD	BB	0	0	1	female	NA	no	yes	no	no	no
PD0359a	COAD	BB	0	0	1	female	NA	no	no	no	no	no
PD0360a	COAD	BB	0	0	1	male	NA	no	no	no	no	no
PD0362a	COAD	BB	0	0	1	male	NA	no	no	no	yes	no
PD0363a	COAD	BB	0	0	1	male	NA	no	no	no	yes	no
PD0364a	COAD	BB	0	0	1	male	NA	no	no	no	no	no
PD0365a	COAD	BB	0	0	1	male	NA	no	no	no	no	no
PD0366a	COAD	BB	0	0	1	male	NA	no	yes	no	yes	no
PD0367a	COAD	BB	0	0	1	female	NA	no	yes	yes	no	no
PD0368a	COAD	BB	0	0	1	male	NA	no	yes	no	yes	no
PD0369a	COAD	BB	0	0	1	male	NA	no	yes	NA	no	no
PD0370a	COAD	BB	0	0	1	female	NA	no	yes	yes	no	no
PD0371a	COAD	BB	0	0	1	female	NA	no	yes	no	no	no
PD0372a	COAD	BB	0	0	1	female	NA	no	no	no	yes	yes
PD0373a	COAD	BB	0	0	1	female	NA	no	no	no	no	no
PD0374a	COAD	BB	0	0	1	male	NA	no	no	no	yes	no
PD0375a	COAD	BB	0	0	1	female	NA	no	no	no	yes	no
PD0376a	COAD	BB	0	0	1	female	NA	no	yes	yes	yes	yes
PD0377a	COAD	BB	0	0	1	male	NA	no	no	no	yes	yes
PD0378a	COAD	BB	0	0	1	male	NA	no	yes	no	yes	no

PD0379a	COAD	BB	0	0	1	male	NA	no	yes	no	yes	yes
PD0380a	COAD	BB	0	0	1	female	NA	no	yes	no	no	no
PD0381a	COAD	BB	0	0	1	male	NA	no	yes	no	no	no
PD0382a	COAD	BB	0	0	1	male	NA	no	no	no	yes	yes
PD0383a	COAD	BB	0	0	1	male	NA	no	no	no	yes	yes
PD0384a	COAD	BB	0	0	1	female	NA	no	no	no	no	no
PD0385a	COAD	BB	0	0	1	male	NA	no	yes	yes	yes	no
PD0386a	COAD	BB	0	0	1	male	NA	no	yes	no	yes	yes
PD0387a	COAD	BB	0	0	1	male	NA	no	no	no	yes	no
PD0388a	COAD	BB	0	0	1	male	NA	no	no	yes	no	no

Supplementary Table 2. Description of laser-capture microdissected (LCM) biopsies included in the study. For each biopsy, the following metadata is reported: sample_id=unique biopsy identifier; donor_id=donor identifier; study=reference of the study to which it belongs; tissue_type=histological tissue type; nb_RT=number of somatic retrotransposition insertions detected; RT_class=categorical variable from the nb_RT; depth=mean sequencing depth; VAF=variant allele frequency of the main clone. Study references: [1] Coorens, T. H. H. et al. Inherent mosaicism and extensive mutation of human placentas. *Nature* 592, 80–85 (2021). [2] Moore, L., Cagan, A., Coorens, T.H.H. et al. The mutational landscape of human somatic and germline cells. *Nature* 597, 381–386 (2021). [3] Coorens, T.H.H., Moore, L., Robinson, P.S. et al. Extensive phylogenies of human development inferred from somatic mutations. *Nature* 597, 387–392 (2021).

sample_id	donor_id	study	tissue_type	nb_RT	RT_class	depth	VAF
PD41760b_lo0022	PD41760	NA	Stomach	20	[10,25)	27,92	0,405
PD40293c_lo0061	PD40293	NA	Stomach	9	[1,10)	NA	NA
PD40295c_lo0056	PD40295	NA	Stomach	9	[1,10)	23,76	0,386
PD40295c_lo0066	PD40295	NA	Stomach	9	[1,10)	21,66	0,418
PD40295c_lo0044	PD40295	NA	Stomach	5	[1,10)	20,99	0,524
PD40295c_lo0064	PD40295	NA	Stomach	4	[1,10)	18,62	0,367
PD45567b_lo0004	PD45567	[1]	Placenta	4	[1,10)	23,31	0,387
PD28690cc_COL_1_B12	PD28690	[2,3]	Colon	3	[1,10)	15,39	0,463
PD28690cc_COL_2_D8	PD28690	[2,3]	Colon	3	[1,10)	13,48	0,394
PD41763b_lo0002	PD41763	NA	Stomach	3	[1,10)	NA	NA
PD41763b_lo0007	PD41763	NA	Stomach	3	[1,10)	29,48	0,478
PD42146b2_lo0008	PD42146	[1]	Placenta	3	[1,10)	NA	NA
PD42787f_lo0002	PD42787	NA	Stomach	3	[1,10)	26,91	0,296
PD43850g_P50_CLN_B9	PD43850	[2,3]	Colon	3	[1,10)	15,11	0,443
PD43851j_P52_DDM_D3	PD43851	[2,3]	Small bowel	3	[1,10)	19,74	0,466
PD45557d_lo0005	PD45557	[1]	Placenta	3	[1,10)	51,62	0,425
PD45567b_lo0005	PD45567	[1]	Placenta	3	[1,10)	20,71	0,318
PD45567b_lo0007	PD45567	[1]	Placenta	3	[1,10)	26,57	0,253
PD28690cc_COL_1_C11	PD28690	[2,3]	Colon	2	[1,10)	18,91	0,416
PD28690cc_COL_2_F9	PD28690	[2,3]	Colon	2	[1,10)	24,23	0,417
PD28690cc_COL_2_H8	PD28690	[2,3]	Colon	2	[1,10)	27,45	0,420
PD41754b_lo0048	PD41754	NA	Stomach	2	[1,10)	22,83	0,491
PD41763b_lo0004	PD41763	NA	Stomach	2	[1,10)	34,11	0,447
PD41764b_lo0006	PD41764	NA	Stomach	2	[1,10)	31,31	0,436
PD42142b4_lo0019	PD42142	[1]	Placenta	2	[1,10)	NA	NA
PD42142b_lo0017	PD42142	[1]	Placenta	2	[1,10)	33,10	0,458
PD42146b4_lo0005	PD42146	[1]	Placenta	2	[1,10)	60,70	0,424
PD43850h_P53_DDM_A10	PD43850	[2,3]	Small bowel	2	[1,10)	27,86	0,441
PD43850w_P53_STM_G9	PD43850	[2,3]	Stomach	2	[1,10)	NA	NA
PD43851j_P52_DDM_D2	PD43851	[2,3]	Small bowel	2	[1,10)	22,07	0,292
PD45557b_lo0007	PD45557	[1]	Placenta	2	[1,10)	62,23	0,461
PD45566b_lo0006	PD45566	[1]	Placenta	2	[1,10)	23,40	0,251
PD45566b_lo0009	PD45566	[1]	Placenta	2	[1,10)	34,14	0,403
PD45566b_lo0010	PD45566	[1]	Placenta	2	[1,10)	NA	NA
PD45566c_lo0008	PD45566	[1]	Placenta	2	[1,10)	34,41	0,337
PD45567b_lo0003	PD45567	[1]	Placenta	2	[1,10)	NA	NA
PD28690bp_SB1_B8	PD28690	[2,3]	Small bowel	1	[1,10)	38,43	0,262
PD28690bv_APP1_G3	PD28690	[2,3]	Appendix	1	[1,10)	40,53	0,446
PD28690bw_APP_3_B2	PD28690	[2,3]	Appendix	1	[1,10)	25,21	0,488
PD28690bw_APP_3_C2	PD28690	[2,3]	Appendix	1	[1,10)	26,59	0,368
PD28690bw_APP_3_D1	PD28690	[2,3]	Appendix	1	[1,10)	27,62	0,438
PD28690bw_APP_3_D2	PD28690	[2,3]	Appendix	1	[1,10)	28,11	0,443
PD28690cc_COL_1_B11	PD28690	[2,3]	Colon	1	[1,10)	20,93	0,468

PD40293c_lo0013	PD40293	NA	Stomach	1	[1,10)	NA	NA
PD40295c_lo0065	PD40295	NA	Stomach	1	[1,10)	20,40	0,416
PD41753c_lo0008	PD41753	NA	Stomach	1	[1,10)	28,02	0,277
PD41753c_lo0012	PD41753	NA	Stomach	1	[1,10)	23,94	0,502
PD41756b_lo0009	PD41756	NA	Stomach	1	[1,10)	28,71	0,413
PD41760b_lo0010	PD41760	NA	Stomach	1	[1,10)	24,19	0,438
PD41760b_lo0025	PD41760	NA	Stomach	1	[1,10)	25,89	0,369
PD41763b_lo0014	PD41763	NA	Stomach	1	[1,10)	NA	NA
PD41763b_lo0015	PD41763	NA	Stomach	1	[1,10)	NA	NA
PD41764b_lo0020	PD41764	NA	Stomach	1	[1,10)	30,45	0,399
PD41764b_lo0025	PD41764	NA	Stomach	1	[1,10)	31,06	0,474
PD41764b_lo0026	PD41764	NA	Stomach	1	[1,10)	33,91	0,442
PD41765a_lo0001	PD41765	NA	Stomach	1	[1,10)	23,33	0,496
PD41766a_lo0034	PD41766	NA	Stomach	1	[1,10)	23,94	0,472
PD42138b2_lo0003	PD42138	[1]	Placenta	1	[1,10)	NA	NA
PD42138b2_lo0004	PD42138	[1]	Placenta	1	[1,10)	NA	NA
PD42138b2_lo0006	PD42138	[1]	Placenta	1	[1,10)	NA	NA
PD42138b2_lo0007	PD42138	[1]	Placenta	1	[1,10)	26,16	0,506
PD42138b2_lo0009	PD42138	[1]	Placenta	1	[1,10)	27,31	0,460
PD42138b3_lo0009	PD42138	[1]	Placenta	1	[1,10)	32,34	0,469
PD42138b3_lo0012	PD42138	[1]	Placenta	1	[1,10)	38,47	0,374
PD42138b3_lo0026	PD42138	[1]	Placenta	1	[1,10)	24,13	0,284
PD42138b_lo0009	PD42138	[1]	Placenta	1	[1,10)	30,92	0,458
PD42138b_lo0015	PD42138	[1]	Placenta	1	[1,10)	29,92	0,463
PD42138b_lo0016	PD42138	[1]	Placenta	1	[1,10)	28,41	0,481
PD42138b_lo0021	PD42138	[1]	Placenta	1	[1,10)	NA	NA
PD42138b_lo0022	PD42138	[1]	Placenta	1	[1,10)	30,16	0,253
PD42142b3_lo0010	PD42142	[1]	Placenta	1	[1,10)	18,94	0,397
PD42142b3_lo0016	PD42142	[1]	Placenta	1	[1,10)	34,33	0,454
PD42142b3_lo0018	PD42142	[1]	Placenta	1	[1,10)	34,14	0,398
PD42142b3_lo0020	PD42142	[1]	Placenta	1	[1,10)	19,09	0,304
PD42142b_lo0013	PD42142	[1]	Placenta	1	[1,10)	23,71	0,466
PD42142b_lo0023	PD42142	[1]	Placenta	1	[1,10)	NA	NA
PD42146b2_lo0007	PD42146	[1]	Placenta	1	[1,10)	48,10	0,420
PD42146b3_lo0011	PD42146	[1]	Placenta	1	[1,10)	28,52	0,492
PD42146b_lo0009	PD42146	[1]	Placenta	1	[1,10)	NA	NA
PD42787e_lo0009	PD42787	NA	Stomach	1	[1,10)	19,12	0,409
PD42787f_lo0008	PD42787	NA	Stomach	1	[1,10)	19,61	0,449
PD43850f_P51_CCM_B3	PD43850	[2,3]	Colon	1	[1,10)	28,97	0,486
PD43850g_P53_CLN_G10	PD43850	[2,3]	Colon	1	[1,10)	20,27	0,417
PD43850l_P50_ILM_B5	PD43850	[2,3]	Small bowel	1	[1,10)	16,20	0,480
PD43850l_P50_ILM_H4	PD43850	[2,3]	Small bowel	1	[1,10)	16,87	0,482
PD43850m_P50_JNM_E7	PD43850	[2,3]	Small bowel	1	[1,10)	16,07	0,495
PD43850u_P50_RTM_A10	PD43850	[2,3]	Colon	1	[1,10)	27,18	0,500
PD43850u_P50_RTM_E10	PD43850	[2,3]	Colon	1	[1,10)	11,75	0,469
PD43850w_P53_STM_D9	PD43850	[2,3]	Stomach	1	[1,10)	22,20	0,361
PD43851i_P52_CLN_B6	PD43851	[2,3]	Colon	1	[1,10)	25,09	0,437
PD43851i_P52_CLN_F5	PD43851	[2,3]	Colon	1	[1,10)	30,88	0,447
PD43851i_P52_CLN_F6	PD43851	[2,3]	Colon	1	[1,10)	29,55	0,389
PD43851i_P52_CLN_G6	PD43851	[2,3]	Colon	1	[1,10)	26,49	0,435
PD43851j_P52_DDM_B4	PD43851	[2,3]	Small bowel	1	[1,10)	16,71	0,464
PD43851j_P52_DDM_C3	PD43851	[2,3]	Small bowel	1	[1,10)	19,08	0,459
PD43851j_P52_DDM_F3	PD43851	[2,3]	Small bowel	1	[1,10)	24,58	0,392
PD43851j_P52_DDM_G2	PD43851	[2,3]	Small bowel	1	[1,10)	20,31	0,406
PD43851j_P52_DDM_G3	PD43851	[2,3]	Small bowel	1	[1,10)	20,27	0,441
PD43851x_P52_STM_B10	PD43851	[2,3]	Stomach	1	[1,10)	18,73	0,432

PD45557b_lo0001	PD45557	[1]	Placenta	1	[1,10)	32,90	0,480
PD45557b_lo0003	PD45557	[1]	Placenta	1	[1,10)	37,96	0,428
PD45557b_lo0006	PD45557	[1]	Placenta	1	[1,10)	33,85	0,458
PD45557b_lo0008	PD45557	[1]	Placenta	1	[1,10)	NA	NA
PD45557c_lo0003	PD45557	[1]	Placenta	1	[1,10)	45,38	0,490
PD45557d_lo0002	PD45557	[1]	Placenta	1	[1,10)	NA	NA
PD45557d_lo0003	PD45557	[1]	Placenta	1	[1,10)	NA	NA
PD45557e_lo0003	PD45557	[1]	Placenta	1	[1,10)	54,37	0,433
PD45557e_lo0006	PD45557	[1]	Placenta	1	[1,10)	33,47	0,431
PD45557e_lo0008	PD45557	[1]	Placenta	1	[1,10)	NA	NA
PD45566b_lo0001	PD45566	[1]	Placenta	1	[1,10)	22,17	0,266
PD45566b_lo0002	PD45566	[1]	Placenta	1	[1,10)	NA	NA
PD45566c_lo0003	PD45566	[1]	Placenta	1	[1,10)	34,65	0,438
PD45566c_lo0005	PD45566	[1]	Placenta	1	[1,10)	24,60	0,290
PD45566d_lo0002	PD45566	[1]	Placenta	1	[1,10)	NA	NA
PD45566d_lo0007	PD45566	[1]	Placenta	1	[1,10)	NA	NA
PD45566d_lo0008	PD45566	[1]	Placenta	1	[1,10)	NA	NA
PD45566d_lo0009	PD45566	[1]	Placenta	1	[1,10)	20,43	0,264
PD45566d_lo0010	PD45566	[1]	Placenta	1	[1,10)	37,70	0,378
PD45566d_lo0011	PD45566	[1]	Placenta	1	[1,10)	28,90	0,449
PD45566e_lo0002	PD45566	[1]	Placenta	1	[1,10)	22,55	0,271
PD45566e_lo0012	PD45566	[1]	Placenta	1	[1,10)	24,52	0,282
PD45567b_lo0012	PD45567	[1]	Placenta	1	[1,10)	22,76	0,301
PD45567c_lo0001	PD45567	[1]	Placenta	1	[1,10)	NA	NA
PD45567c_lo0007	PD45567	[1]	Placenta	1	[1,10)	27,72	0,390
PD45567d_lo0003	PD45567	[1]	Placenta	1	[1,10)	33,05	0,447
PD45567e_lo0004	PD45567	[1]	Placenta	1	[1,10)	NA	NA
PD45567e_lo0009	PD45567	[1]	Placenta	1	[1,10)	27,98	0,464
PD28690bp_SB1_A9	PD28690	[2,3]	Small bowel	0	[0,1)	19,27	0,361
PD28690bp_SB1_B9	PD28690	[2,3]	Small bowel	0	[0,1)	26,65	0,307
PD28690bp_SB1_C9	PD28690	[2,3]	Small bowel	0	[0,1)	5,38	0,495
PD28690bp_SB1_D8	PD28690	[2,3]	Small bowel	0	[0,1)	27,18	0,449
PD28690bp_SB1_D9	PD28690	[2,3]	Small bowel	0	[0,1)	3,90	0,591
PD28690bp_SB1_E9	PD28690	[2,3]	Small bowel	0	[0,1)	26,91	0,255
PD28690bp_SB1_G8	PD28690	[2,3]	Small bowel	0	[0,1)	17,77	0,282
PD28690bp_SB1_G9	PD28690	[2,3]	Small bowel	0	[0,1)	28,98	0,274
PD28690bp_SB1_H8	PD28690	[2,3]	Small bowel	0	[0,1)	15,55	0,361
PD28690bp_SB1_H9	PD28690	[2,3]	Small bowel	0	[0,1)	50,62	0,419
PD28690bt_SB2_A11	PD28690	[2,3]	Small bowel	0	[0,1)	21,39	0,348
PD28690bt_SB2_C11	PD28690	[2,3]	Small bowel	0	[0,1)	13,68	0,441
PD28690bt_SB2_E11	PD28690	[2,3]	Small bowel	0	[0,1)	14,69	0,432
PD28690bt_SB2_F10	PD28690	[2,3]	Small bowel	0	[0,1)	23,80	0,482
PD28690bt_SB2_F11	PD28690	[2,3]	Small bowel	0	[0,1)	23,96	0,488
PD28690bt_SB2_G10	PD28690	[2,3]	Small bowel	0	[0,1)	17,53	0,394
PD28690bt_SB2_G11	PD28690	[2,3]	Small bowel	0	[0,1)	15,26	0,349
PD28690bt_SB2_H10	PD28690	[2,3]	Small bowel	0	[0,1)	17,16	0,422
PD28690bt_SB3_B5	PD28690	[2,3]	Small bowel	0	[0,1)	19,17	0,404
PD28690bt_SB3_C5	PD28690	[2,3]	Small bowel	0	[0,1)	7,36	0,498
PD28690bt_SB3_E5	PD28690	[2,3]	Small bowel	0	[0,1)	12,62	0,436
PD28690bt_SB3_F5	PD28690	[2,3]	Small bowel	0	[0,1)	22,29	0,441
PD28690bv_APP1_B1	PD28690	[2,3]	Appendix	0	[0,1)	13,72	0,488
PD28690bv_APP1_B2	PD28690	[2,3]	Appendix	0	[0,1)	22,10	0,509
PD28690bv_APP1_C3	PD28690	[2,3]	Appendix	0	[0,1)	53,98	0,428
PD28690bv_APP1_F1	PD28690	[2,3]	Appendix	0	[0,1)	5,06	0,575
PD28690bv_APP1_F2	PD28690	[2,3]	Appendix	0	[0,1)	55,56	0,469
PD28690bv_APP1_F3	PD28690	[2,3]	Appendix	0	[0,1)	4,20	0,574

PD28690bv_APP1_H2	PD28690	[2,3]	Appendix	0	[0,1)	12,05	0,497
PD28690bv_APP_4_A7	PD28690	[2,3]	Appendix	0	[0,1)	35,51	0,469
PD28690bv_APP_4_A8	PD28690	[2,3]	Appendix	0	[0,1)	33,91	0,477
PD28690bv_APP_4_B7	PD28690	[2,3]	Appendix	0	[0,1)	4,11	0,575
PD28690bv_APP_4_C7	PD28690	[2,3]	Appendix	0	[0,1)	25,16	0,506
PD28690bv_APP_4_E7	PD28690	[2,3]	Appendix	0	[0,1)	NA	NA
PD28690bv_APP_4_F7	PD28690	[2,3]	Appendix	0	[0,1)	7,92	0,382
PD28690bv_APP_4_H7	PD28690	[2,3]	Appendix	0	[0,1)	8,95	0,358
PD28690bw_APP_3_A1	PD28690	[2,3]	Appendix	0	[0,1)	26,67	0,434
PD28690bw_APP_3_A2	PD28690	[2,3]	Appendix	0	[0,1)	26,75	0,348
PD28690bw_APP_3_C4	PD28690	[2,3]	Appendix	0	[0,1)	NA	NA
PD28690bw_APP_3_C5	PD28690	[2,3]	Appendix	0	[0,1)	30,66	0,440
PD28690bw_APP_3_D4	PD28690	[2,3]	Appendix	0	[0,1)	32,13	0,296
PD28690bw_APP_3_D5	PD28690	[2,3]	Appendix	0	[0,1)	27,06	0,464
PD28690bw_APP_3_E3	PD28690	[2,3]	Appendix	0	[0,1)	23,21	0,430
PD28690bw_APP_3_F1	PD28690	[2,3]	Appendix	0	[0,1)	30,61	0,459
PD28690bw_APP_3_F2	PD28690	[2,3]	Appendix	0	[0,1)	35,25	0,416
PD28690bw_APP_3_F3	PD28690	[2,3]	Appendix	0	[0,1)	23,80	0,394
PD28690bw_APP_3_F4	PD28690	[2,3]	Appendix	0	[0,1)	26,77	0,378
PD28690bw_APP_3_G3	PD28690	[2,3]	Appendix	0	[0,1)	23,66	0,499
PD28690bw_APP_3_G4	PD28690	[2,3]	Appendix	0	[0,1)	30,31	0,454
PD28690bw_APP_3_H3	PD28690	[2,3]	Appendix	0	[0,1)	24,77	0,471
PD28690bw_APP_3_H4	PD28690	[2,3]	Appendix	0	[0,1)	27,19	0,463
PD28690bx_2_a5	PD28690	[2,3]	Colon	0	[0,1)	16,26	0,410
PD28690bx_2_d5	PD28690	[2,3]	Colon	0	[0,1)	17,03	0,447
PD28690cb_2_g1	PD28690	[2,3]	Colon	0	[0,1)	16,54	0,396
PD28690cb_2_g3	PD28690	[2,3]	Colon	0	[0,1)	15,59	0,445
PD28690cb_2_h1	PD28690	[2,3]	Colon	0	[0,1)	16,71	0,354
PD28690cb_2_h3	PD28690	[2,3]	Colon	0	[0,1)	18,40	0,422
PD28690cc_COL_1_A11	PD28690	[2,3]	Colon	0	[0,1)	NA	NA
PD28690cc_COL_1_C12	PD28690	[2,3]	Colon	0	[0,1)	15,30	0,415
PD28690cc_COL_1_D12	PD28690	[2,3]	Colon	0	[0,1)	10,69	0,343
PD28690cc_COL_1_E12	PD28690	[2,3]	Colon	0	[0,1)	8,77	0,514
PD28690cc_COL_1_G11	PD28690	[2,3]	Colon	0	[0,1)	6,61	0,584
PD28690cc_COL_1_H11	PD28690	[2,3]	Colon	0	[0,1)	4,75	0,605
PD28690cc_COL_2_B8	PD28690	[2,3]	Colon	0	[0,1)	29,92	0,438
PD28690cc_COL_2_E8	PD28690	[2,3]	Colon	0	[0,1)	8,73	0,496
PD28690cc_COL_2_G8	PD28690	[2,3]	Colon	0	[0,1)	33,09	0,450
PD28690cc_COL_2_G9	PD28690	[2,3]	Colon	0	[0,1)	25,91	0,466
PD28690cc_COL_4_A3	PD28690	[2,3]	Colon	0	[0,1)	16,49	0,475
PD28690cc_COL_4_B3	PD28690	[2,3]	Colon	0	[0,1)	8,46	0,498
PD28690cc_COL_4_C2	PD28690	[2,3]	Colon	0	[0,1)	13,57	0,452
PD28690cc_COL_4_C4	PD28690	[2,3]	Colon	0	[0,1)	12,58	0,475
PD28690cc_COL_4_E3	PD28690	[2,3]	Colon	0	[0,1)	11,55	0,474
PD28690cf_2_a10	PD28690	[2,3]	Colon	0	[0,1)	NA	NA
PD28690cf_2_d10	PD28690	[2,3]	Colon	0	[0,1)	17,55	0,277
PD28690cf_2_g11	PD28690	[2,3]	Colon	0	[0,1)	NA	NA
PD28690cf_2_h11	PD28690	[2,3]	Colon	0	[0,1)	NA	NA
PD28690id_T3_L1_A1	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L1_A2	PD28690	[2,3]	Testis	0	[0,1)	26,41	0,268
PD28690id_T3_L1_B1	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L1_B2	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L1_C2	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L1_D2	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L1_E1	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L1_E2	PD28690	[2,3]	Testis	0	[0,1)	NA	NA

PD28690id_T3_L1_F1	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L1_F2	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L1_G1	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L1_G2	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L2_A4	PD28690	[2,3]	Testis	0	[0,1)	18,99	0,341
PD28690id_T3_L2_A5	PD28690	[2,3]	Testis	0	[0,1)	13,03	0,435
PD28690id_T3_L2_B5	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L2_C3	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L2_C4	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L2_C5	PD28690	[2,3]	Testis	0	[0,1)	11,16	0,474
PD28690id_T3_L2_D3	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L2_E3	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L2_E4	PD28690	[2,3]	Testis	0	[0,1)	11,56	0,470
PD28690id_T3_L2_F3	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L2_F4	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L2_G3	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L2_G4	PD28690	[2,3]	Testis	0	[0,1)	9,35	0,528
PD28690id_T3_L4_A6	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L4_A7	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L4_B6	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L4_B7	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L4_C6	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L4_D6	PD28690	[2,3]	Testis	0	[0,1)	11,00	0,479
PD28690id_T3_L4_D7	PD28690	[2,3]	Testis	0	[0,1)	11,80	0,466
PD28690id_T3_L4_F5	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L4_F7	PD28690	[2,3]	Testis	0	[0,1)	11,94	0,461
PD28690id_T3_L4_G5	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L4_H5	PD28690	[2,3]	Testis	0	[0,1)	NA	NA
PD28690id_T3_L4_H6	PD28690	[2,3]	Testis	0	[0,1)	11,53	0,474
PD40293c_lo0067	PD40293	NA	Stomach	0	[0,1)	21,61	0,391
PD40293c_lo0068	PD40293	NA	Stomach	0	[0,1)	16,79	0,375
PD40293c_lo0069	PD40293	NA	Stomach	0	[0,1)	22,81	0,315
PD40293c_lo0071	PD40293	NA	Stomach	0	[0,1)	22,43	0,286
PD40293c_lo0073	PD40293	NA	Stomach	0	[0,1)	NA	NA
PD40293c_lo0076	PD40293	NA	Stomach	0	[0,1)	NA	NA
PD40293c_lo0173	PD40293	NA	Stomach	0	[0,1)	21,39	0,378
PD40293d_lo0013	PD40293	NA	Stomach	0	[0,1)	26,79	0,286
PD40293d_lo0023	PD40293	NA	Stomach	0	[0,1)	23,76	0,285
PD40293d_lo0025	PD40293	NA	Stomach	0	[0,1)	20,62	0,281
PD40294c_lo0003	PD40294	NA	Stomach	0	[0,1)	NA	NA
PD40294c_lo0040	PD40294	NA	Stomach	0	[0,1)	22,66	0,308
PD40294c_lo0044	PD40294	NA	Stomach	0	[0,1)	18,97	0,266
PD40294c_lo0046	PD40294	NA	Stomach	0	[0,1)	24,06	0,276
PD40294c_lo0048	PD40294	NA	Stomach	0	[0,1)	20,40	0,375
PD40294c_lo0058	PD40294	NA	Stomach	0	[0,1)	22,86	0,301
PD40294c_lo0072	PD40294	NA	Stomach	0	[0,1)	NA	NA
PD40294c_lo0076	PD40294	NA	Stomach	0	[0,1)	26,38	0,420
PD40294c_lo0078	PD40294	NA	Stomach	0	[0,1)	NA	NA
PD40294c_lo0087	PD40294	NA	Stomach	0	[0,1)	21,14	0,370
PD40294c_lo0088	PD40294	NA	Stomach	0	[0,1)	NA	NA
PD40294d_lo0017	PD40294	NA	Stomach	0	[0,1)	24,59	0,348
PD40294d_lo0019	PD40294	NA	Stomach	0	[0,1)	NA	NA
PD40294d_lo0032	PD40294	NA	Stomach	0	[0,1)	NA	NA
PD40295c_lo0051	PD40295	NA	Stomach	0	[0,1)	21,25	0,380
PD40296c_lo0018	PD40296	NA	Stomach	0	[0,1)	24,66	0,300
PD40296c_lo0034	PD40296	NA	Stomach	0	[0,1)	16,40	0,304

PD40297c_lo0005	PD40297	NA	Stomach	0	[0,1)	29,40	0,408
PD40297c_lo0049	PD40297	NA	Stomach	0	[0,1)	21,90	0,274
PD40297c_lo0059	PD40297	NA	Stomach	0	[0,1)	20,76	0,375
PD40297c_lo0061	PD40297	NA	Stomach	0	[0,1)	22,84	0,469
PD40297c_lo0069	PD40297	NA	Stomach	0	[0,1)	23,74	0,346
PD40297c_lo0073	PD40297	NA	Stomach	0	[0,1)	19,44	0,256
PD40297c_lo0074	PD40297	NA	Stomach	0	[0,1)	18,46	0,298
PD40297c_lo0077	PD40297	NA	Stomach	0	[0,1)	23,18	0,250
PD41751c_lo0017	PD41751	NA	Stomach	0	[0,1)	22,58	0,305
PD41751c_lo0030	PD41751	NA	Stomach	0	[0,1)	22,27	0,284
PD41751c_lo0031	PD41751	NA	Stomach	0	[0,1)	NA	NA
PD41752c_lo0005	PD41752	NA	Stomach	0	[0,1)	24,95	0,298
PD41752c_lo0006	PD41752	NA	Stomach	0	[0,1)	19,19	0,369
PD41752c_lo0019	PD41752	NA	Stomach	0	[0,1)	21,90	0,274
PD41752c_lo0022	PD41752	NA	Stomach	0	[0,1)	19,60	0,285
PD41752c_lo0023	PD41752	NA	Stomach	0	[0,1)	21,14	0,298
PD41753c_lo0003	PD41753	NA	Stomach	0	[0,1)	32,34	0,477
PD41753c_lo0004	PD41753	NA	Stomach	0	[0,1)	24,97	0,472
PD41753c_lo0005	PD41753	NA	Stomach	0	[0,1)	30,34	0,428
PD41753c_lo0006	PD41753	NA	Stomach	0	[0,1)	34,39	0,466
PD41753c_lo0007	PD41753	NA	Stomach	0	[0,1)	27,44	0,474
PD41753c_lo0009	PD41753	NA	Stomach	0	[0,1)	27,52	0,483
PD41753c_lo0010	PD41753	NA	Stomach	0	[0,1)	27,11	0,468
PD41753c_lo0011	PD41753	NA	Stomach	0	[0,1)	27,76	0,480
PD41754b_lo0025	PD41754	NA	Stomach	0	[0,1)	21,61	0,522
PD41754b_lo0033	PD41754	NA	Stomach	0	[0,1)	21,01	0,497
PD41754b_lo0045	PD41754	NA	Stomach	0	[0,1)	19,12	0,256
PD41757b_lo0002	PD41757	NA	Stomach	0	[0,1)	21,16	0,318
PD41757b_lo0003	PD41757	NA	Stomach	0	[0,1)	25,90	0,452
PD41757b_lo0006	PD41757	NA	Stomach	0	[0,1)	NA	NA
PD41757b_lo0007	PD41757	NA	Stomach	0	[0,1)	22,79	0,286
PD41757b_lo0008	PD41757	NA	Stomach	0	[0,1)	18,58	0,432
PD41760b_lo0020	PD41760	NA	Stomach	0	[0,1)	27,49	0,365
PD41760b_lo0030	PD41760	NA	Stomach	0	[0,1)	NA	NA
PD41763b_lo0003	PD41763	NA	Stomach	0	[0,1)	27,19	0,381
PD41763b_lo0005	PD41763	NA	Stomach	0	[0,1)	31,73	0,400
PD41763b_lo0006	PD41763	NA	Stomach	0	[0,1)	33,04	0,392
PD41763b_lo0010	PD41763	NA	Stomach	0	[0,1)	NA	NA
PD41763b_lo0011	PD41763	NA	Stomach	0	[0,1)	24,54	0,283
PD41764b_lo0004	PD41764	NA	Stomach	0	[0,1)	30,00	0,456
PD41765a_lo0008	PD41765	NA	Stomach	0	[0,1)	NA	NA
PD41765a_lo0023	PD41765	NA	Stomach	0	[0,1)	NA	NA
PD41765b_lo0004	PD41765	NA	Stomach	0	[0,1)	15,63	0,289
PD41765b_lo0005	PD41765	NA	Stomach	0	[0,1)	18,70	0,432
PD41765b_lo0010	PD41765	NA	Stomach	0	[0,1)	NA	NA
PD41765b_lo0011	PD41765	NA	Stomach	0	[0,1)	21,50	0,415
PD41765b_lo0012	PD41765	NA	Stomach	0	[0,1)	23,43	0,351
PD41766a_lo0032	PD41766	NA	Stomach	0	[0,1)	21,02	0,263
PD41766a_lo0033	PD41766	NA	Stomach	0	[0,1)	18,92	0,293
PD41766a_lo0035	PD41766	NA	Stomach	0	[0,1)	20,12	0,272
PD41766a_lo0036	PD41766	NA	Stomach	0	[0,1)	25,53	0,385
PD41766b_lo0004	PD41766	NA	Stomach	0	[0,1)	25,64	0,464
PD41766b_lo0005	PD41766	NA	Stomach	0	[0,1)	NA	NA
PD41766b_lo0006	PD41766	NA	Stomach	0	[0,1)	29,84	0,468
PD41766b_lo0007	PD41766	NA	Stomach	0	[0,1)	24,30	0,466
PD41766b_lo0012	PD41766	NA	Stomach	0	[0,1)	NA	NA

PD42138b3_lo0003	PD42138	[1]	Placenta	0	[0,1)	NA	NA
PD42138b3_lo0017	PD42138	[1]	Placenta	0	[0,1)	NA	NA
PD42138b3_lo0024	PD42138	[1]	Placenta	0	[0,1)	24,38	0,288
PD42138b4_lo0010	PD42138	[1]	Placenta	0	[0,1)	29,01	0,422
PD42138b4_lo0018	PD42138	[1]	Placenta	0	[0,1)	27,72	0,273
PD42138b_lo0018	PD42138	[1]	Placenta	0	[0,1)	NA	NA
PD42142b2_lo0012	PD42142	[1]	Placenta	0	[0,1)	NA	NA
PD42142b2_lo0014	PD42142	[1]	Placenta	0	[0,1)	33,08	0,459
PD42142b3_lo0021	PD42142	[1]	Placenta	0	[0,1)	21,59	0,337
PD42142b4_lo0006	PD42142	[1]	Placenta	0	[0,1)	NA	NA
PD42142b4_lo0009	PD42142	[1]	Placenta	0	[0,1)	NA	NA
PD42142b4_lo0020	PD42142	[1]	Placenta	0	[0,1)	23,21	0,313
PD42142b_lo0003	PD42142	[1]	Placenta	0	[0,1)	NA	NA
PD42142b_lo0007	PD42142	[1]	Placenta	0	[0,1)	NA	NA
PD42142b_lo0008	PD42142	[1]	Placenta	0	[0,1)	NA	NA
PD42142b_lo0010	PD42142	[1]	Placenta	0	[0,1)	NA	NA
PD42142b_lo0015	PD42142	[1]	Placenta	0	[0,1)	29,76	0,443
PD42142b_lo0019	PD42142	[1]	Placenta	0	[0,1)	30,32	0,425
PD42142b_lo0020	PD42142	[1]	Placenta	0	[0,1)	24,99	0,259
PD42146b2_lo0001	PD42146	[1]	Placenta	0	[0,1)	32,17	0,481
PD42146b3_lo0005	PD42146	[1]	Placenta	0	[0,1)	28,50	0,434
PD42146b3_lo0009	PD42146	[1]	Placenta	0	[0,1)	28,20	0,455
PD42146b3_lo0010	PD42146	[1]	Placenta	0	[0,1)	NA	NA
PD42146b3_lo0018	PD42146	[1]	Placenta	0	[0,1)	NA	NA
PD42146b3_lo0020	PD42146	[1]	Placenta	0	[0,1)	30,53	0,461
PD42146b3_lo0022	PD42146	[1]	Placenta	0	[0,1)	29,06	0,436
PD42146b4_lo0006	PD42146	[1]	Placenta	0	[0,1)	NA	NA
PD42146b4_lo0008	PD42146	[1]	Placenta	0	[0,1)	35,63	0,419
PD42146b_lo0005	PD42146	[1]	Placenta	0	[0,1)	33,75	0,426
PD42787b_lo0038	PD42787	NA	Stomach	0	[0,1)	17,37	0,307
PD42787c_lo0004	PD42787	NA	Stomach	0	[0,1)	22,46	0,398
PD42787c_lo0005	PD42787	NA	Stomach	0	[0,1)	17,52	0,388
PD42787c_lo0006	PD42787	NA	Stomach	0	[0,1)	NA	NA
PD42787c_lo0046	PD42787	NA	Stomach	0	[0,1)	23,46	0,499
PD42787c_lo0047	PD42787	NA	Stomach	0	[0,1)	24,33	0,372
PD42787d_lo0017	PD42787	NA	Stomach	0	[0,1)	24,57	0,284
PD42787d_lo0030	PD42787	NA	Stomach	0	[0,1)	NA	NA
PD42787d_lo0031	PD42787	NA	Stomach	0	[0,1)	18,21	0,393
PD42787d_lo0032	PD42787	NA	Stomach	0	[0,1)	13,23	0,327
PD42787e_lo0001	PD42787	NA	Stomach	0	[0,1)	NA	NA
PD42787e_lo0002	PD42787	NA	Stomach	0	[0,1)	22,46	0,383
PD42787e_lo0006	PD42787	NA	Stomach	0	[0,1)	27,70	0,457
PD42787e_lo0012	PD42787	NA	Stomach	0	[0,1)	19,31	0,459
PD42787f_lo0006	PD42787	NA	Stomach	0	[0,1)	27,66	0,455
PD42787f_lo0013	PD42787	NA	Stomach	0	[0,1)	22,05	0,296
PD42787f_lo0016	PD42787	NA	Stomach	0	[0,1)	18,72	0,408
PD42787f_lo0017	PD42787	NA	Stomach	0	[0,1)	19,10	0,378
PD43850f_P51_CCM_E3	PD43850	[2,3]	Colon	0	[0,1)	34,96	0,486
PD43850f_P51_CCM_F3	PD43850	[2,3]	Colon	0	[0,1)	27,28	0,455
PD43850f_P51_CCM_G3	PD43850	[2,3]	Colon	0	[0,1)	15,14	0,490
PD43850g_P50_CLN_C9	PD43850	[2,3]	Colon	0	[0,1)	27,54	0,436
PD43850g_P53_CLN_F10	PD43850	[2,3]	Colon	0	[0,1)	23,97	0,511
PD43850h_P50_DNM_A2	PD43850	[2,3]	Small bowel	0	[0,1)	12,69	0,510
PD43850h_P50_DNM_B4	PD43850	[2,3]	Small bowel	0	[0,1)	13,34	0,449
PD43850h_P50_DNM_C2	PD43850	[2,3]	Small bowel	0	[0,1)	10,75	0,506
PD43850h_P50_DNM_D1	PD43850	[2,3]	Small bowel	0	[0,1)	20,51	0,429

PD43850h_P53_DDM_B10	PD43850	[2,3]	Small bowel	0	[0,1)	29,21	0,450
PD43850h_P53_DDM_D10	PD43850	[2,3]	Small bowel	0	[0,1)	20,70	0,325
PD43850l_P50_ILM_A5	PD43850	[2,3]	Small bowel	0	[0,1)	14,81	0,488
PD43850l_P50_ILM_C5	PD43850	[2,3]	Small bowel	0	[0,1)	18,71	0,485
PD43850l_P50_ILM_E4	PD43850	[2,3]	Small bowel	0	[0,1)	20,79	0,465
PD43850l_P50_ILM_F4	PD43850	[2,3]	Small bowel	0	[0,1)	14,93	0,500
PD43850l_P50_ILM_G4	PD43850	[2,3]	Small bowel	0	[0,1)	28,58	0,449
PD43850m_P50_JNM_B7	PD43850	[2,3]	Small bowel	0	[0,1)	18,09	0,492
PD43850m_P50_JNM_C7	PD43850	[2,3]	Small bowel	0	[0,1)	14,44	0,499
PD43850m_P50_JNM_D6	PD43850	[2,3]	Small bowel	0	[0,1)	15,46	0,489
PD43850m_P50_JNM_E6	PD43850	[2,3]	Small bowel	0	[0,1)	16,20	0,489
PD43850m_P50_JNM_F6	PD43850	[2,3]	Small bowel	0	[0,1)	20,33	0,470
PD43850t_P51_PNC_A7	PD43850	[2,3]	Pancreas	0	[0,1)	21,72	0,327
PD43850t_P51_PNC_A8	PD43850	[2,3]	Pancreas	0	[0,1)	17,67	0,299
PD43850t_P51_PNC_B7	PD43850	[2,3]	Pancreas	0	[0,1)	22,09	0,254
PD43850t_P51_PNC_C7	PD43850	[2,3]	Pancreas	0	[0,1)	19,52	0,295
PD43850t_P51_PNC_C8	PD43850	[2,3]	Pancreas	0	[0,1)	13,09	0,399
PD43850t_P51_PNC_D7	PD43850	[2,3]	Pancreas	0	[0,1)	NA	NA
PD43850t_P51_PNC_D8	PD43850	[2,3]	Pancreas	0	[0,1)	16,88	0,339
PD43850t_P51_PNC_D9	PD43850	[2,3]	Pancreas	0	[0,1)	18,47	0,320
PD43850t_P51_PNC_E7	PD43850	[2,3]	Pancreas	0	[0,1)	NA	NA
PD43850t_P51_PNC_E9	PD43850	[2,3]	Pancreas	0	[0,1)	13,86	0,374
PD43850t_P51_PNC_F7	PD43850	[2,3]	Pancreas	0	[0,1)	19,16	0,327
PD43850t_P51_PNC_F8	PD43850	[2,3]	Pancreas	0	[0,1)	12,47	0,395
PD43850t_P51_PNC_G7	PD43850	[2,3]	Pancreas	0	[0,1)	NA	NA
PD43850t_P51_PNC_G8	PD43850	[2,3]	Pancreas	0	[0,1)	15,38	0,390
PD43850t_P51_PNC_H8	PD43850	[2,3]	Pancreas	0	[0,1)	17,26	0,327
PD43850u_P50_RTM_B10	PD43850	[2,3]	Colon	0	[0,1)	21,88	0,494
PD43850u_P50_RTM_C10	PD43850	[2,3]	Colon	0	[0,1)	12,09	0,532
PD43850u_P50_RTM_H9	PD43850	[2,3]	Colon	0	[0,1)	27,94	0,493
PD43850v_P50_SKN_F12	PD43850	[2,3]	Skin	0	[0,1)	13,61	0,307
PD43850v_P50_SKN_G12	PD43850	[2,3]	Skin	0	[0,1)	19,07	0,336
PD43850v_P50_SKN_H12	PD43850	[2,3]	Skin	0	[0,1)	16,24	0,392
PD43850v_P53_SKN_D8	PD43850	[2,3]	Skin	0	[0,1)	NA	NA
PD43850w_P53_STM_B9	PD43850	[2,3]	Stomach	0	[0,1)	14,80	0,362
PD43850w_P53_STM_F9	PD43850	[2,3]	Stomach	0	[0,1)	12,39	0,432
PD43850w_P53_STM_H9	PD43850	[2,3]	Stomach	0	[0,1)	13,61	0,493
PD43851f_P52_CLN_C11	PD43851	[2,3]	Colon	0	[0,1)	22,38	0,470
PD43851f_P53_CLN_A6	PD43851	[2,3]	Colon	0	[0,1)	14,17	0,510
PD43851f_P53_CLN_B6	PD43851	[2,3]	Colon	0	[0,1)	17,05	0,479
PD43851f_P53_CLN_C6	PD43851	[2,3]	Colon	0	[0,1)	16,31	0,501
PD43851i_P52_CLN_A6	PD43851	[2,3]	Colon	0	[0,1)	23,33	0,440
PD43851i_P52_CLN_C6	PD43851	[2,3]	Colon	0	[0,1)	19,55	0,461
PD43851i_P52_CLN_D6	PD43851	[2,3]	Colon	0	[0,1)	20,60	0,472
PD43851i_P52_CLN_E5	PD43851	[2,3]	Colon	0	[0,1)	19,39	0,461
PD43851i_P52_CLN_E6	PD43851	[2,3]	Colon	0	[0,1)	19,64	0,465
PD43851i_P52_CLN_H5	PD43851	[2,3]	Colon	0	[0,1)	23,60	0,441
PD43851i_P52_CLN_H6	PD43851	[2,3]	Colon	0	[0,1)	18,97	0,369
PD43851j_P52_DDM_B5	PD43851	[2,3]	Small bowel	0	[0,1)	16,03	0,347
PD43851j_P52_DDM_C4	PD43851	[2,3]	Small bowel	0	[0,1)	12,02	0,463
PD43851j_P52_DDM_E2	PD43851	[2,3]	Small bowel	0	[0,1)	NA	NA
PD43851j_P52_DDM_E3	PD43851	[2,3]	Small bowel	0	[0,1)	16,02	0,436
PD43851j_P52_DDM_H2	PD43851	[2,3]	Small bowel	0	[0,1)	24,61	0,347
PD43851j_P52_DDM_H3	PD43851	[2,3]	Small bowel	0	[0,1)	15,38	0,423
PD43851m_P52_ILM_A1	PD43851	[2,3]	Small bowel	0	[0,1)	13,14	0,499
PD43851m_P52_ILM_D1	PD43851	[2,3]	Small bowel	0	[0,1)	18,64	0,485

PD43851m_P52_ILM_H1	PD43851	[2,3]	Small bowel	0	[0,1)	15,83	0,341
PD43851s_P53_PNC_B4	PD43851	[2,3]	Pancreas	0	[0,1)	14,51	0,332
PD43851s_P53_PNC_D4	PD43851	[2,3]	Pancreas	0	[0,1)	15,74	0,286
PD43851s_P53_PNC_E4	PD43851	[2,3]	Pancreas	0	[0,1)	NA	NA
PD43851s_P53_PNC_F4	PD43851	[2,3]	Pancreas	0	[0,1)	15,57	0,293
PD43851w_P52_SKN_A12	PD43851	[2,3]	Skin	0	[0,1)	NA	NA
PD43851w_P52_SKN_B12	PD43851	[2,3]	Skin	0	[0,1)	NA	NA
PD43851w_P52_SKN_D12	PD43851	[2,3]	Skin	0	[0,1)	NA	NA
PD43851w_P53_SKN_A1	PD43851	[2,3]	Skin	0	[0,1)	NA	NA
PD43851w_P53_SKN_B1	PD43851	[2,3]	Skin	0	[0,1)	NA	NA
PD43851w_P53_SKN_E1	PD43851	[2,3]	Skin	0	[0,1)	23,48	0,252
PD43851w_P53_SKN_F1	PD43851	[2,3]	Skin	0	[0,1)	23,40	0,256
PD43851w_P53_SKN_H1	PD43851	[2,3]	Skin	0	[0,1)	22,36	0,290
PD43851x_P52_STM_C10	PD43851	[2,3]	Stomach	0	[0,1)	25,05	0,496
PD43851x_P52_STM_C9	PD43851	[2,3]	Stomach	0	[0,1)	13,27	0,431
PD43851x_P52_STM_D10	PD43851	[2,3]	Stomach	0	[0,1)	18,16	0,449
PD43851x_P52_STM_G10	PD43851	[2,3]	Stomach	0	[0,1)	17,98	0,485
PD43851x_P52_STM_G9	PD43851	[2,3]	Stomach	0	[0,1)	18,52	0,379
PD43851x_P52_STM_H10	PD43851	[2,3]	Stomach	0	[0,1)	16,90	0,395
PD43851y_P52_TST_A7	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P52_TST_A8	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P52_TST_B7	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P52_TST_B8	PD43851	[2,3]	Testis	0	[0,1)	44,96	0,259
PD43851y_P52_TST_C7	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P52_TST_C8	PD43851	[2,3]	Testis	0	[0,1)	40,25	0,280
PD43851y_P52_TST_D7	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P52_TST_D8	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P52_TST_E7	PD43851	[2,3]	Testis	0	[0,1)	46,12	0,353
PD43851y_P52_TST_E8	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P52_TST_F7	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P52_TST_G8	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P52_TST_H7	PD43851	[2,3]	Testis	0	[0,1)	59,96	0,435
PD43851y_P52_TST_H8	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P53_TST_A5	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P53_TST_B5	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P53_TST_C5	PD43851	[2,3]	Testis	0	[0,1)	65,29	0,259
PD43851y_P53_TST_D5	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P53_TST_E5	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P53_TST_F5	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P53_TST_G5	PD43851	[2,3]	Testis	0	[0,1)	NA	NA
PD43851y_P53_TST_H5	PD43851	[2,3]	Testis	0	[0,1)	74,31	0,438
PD45518d_lo0025	PD45518	NA	Stomach	0	[0,1)	NA	NA
PD45518d_lo0037	PD45518	NA	Stomach	0	[0,1)	15,27	0,398
PD45518d_lo0038	PD45518	NA	Stomach	0	[0,1)	14,98	0,460
PD45518d_lo0042	PD45518	NA	Stomach	0	[0,1)	13,45	0,414
PD45557c_lo0002	PD45557	[1]	Placenta	0	[0,1)	20,39	0,300
PD45557c_lo0004	PD45557	[1]	Placenta	0	[0,1)	27,61	0,501
PD45557c_lo0006	PD45557	[1]	Placenta	0	[0,1)	87,48	0,457
PD45557c_lo0009	PD45557	[1]	Placenta	0	[0,1)	NA	NA
PD45557c_lo0010	PD45557	[1]	Placenta	0	[0,1)	NA	NA
PD45557d_lo0007	PD45557	[1]	Placenta	0	[0,1)	NA	NA
PD45557e_lo0002	PD45557	[1]	Placenta	0	[0,1)	NA	NA
PD45557e_lo0005	PD45557	[1]	Placenta	0	[0,1)	NA	NA
PD45566b_lo0004	PD45566	[1]	Placenta	0	[0,1)	NA	NA
PD45566c_lo0004	PD45566	[1]	Placenta	0	[0,1)	NA	NA
PD45566c_lo0009	PD45566	[1]	Placenta	0	[0,1)	19,74	0,315

PD45566c_lo0011	PD45566	[1]	Placenta	0	[0,1)	NA	NA
PD45566e_lo0001	PD45566	[1]	Placenta	0	[0,1)	20,97	0,267
PD45566e_lo0007	PD45566	[1]	Placenta	0	[0,1)	NA	NA
PD45566e_lo0008	PD45566	[1]	Placenta	0	[0,1)	NA	NA
PD45566e_lo0009	PD45566	[1]	Placenta	0	[0,1)	27,92	0,461
PD45567b_lo0008	PD45567	[1]	Placenta	0	[0,1)	NA	NA
PD45567c_lo0003	PD45567	[1]	Placenta	0	[0,1)	32,60	0,437
PD45567c_lo0005	PD45567	[1]	Placenta	0	[0,1)	15,81	0,273
PD45567c_lo0006	PD45567	[1]	Placenta	0	[0,1)	25,11	0,276
PD45567c_lo0009	PD45567	[1]	Placenta	0	[0,1)	NA	NA
PD45567e_lo0002	PD45567	[1]	Placenta	0	[0,1)	40,03	0,465
PD45567e_lo0008	PD45567	[1]	Placenta	0	[0,1)	36,55	0,464
PD45567e_lo0011	PD45567	[1]	Placenta	0	[0,1)	24,11	0,269

	O/E length ratio			*Pairwise identity*		
	Median	**Mean**	**Sd**	**Median**	**Mean**	**Sd**
Alu-solo	1,000	0,999	0,0039	100,000	99,926	0,4372
L1-orphan	1,000	1,000	0,0035	100,000	99,947	0,4567
L1-partnered	1,000	1,000	0,0141	99,978	99,873	0,7745
L1-solo	1,000	1,000	0,0047	100,000	99,898	0,5104
SVA-solo	1,000	1,000	0,0039	100,000	99,888	0,6777

Supplementary Table 3. Statistics of the sequence reconstructions performed by MEIGA across different RT insertion types. Stat values were estimated using simulated data. These included pairwise identity percentage and observed-to-expected length ratio (O/E). Long Interspersed Nucleotide Element 1; RT: Retrotransposition; Sd: Standard deviation; SVA: SINE-VNTR-*Alu*.

Samples	DEL	DUP	INV	TRA	RT/None_INV+TRA	RT/None_rINV	RT/None_rTRA	RT/RT_3REF rTRA	RT/RT_DEL+DUP	RT/RT_DUP+TRA	RT/RT_INV+TRA	RT/RT_rINV	RT/RT_rTRA	TOTAL
PD0266a	2	0	0	2	0	0	0	0	0	0	0	0	0	**4**
PD0270a	8	0	1	1	0	1	0	0	0	1	0	0	4	**16**
PD0274a	9	0	2	1	0	0	0	0	0	0	0	1	0	**13**
PD0277a	9	2	2	0	0	1	0	0	0	0	1	0	1	**16**
PD0287a	4	1	5	4	0	0	1	1	0	0	0	0	1	**17**
PD0307a	4	5	2	0	0	0	0	0	0	0	0	1	0	**12**
PD0312a	2	0	0	2	0	1	0	0	0	0	0	0	0	**5**
PD0319a	1	0	0	1	1	0	0	0	0	0	0	0	0	**3**
PD0331a	23	4	4	5	0	1	0	1	1	0	1	0	4	**44**
PD0351a	3	0	0	1	0	0	1	0	0	0	0	0	1	**6**
TOTAL	**65**	**12**	**16**	**17**	**1**	**4**	**2**	**2**	**1**	**1**	**2**	**2**	**11**	**136**

Supplementary Table 4. Number of RT-mediated rearrangements attributed to each rearrangement class across the 10 tumours in the cohort. Simple rearrangements are highlighted in blue, while more complex patterns resulting from DSB and involving several junctions are coloured in green. A red colour scale is used within each rearrangement class. DEL: Deletion; DSB: Double-strand break; DUP: Duplication; INV: Inversion; rINV: Reciprocal inversion; RT: Retrotransposition; RT/None: Only one of the two junctions within a reciprocal rearrangement features a RT bridge; RT/RT: Both junctions within a reciprocal rearrangement feature a RT bridge; 3ref TRA: Two interchromosomal junctions connect three different chromosomes at a given DSB; rTRA: Reciprocal translocation; TRA: Translocation.

13.3 ETHICS COMMITTEE APPROVAL

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

Secretaria Técnica
Comité Autonómico de Ética da Investigación de Galicia
Secretaria Xeral. Consellería de Sanidade
Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Tel: 881546425. Correo-e: ceic@sergas.es

galicia

DICTAMEN DEL COMITÉ DE ÉTICA DE LA INVESTIGACIÓN DE SANTIAGO-LUGO

Cristina Márquez Riveras, Secretaria del Comité de Ética de la Investigación de Santiago-Lugo,

CERTIFICA:

Que este Comité evaluó en su reunión del día 22/01/19 el estudio**:**

Título: The impact of L1-mediated somatic rearrangements in the origin and development of human cancer
Versión:
Promotor/a:José Manuel Castro Tubío
Investigador/a: José Manuel Castro Tubío
Código de Registro:2018/586

Y que este Comité, tomando en consideración la pertinencia del estudio, el conocimiento disponible, los requisitos legales aplicables y los Procedimientos Normalizados de Trabajo del Comité, emite un dictamen **FAVORABLE** para la realización del citado estudio.

Documento firmado digitalmente por:
Márquez Riveras, Cristina: 31/01/2019 10:39
SAOC-Z4G4-BOAH-OFHK-6M15-4892-7577-549

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

galicia

Y HACE CONSTAR QUE:

1.- El Comité Territorial de Ética de la Investigación de Santiago-Lugo cumple tanto en su composición como en sus PNTs los requisitos legales vigentes.

2.- La composición actual del Comité Territorial de Ética de la Investigación de Santiago-Lugo es:

- **Juan Manuel Vázquez Lago (Presidente).** Médico especialista en Medicina Preventiva y Salud Pública. Área de Gestión Integrada de Santiago**.**
- **Pilar Rodríguez Ledo (Vicepresidenta).** Médico especialista en Medicina Familiar y Comunitaria. Área de Gestión Integrada de Lugo.
- **Cristina Márquez Riveras (Secretaria).** Enfermera. Dirección Xeral de Saúde Pública.
- **Lorenzo Armenteros del Olmo (Secretario Suplente).** Médico especialista en Medicina Familiar y Comunitaria. Área de Gestión Integrada de Lugo.
- **Francisco Campos Pérez.** Biólogo. Fundación Instituto de Investigación Sanitaria de Santiago de Compostela.
- **Rosana Castelo Domínguez.** Farmacéutica de Atención Primaria. Área de Gestión Integrada de Santiago.
- **Ricardo García Martínez.** Licenciado en Derecho. Área de Gestión Integrada de Lugo.
- **Jaime Gulín Dávila**. Farmacéutico especialista en Farmacia Hospitalaria. Área de Gestión Integrada de Lugo.
- **María Jesús Lamas Díaz.** Farmacéutica especialista en Farmacia Hospitalaria. Área de Gestión Integrada de Santiago.
- **Guillermo José Prada Ramallal** Médico especialista en Farmacología Clínica. Área de Gestión Integrada de Santiago. Fundación Instituto de Investigación Sanitaria de Santiago de Compostela.
- **Carlos Rodríguez Moreno.** Médico especialista en Farmacología Clínica. Área de Gestión Integrada de Santiago.
- **Sandra Vidal Martínez.** Enfermera. Área de Gestión Integrada de Santiago
- **Rafael Carlos Vidal Pérez.** Médico especialista en Cardiología. Área de Gestión Integrada de Lugo.

Para que conste donde proceda, y a petición de quien proceda, en Santiago de Compostela,

La Secretaria del Comité Territorial de Ética de la Investigación de Santiago Lugo,

Cristina Márquez Riveras

Documento firmado digitalmente por:
Márquez Riveras, Cristina: 31/01/2019 10:39
SAOC-Z4G4-BOAH-OFHK-6M15-4892-7577-549

13.4 EXTENDED ABSTRACT

A retrotransposición somática, un proceso mutacional no que os elementos retrotranspoñibles son copiados e inseridos en novos sitios do xenoma dentro de células somáticas, está implicado na iniciación e progresión de certos tumores humanos. Nunha análise previa que englobou 2,954 xenomas de cancros pertencentes a 38 subtipos diferentes dentro do proxecto *Pan-Cancer Analysis of Whole Genomes* (PCAWG), detectamos altas taxas de retrotransposición somática en varios tipos de cancros. Estes incluíron o adenocarcinoma de esófago (ESAD), o carcinoma escamoso de cabeza e pescozo (HNSC), o carcinoma escamoso de pulmón (LUSC) e o adenocarcinoma colorectal (COAD). De maneira notábel, estes catro tipos de cancros representaron o 70% de todas as retrotransposicións somáticas, aínda que só constituían o 9% das mostras. Con todo, algúns aspectos relevantes deste mecanismo mutacional quedaron sen caracterizar debido ás limitacións tecnolóxicas da secuenciación de lectura curta de Illumina.

Este proxecto de doutoramento foi concibido para explorar o potencial das tecnoloxías de secuenciación de lectura longa e desenvolver ferramentas computacionais específicas para superar as limitacións previas, mellorando a nosa comprensión dos patróns mutacionais derivados da retrotransposición somática. Como resultado, desenvolvemos MEIGA (*Mobile Element Integrations Genome Analyzer)*, un novo método computacional especificamente deseñado para detectar eventos de retrotransposición somática en datos de secuenciación de lectura longa no contexto do cancro. Aínda que existen ferramentas como xTea, rMETL e PALMER para detectar ditos eventos en lecturas longas, estas están principalmente deseñadas para identificar inserción canónicas, e presentan limitada capacidade para detectar eventos non canónicos. Ademais, ningunha destas ferramentas está adaptada para identificar retrotransposicións somáticas en análises emparelladas de tumores e tecidos normais, un aspecto crítico na investigación do cancro. MEIGA foi deseñado para abordar estas limitacións, proporcionando unha análise máis exhaustiva do impacto destes elementos no xenoma de tumores.

Avaliamos a precisión e o *recall* de MEIGA e outras ferramentas dispoñibles a través da análise de datos simulados. Simulamos un xenoma ficticio con 6,420 inserción de retrotransposóns empregando lecturas longas de Oxford Nanopore Technologies (ONT) a 30x e un VAF do 50%. En comparación con xTea, rMETL e PALMER, MEIGA destacou cun *recall* do 96,04% para inserción de L1, sobresaíndo especialmente na identificación de transducións. Para transducións orfas, MEIGA acadou unha valor de *recall* do 97.6%, mentres que rMETL, xTea, e PALMER obtiveron valores por debaixo do 8%. Con respecto ás transducións emparelladas de L1, MEIGA e xTea acadaron un *recall* do 92.28% e 95.26%, mentres que rMETL e PALMER obtiveron valores do 78.77% e 57.89%, respectivamente. Para os elementos *Alu*s e SVA, os valores de *recall* non mostraron diferenzas relevantes entre as metodoloxías (intervalo=[95,9, 99,9]). Sen embargo, as taxas de precisión foron lixeiramente máis baixas para *Alu*s nas ferramentas alternativas (rMETL: 92.59%, xTea: 93.66%, PALMER: 95.23%, MEIGA: 99.8%).

Ademais, MEIGA mostrou unha maior sensibilidade en niveis decrecentes de VAF, indicando unha detección superior de eventos subclonais. Así mesmo, MEIGA destacouse por realizar unha reconstrución precisa das secuencias das inserción simuladas, permitíndolle describir en detalle as características estruturais dos eventos de retrotransposición detectados. No que se refire ao rendemento en termos de tempo de execución e uso de memoria, a pesar de que MEIGA ocupou o segundo lugar por detrás de rMETL, MEIGA demostrou eficiencia suficiente

para ser aplicada sobre grandes volumes de datos xenómicos. Nas condicións simuladas a 30x, o seu tempo de execución foi de 1.56 horas, con un uso de memoria de 26.2 GB. Xa que non hai outras ferramentas de referencia para a detección de reordenamentos mediados por retrotransposóns, só avaliamos o rendemento de MEIGA na detección de tales eventos. A avaliación de MEIGA resultou en ningunha detección errónea e alcanzou as seguintes niveles de *recall*: 20/20 para delecións, 18/20 para duplicacións, 17/20 para inversións e 9/10 para translocacións. Polo tanto, MEIGA emerxe como unha ferramenta robusta e fiábel para a identificación e precisa caracterización non só de inserción, senón tamén de reordenamentos mediados por retrotransposóns, destacando significativamente sobre outros métodos existentes.

Para identificar tumores con elevadas taxas de retrotransposición somática, levouse a cabo un cribado prospectivo nunha cohorte de 150 pacientes con cancros primarios de cabeza e pescozo, pulmón e colorectal (50 HNSC, 50 LUSC e 50 COAD). A análise inicial implicou a secuenciación do xenoma completo destes tumores usando lecturas curtas a baixa cobertura para identificar un subconxunto de 10 tumores con máis de 100 retrotransposicións somáticas. Estes tumores incluíron cinco HNSC, catro LUSC e un COAD. Posteriormente, realizouse a secuenciación de xenoma completo empregando a tecnoloxía de secuenciación de lecturas longas de ONT nestes 10 tumores con altas taxas de retrotransposición, así como nos seus tecidos adxacentes non tumorais. Analizamos con MEIGA os datos resultantes, identificando un total de 6,266 inserción somáticas de retrotransposóns. As inserción solitarias de L1 foron o tipo máis común (56.3%), seguidas por transducións de L1 (34.9%), tractos de poli(A/T) (2.4%), pseudoxenes (2.1%), *Alu*s (1.9%) e SVAs (<0.1%). Destaca a identificación de 1,311 retrotransposicións somáticas nun tumor de HNSC, PD0270a, representando o 20% de todos os eventos na cohorte.

Empregáronse dúas ferramentas adicionais para analizar estes 10 tumores co obxectivo de avaliar os resultados de MEIGA en datos reais, o módulo de xTea para datos de lectura curta e o módulo de xTea para datos de lectura longa. Ao comparar os resultados, observouse que a maioría (76.5%) dos eventos detectados por MEIGA foron confirmados de maneira consistente por polo menos un algoritmo adicional, indicando que representan eventos de retrotransposición xenuína. No caso das retrotransposicións específicas de MEIGA, o 97.3% foron confirmados como verdadeiros mediante inspección visual no Integrative Genomics Viewer. Así, cos resultados obtidos por MEIGA nesta cohorte, estimouse unha sensibilidade do 96.6%, e unha taxa de falsos positivos do 0.6%. En comparación cos outros dous métodos, MEIGA identificou un 22.2% de retrotransposicións somáticas xenuínas que os outros non detectaron. Polo tanto, MEIGA demostrou ser unha ferramenta sensible e sólida non só con datos simulados, senón tamén sobre datos reais.

MEIGA permitiu reconstruír con precisión as secuencias dos 6,266 eventos de retrotransposición somática detectados, facilitando a análise das inserción cunha resolución sen precedentes. A lonxitude media das inserción variou segundo o tipo de retrotransposón analizado. A análise da estrutura das inserción mostrou que case a metade das inserción solitarias de L1 contiñan unha inversión interna, e que estas eran máis longas en comparación coas non invertidas. Observáronse tamén patróns complexos que implicaban saltos da transcriptasa inversa non só ao longo da secuencia dunha molécula de ARN mensaxeiro, senón tamén entre varias moléculas precursoras. A análise da estrutura das inserción de pseudoxenes procesados revelou que, aínda que a maioría destas inserción consisten en exóns únicos, normalmente o 3'-UTR, poden involucrar múltiples exóns, sendo 15 exóns o máximo observado neste estudo.

Ademais, MEIGA permitiu a reconstrución das secuencias de elementos L1 potencialmente activos —os chamados elementos fonte— no xenoma dos dez pacientes en estudo. Este proceso abriu a porta a unha nova metodoloxía focalizada na identificación de variantes nucleótidicas específicas de cada elemento fonte —os chamados nucleótidos diagnóstico— dentro das secuencias retrotranspostas. Durante a transcrición e a retrotranscrición, as variantes xenéticas son transferidas do elemento fonte ás súas copias derivadas, o que nos dá a posibilidade de asociar cada inserción somática co seu elemento de orixe.

Investigacións previas utilizaron as transducións de L1, que conteñen secuencias xenómicas únicas adxacentes ao seu L1 de orixe, para identificar de forma inequívoca o elemento fonte. Sen embargo, estes estudos non abarcaron a orixe das insercións solitarias de L1 que constitúen a maior parte das retrotransposicións somáticas, polo que posiblemente se subestimou o número e a actividade dos elementos fonte. Para abordar esta cuestión, empregamos MEIGA para reconstruír con precisión as secuencias dos elementos L1 potencialmente activos nos tecidos sans adxacentes, así como as secuencias retrotranspostas nos tumores, e aplicamos a estratexia de inferencia de elementos fonte baseada en nucleótidos diagnóstico deseñada por Martin Santamarina. Deste modo, conseguimos atribuír o 25.1% das insercións solitarias de L1 detectadas na cohorte a elementos fonte concretos. Esta análise transformou significativamente a nosa comprensión do espectro de actividade dos elementos fonte de L1. Desvelou patróns de actividade diferenciados, con algúns loci propagándose principalmente a través de transducións somáticas, mentres que outros tendían a xerar insercións solitarias. Ademais, o estudo dos sinais de poliadenilación (PA) vinculados a cada locus de L1 revelou que a potencia destes sinais é crucial na determinación da frecuencia de insercións solitarias e de transducións para cada elemento.

MEIGA non só detectou insercións, senón que tamén identificou 152 reordenamentos cromosómicos nos 10 tumores examinados, todos eles resultantes da actividade dos retrotransposóns. Os reordenamentos causados por retrotransposóns caracterízanse pola unión de dous puntos de ruptura distantes no xenoma a través de pontes formadas por retrotransposóns. Estes reordenamentos foron divididos en catro categorías dependendo de se eran eventos intercromosómicos ou intracromosómicos e da orientación dos puntos de ruptura implicados. As delecións resultaron ser as máis frecuentes, constituíndo o 43% (n=66) do conxunto, seguidas das translocacións co 30% (n=45), inversións co 18% (n=27), e duplicacións co 9% (n=14). Estas cifras varían considerablemente das reportadas no estudo de PCAWG, que só revelou 90 delecións, unha duplicación, unha translocación e unha inversión nos 2,954 tumores analizados. Isto indica que, con excepción das delecións, é probable que as demais categorías estivesen infraestimadas nos estudos previos que empregaban secuenciación de lecturas curtas.

Os reordenamentos mediados por retrotransposóns están presentes en todas as mostras de tumores, pero amosaron unha considerable variabilidade entre pacientes en termos de número e tipos (mediana=13.5; rango [3-49]). Nun tumor LUSC salientable, PD0331a, identificamos ata 49 reordenamentos mediados por retrotransposóns, constituíndo o 32% (49 de 152) dos eventos totais dentro da cohorte. Para analizar os reordenamentos dentro dun contexto xenómico máis amplo, clasificámolos baseándonos na proximidade dos seus puntos de ruptura. Así, observamos que o 13.2% (20/152) dos reordenamentos mediados por retrotransposóns formaban parte de 13 translocacións recíprocas. Observamos este patrón en cinco dos dez tumores analizados, subliñando a súa prevalencia e a necesidade dunha análise máis detallada.

Por exemplo, nun caso destacable, o tumor PD0270a, descubrimos un total de catro translocacións recíprocas, xunto con dúas translocacións non recíprocas, todas mediadas por retrotransposóns.

Exploramos en detalle o mecanismo de formación das translocacións recíprocas mediadas por retrotransposóns. As translocacións recíprocas caracterízanse por un intercambio equilibrado de material xenético entre dous cromosomas non homólogos, o que resulta na formación de dous cromosomas derivados. Nos casos que identificamos, estes cromosomas derivados están conectados por pontes de retrotransposóns. O exame pormenorizado das pontes de ambos cromosomas derivados revelou que en 10 de 13 translocacións recíprocas, dous eventos de retrotransposición independentes estaban implicados. Isto foi evidenciado por un deseño singular coas seguintes características. Primeiro, atopáronse distintas secuencias retrotranspostas en cada cromosoma derivado, como evidenciaron as secuencias transducidas que se orixinan de diferentes elementos fonte. Segundo, a presenza dun motivo da endonucleasa xunto cunha cola de poli(A/T) en cada cromosoma confirmou a utilización independente do mecanismo de integración canónico (TPRT, polas súas siglas en inglés: *Target-Primed Reverse Transcription*). Terceiro, a análise das duplicacións (TSDs, polas súas siglas en inglés: *Target Site Duplications*) resultantes dos dous eventos de retrotransposición independentes revelou que están entrelazadas de tal maneira que cada ponte individual está flanqueada por unha heteroduplicación. É dicir, unha das dúas copias de cada TSD están localizadas nun cromosoma derivado diferente.

Ademais, identificamos tres casos de translocacións recíprocas onde parece que un único evento de retrotransposición é empregado para reparar simultaneamente ambos extremos dunha ruptura de dobre cadea (DSB, polas súas siglas en inglés: *Double-strand break*) situada nun cromosoma non homólogo, tal como se evidencia polo estudo das secuencias retrotranspostas. Polo tanto, os nosos achados suxiren que as translocacións recíprocas mediadas por retrotransposóns poden resultar non só da unión de dous eventos independentes de retrotransposición, senón tamén dun único evento de retrotransposición. Máis aló das translocacións recíprocas, observamos que os retrotransposóns tamén poden mediar outras formas de reordenamentos recíprocos, como inversións (n=6) e outros eventos máis complexos tamén resultantes de DSBs. En canto ás inversións recíprocas e os eventos máis complexos, tamén observamos que poden ser mediados por dous eventos de retrotransposición independentes, así como por un único evento.

O xenoma dunha célula cancerosa é un rexistro valioso das mutacións somáticas que se foron acumulando ao longo do tempo dentro da súa liñaxe clonal. Estas mutacións teñen a súa orixe nun único cromosoma dentro dunha célula precursora que da lugar a unha descendencia celular que comparte a mutación inicial. Cando se produce a duplicación dunha rexión cromosómica, todas as mutacións xa existentes no alelo que será duplicado tamén se replicarán. Non obstante, as mutacións que ocorran despois desta duplicación, ou aquelas que estean no alelo oposto, non seguirán ese mesmo patrón. Utilizando datos de secuenciación, podemos determinar con precisión a frecuencia alélica de cada mutación específica, permitíndonos clasificar as mutacións como variantes clonais temperás ou tardías. As variantes temperás preceden ós aumentos do número de copias, mentres que as variantes tardías suceden despois destes. Ademais, hai un subconxunto de mutacións referidas como variantes clonais non asignadas, que están presentes en todas as células cancerosas pero non poden ser clasificadas máis detalladamente como variantes temperás ou tardías. Por último, as mutacións subclonais son

aquelas que xurdiron en etapas posteriores no desenvolvemento do tumor e só están presentes nun subconxunto de células tumorais.

Para profundar na análise da retrotransposición somática nas distintas fases da evolución do cancro, modificamos as técnicas actuais de datación de mutacións somáticas para o estudo dos eventos de retrotransposición. Dado que os métodos de datación existentes están adaptados para datos de secuenciación de lectura curta e baséanse na determinación precisa da frecuencia alélica da variante (VAF, das súas siglas en inglés: *Variant Allele Frequency*), aproveitamos o código desenvolvido para MEIGA e fixemos unha implementación para datos de secuenciación de lectura curta, MEIGA-SR. Neste método, introducimos unha funcionalidade de xenotipado deseñada especificamente para calcular a frecuencia alélica de eventos de retrotransposición en tumores. A nosa validación con datos simulados confirmou que MEIGA-SR calcula con exactitude a VAF destes eventos. As estimacións de VAF centráronse ao redor do valor esperado e a dispersión observada axustouse a unha distribución normal.

Ademais, empregamos as VAFs estimadas con MEIGA-SR para elaborar predicións de datación relativa de inserción de retrotransposóns en datos reais, e comprobamos a coherencia destas predicións entre seis áreas distintas dun mesmo tumor. Esta avaliación demostrou unha eficacia do 97.94% na correcta asignación da etiqueta temporal máis verosímil a unha área específica, confirmado a solidez do noso enfoque. Posteriormente, aplicamos esta metodoloxía para datar os eventos de retrotransposición somática identificados con lecturas longas no conxunto de 10 tumores con altas taxas de retrotransposición. A nosa análise revelou que 2,975 inserción de retrotransposóns (64.1% de 4,644) ocorreran nas etapas iniciais do desenvolvemento tumoral, mentres que 692 (14.9%) foron eventos clonais tardíos, 674 (14.5%) caeron na categoría de clonais non asignados, e 303 (6.5%) foron designados como eventos subclonais que ocorreron nas etapas posteriores da tumorixénese. Isto suxire que a retrotransposición somática non é unha consecuencia do caótico ambiente xenómico característico das etapas tardías, senón un proceso mutacional que pode ser moi activo durante as etapas temperás do desenvolvemento tumoral. Ademais, a pesar da maior dificultade para detectar eventos tardíos e subclonais, tamén parece que a retrotransposición somática permanece como un mecanismo mutacional activo ao longo da evolución do tumores.

A duplicación completa do xenoma (WGD, polas súas siglas en inglés: *Whole Genome Doubling*) é un evento xenético no que unha célula adquire un conxunto adicional de cromosomas, resultando así na duplicación de todo o seu xenoma. Este fenómeno é común en células cancerosas e xorde debido a erros na división celular. O momento exacto dos eventos de WGD, cuantificado en anos, pode ser deducido ao avaliar o número de copias das mutacións 'tipo reloxo', como as transicións de CpG a TpG. Denomínanse mutacións tipo reloxo a aquelas que se acumulan a unha taxa relativamente constante e predecible ao longo do tempo. A relación entre mutacións de copia única e duplicadas permite obter estimacións de tempo real do evento de WGD, e clasificar outras mutacións en relación a este. Así, a nosa análise determinou que os eventos de WGD ocorreran nun tempo promedio de 4.77 anos antes da biopsia na cohorte estudada, cun rango que ía dende 1.77 ata 8.87 anos. É notable que todos os tumores mostraron máis de 100 inserción somáticas de retrotransposóns ocorrendo antes do evento de WGD. Nun tumor relevante, PD0270a, identificamos ata 836 inserción de retrotransposóns ocorrendo antes do WGD, remontándose polo menos a 1.77 anos antes da biopsia, e 175 despois do WGD. Por outra banda, o tumor PD0287a exhibiu 121 inserción de retrotransposóns que ocorreran polo menos 8.87 anos antes da biopsia e 166 despois dese punto temporal. Polo tanto, os nosos

achados demostran que a retrotransposición somática constitúe un proceso mutacional activo que pode ocorrer anos antes do diagnóstico.

A continuación, procedemos a examinar a contribución das diferentes familias de retrotransposóns ao longo do desenvolvemento tumoral. Esta análise revelou unha tendencia á mobilización de elementos *Alu* durante as etapas tardías da tumorixénese. Un patrón similar observouse para a categoría de pseudoxenes procesados, aínda que sen significación estatística. Estes achados conxuntamente suxiren que nas etapas iniciais, a maquinaria de retrotransposición L1 mostra unha forte preferencia por inserción cis que involucran o seu propio ARN. Con todo, esta preferencia cambia durante o desenvolvemento do tumor, tornándose máis permisiva e facilitando a trans-mobilización de elementos *Alu* e outros transcritos derivados de xenes nucleares. Para obter información sobre como as presións evolutivas modelan a paisaxe da retrotransposición, investigamos posibles correlacións entre os eventos de retrotransposición e varias características xenómicas en distintas etapas do tumor. A nosa análise non atopou diferenzas significativas na distribución das inserción somáticas de retrotransposóns dentro das rexións xénicas entre as etapas temperás e tardías. Tampouco observamos diferenzas na orientación das inserción en rexións xénicas. Con todo, cando examinamos exclusivamente a frecuencia das inserción somáticas de retrotransposóns en xenes transcripcionalmente activos, observamos un sinal APARENTe de selección negativa contra as inserción de retrotransposóns en xenes activos.

A datación dos reordenamentos mediados por retrotransposóns revelou que ata 75 reordenamentos correspondían a eventos temperáns, mentres que 28 foron identificados como eventos tardíos, dun total de 103 casos nos que se logrou unha datación exitosa. Curiosamente, observamos que reordenamentos cromosómicos de gran envergadura, como nove translocacións recíprocas, ocorreran de maneira temperá na tumorixénese. Por exemplo, no tumor LUSC PD0331a, atopamos ata 26 reordenamentos ocorrendo cedo no desenvolvemento do tumor. Aínda que o impacto segue sendo incerto para moitos destes eventos, teñen o potencial de perturbar significativamente os dominios funcionais dentro do xenoma. Entre eles, unha translocación recíproca afectou ao xene supresor tumoral *EP300*. A estimación en tempo real do evento de WGD para este paciente indicou que todos estes reordenamentos ocorreran polo menos 3.4 anos antes da biopsia do tumor (intervalo de confianza=[2.5, 5.8]). No que se refire aos eventos nas etapas tardías, tamén observamos eventos tardíos con potencial impacto no cancro. Por exemplo, no tumor PD0274a, datamos unha inversión recíproca que afecta ao supresor tumoral *LRP1B* como un evento tardío, que presumiblemente ofrece unha vantaxe selectiva máis tarde na tumorixénese. Ademais, observamos que eventos con potencial de exercer un gran impacto, como as translocacións recíprocas, tamén poden xurdir tardiamente no desenvolvemento do tumor. Estes achados apoian a noción de que a retrotransposición somática, a través de eventos de integración aberrante que resultan en reordenamentos xenómicos, desempeñan un papel activo na inestabilidade do tumor e na súa progresión; e que isto ocorre dende as etapas temperás da evolución tumoral, podendo ter lugar incluso anos antes do diagnóstico.

Utilizando tanto a orixe de transducións como as inserción solitarias de L1, realizamos unha investigación detallada sobre o número de elementos L1 activos en cada tumor nas etapas temperás e tardías da tumorixénese. Primeiramente, a nosa análise descubriu que as taxas individuais de actividade dos elementos fonte fluctúan ao longo do desenvolvemento do tumor, con períodos de activación e desactivación, aínda que o número total de elementos fonte activos mantense constante. Estes resultados están en liña co modelo que propón unha diversidade

constante de elementos activos ao longo da tumorixénese, parecido á dinámica observada nos procesos da liña xerminal. Adicionalmente, observamos que a contribución relativa dos elementos fonte L1 non só varía co tempo, senón tamén entre diferentes tumores, o que indica un escenario dinámico de regulación e desregulación destes potenciais axentes mutaxénicos nos xenomas tumorais.

Aínda que a retrotransposición somática xoga un papel significativo na tumorixénese, dispoñemos de coñecementos limitados sobre a súa relevancia e alcance nos estadios previos ao proceso tumoral, nos tecidos normais. A identificación de mutacións somáticas en tecidos normais que non experimentaron expansión clonal, representou un gran desafío; con todo, os avances tecnolóxicos recentes comezaron a posibilitar a súa detección. Un destes avances é a técnica de microdisección láser (LCM, das súas siglas en inglés: *Laser-Capture Microdissection*), que facilita o illamento de pequenas seccións de tecido, tipicamente incluíndo só uns poucos centos de células de interese, as cales logo poden ser sometidas a técnicas de secuenciación. Como resultado, agora é posible caracterizar a paisaxe mutacional das células illadas cunha resolución sen precedentes.

Aínda que hai numerosos métodos dispoñibles para estudar a retrotransposición somática en cancro empregando lecturas curtas, nós centrámonos en crear un algoritmo especificamente deseñado para a detección destas mutacións en tecidos sans estudados con LCM. Con este fin, o noso obxectivo era deseñar un método altamente sensible capaz de identificar calquera sinal de retrotransposición somática. Dado que os tumores frecuentemente amosan taxas de mutación elevadas, resulta imprescindible a implementación de filtros estritos para minimizar o ruído de fondo, unha medida que non se require nos tecidos sans. Na súa avaliación, MEIGA-SR superou as outras dúas ferramentas probadas, xTea e TraFiC, especialmente en canto ao recall, onde MEIGA-SR acadou o 95.7%, xTea 83.5% e TraFiC 65%. A precisión mantívose comparable entre os tres métodos, variando do 99.4% ao 99.9%. Aínda que se necesitan esforzos adicionais para mellorar a robustez de MEIGA-SR para aplicación máis xeneralizada, MEIGA-SR é un método sensible e versátil para detectar inserción de retrotransposóns.

Para examinar a retrotransposición somática en tecidos normais, empregamos MEIGA-SR para analizar datos de secuenciación de xenoma completo de Illumina obtidos de microbiopsias LCM procedentes de individuos sans. No total, o noso estudo consistiu na análise de 504 microbiopsias. Estas mostras abarcaban oito tipos de tecido, incluíndo placenta (n=117), estómago (n=145), apéndice (n=34), colon (n=59), páncreas (n=19), pel (n=12), intestino delgado (n=59) e testículo (n=59). A nosa abordaxe identificou un total de 220 inserción de retrotransposóns adquiridas somaticamente. Os elementos L1-solo representaron o tipo máis común de inserción, constituíndo o 62.3% do total dos eventos de retrotransposición.

A comparación das taxas de retrotransposición entre os distintos tipos de tecidos revelou que a placenta foi o tecido con maior proporción de biopsias afectadas por un ou máis eventos de retrotransposición somática (54.7%), destacando que a retrotransposición estivo activa en máis da metade das biopsias nalgún momento durante o desenvolvemento placentario. Notablemente, o colon seguiu con unha taxa do 25.4%, xunto co estómago co 21.4%, intestino delgado co 20.3% e apéndice co 14.7%. De forma destacada, o estómago mostrou o maior número de inserción de retrotransposóns somáticos por biopsia, revelando ata 20 inserción somáticas distintas nunha única microbiopsia. En xeral, estes achados indican que a retrotransposición é un mecanismo mutacional activo en varios tecidos sans. Non obstante, a capacidade de detectar mutacións somáticas está directamente vinculada á profundidade da

secuenciación e á clonalidade do tecido estudado. Para avaliar se o poder de detección foi consistente entre os tecidos, empregamos unha función de distribución acumulativa binomial. Os nosos resultados mostraron que o poder de detección semella consistente en todos os tecidos agás para a pel e o páncreas, onde foi notablemente máis baixo.

En consecuencia, os nosos resultados suxiren que as diferenzas observadas nas taxas de actividade para a placenta, apéndice, colon, estómago e testículos son xenuínas, pero isto non é certo para as biopsias de pel e páncreas. Así, en liña cos estudos previos que demostran unha extensa mutaxénese nos tecidos placentarios, as nosas observacións indicaron que máis da metade das microbiopsias placentarias foron afectadas pola retrotransposición somática, converténdoa no tipo de tecido máis afectado. Isto suxire un escenario onde quizais non exista unha presión selectiva significativa para preservar o material xenético deste órgano de rápido crecemento, curta vida e descartable. En contraste, non houbo actividade APARENTe de retrotransposón en testículo, apoiando a baixa carga mutacional xa informada neste tecido. En xeral, os nosos resultados suxiren que a retrotransposición somática está activa anos antes do inicio de enfermidades como o cancro, presumiblemente contribuíndo á xeración de variabilidade xenómica que finalmente pode contribuír á progresión do tumor.

En conxunto, esta tese ofrece novas perspectivas sobre a dinámica da retrotransposición somática. As nosas investigacións demostran que é un proceso mutaxénico activo anos antes do diagnóstico do tumor, mesmo en tecidos non tumorais, e capaz de producir reordenamentos xenómicos de gran impacto. Estes resultados demostran que a retrotransposición somática pode xogar un papel relevante no inicio e desenvolvemento de certos tumores. Este estudo, polo tanto, amplía a nosa comprensión da retrotransposición somática no contexto do cancro, proporcionando perspectivas valiosas para futuras investigacións e o desenvolvemento de novas estratexias terapéuticas.

Somatic retrotransposition can significantly contribute to tumorigenesis. Yet, limitations imposed by short-read sequencing have hindered a thorough understanding of the extent and impact of this mutational process. To overcome this, my PhD project focused on implementing innovative long-read sequencing techniques. We developed MEIGA, a bioinformatic pipeline for comprehensively analysing somatic retrotransposition in long-read sequencing data. Notably, MEIGA revealed previously undisclosed levels of somatic retrotransposition in human tumours, including a hidden landscape of reciprocal genomic rearrangements. Additionally, this work unveiled remarkable rates of somatic retrotransposition occurring years before tumour diagnosis, even within normal cells, providing valuable insights into somatic retrotransposition dynamics.