INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Bernardo
Rodríguez Martín

PhD Thesis

The impact of transposable elements on the structure and function of the cancer genome

Santiago de Compostela, 2022

Doctoral Programme in Molecular Medicine

TESE DE DOUTORAMENTO

THE IMPACT OF TRANSPOSABLE ELEMENTS ON THE STRUCTURE AND FUNCTION OF THE CANCER GENOME

Bernardo Rodríguez Martín

ESCOLA DE DOUTORAMENTO INTERNACIONAL DA UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOUTORAMENTO EN MEDICINA MOLECULAR

SANTIAGO DE COMPOSTELA / LUGO

2022

DECLARACIÓN DO AUTOR/A DA TESE

D./Dna. **Bernardo Rodríguez Martín**

Título da tese: **The impact of transposable elements on the structure and function of the cancer genome**

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1) A tese abarca os resultados da elaboración do meu traballo.
2) De ser o caso, na tese faise referencia ás colaboracións que tivo este traballo.
3) Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.
4) A tese é a versión definitiva presentada para a súa defensa e coincide a versión impresa coa presentada en formato electrónico

E comprométome a presentar o Compromiso Documental de Supervisión no caso de que o orixinal non estea na Escola.

En **Heidelberg**, **02 de Decembro de 2021**.

AUTORIZACIÓN DO DIRECTOR / TITOR DA TESE

The impact of transposable elements on the structure and function of the cancer genome

D./Dna. José Manuel Castro Tubío

INFORMA/N:

Que a presente tese, correspóndese co traballo realizado por D/Dna. Bernardo Rodríguez Martín, baixo a miña dirección/titorización, e autorizo a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como director desta non incorre nas causas de abstención establecidas na Lei 40/2015.

De acordo co indicado no Regulamento de Estudos de Doutoramento, declara tamén que a presente tese de doutoramento é idónea para ser defendida en base á modalidade de Monográfica con reproducción de publicaciones, nos que a participación do/a doutorando/a foi decisiva para a súa elaboración e as publicacións se axustan ao Plan de Investigación.

En Santiago de Compostela, 2 de Diciembre de 2021

Esta tesis fue financiada por la Xunta de Galicia, en el marco de su programa de contratos predoctorales. Convocatoria 2015. Referencia: ED481A-2016/151

AKNOWLEDGMENTS

Human beings, as mobile elements, jump through multiple stages of our own lives. In 2016, I was lucky enough to jump into José Tubio laboratory to start a PhD under his supervision on cancer genomics and somatic retrotransposition. Like many other young scientists, I recall the beginning of my PhD as a cocktail of excitement, enthusiasm, but also uncertainty regarding the success and impact of my research. In 2021, while I am writing this PhD thesis and looking back to these 5 years, I can only feel satisfied about what I have learnt and achieved, happy about how much I have enjoyed this journey and grateful with the excellent researchers that have contributed to this PhD dissertation.

I am greatly indebted to José Tubio for introducing me to the fascinating world of mobile DNA, sharing his passion about scientific research and guiding me along these years. I am totally aware that the level of success of a PhD student is highly dependent on his advisor, so thanks for your great mentorship. I will always remember these 4 years working at “Genomes & Disease Lab” as an exciting and broadly enjoyable period of my life. This was only possible thanks to the amazing team of scientists I had the opportunity to work with. Thanks to Alicia, Martin and Eva, my very first partners in crime at the lab. Besides their important scientific contributions, which are described along this thesis, my PhD memories are full of amazing moments with you folks. These range from surfing evenings after an exhausting working day to dinners, parties and hikes, including camping on the seaside at Costa Da Morte. A warm thanks to Paula. It was a big pleasure to supervise your master thesis and I wish you the same level of success as your talent. Thanks to crazy Jorge for your major contributions from the computational angle. I will always remember your visits to Santiago as fun. Thanks to Javi for his major contributions managing sequencing data, setting up nanopore sequencers and, more importantly, for your disinterested intellectual input in the development of MEIGA. Thanks to Sonia, with whom, together with Eva, Martin and Javi; I joined forces to investigate somatic retrotransposition in cancer genomes under the lens of emerging long-read sequencing technologies. Thanks to Jeni, Jorge R, Pili, Seila, Mónica and Dani for citing some of the great lab colleagues I had the pleasure to work with.

I am extremely grateful to David Posada and his team at University of Vigo. It was really inspiring to work alongside you and to know about your exciting projects to model cancer evolution using single-cell sequencing technologies. As a consequence, I really missed our daily interaction when I moved to Santiago de Compostela. Special mentions for Sonia and her sharp sense of humor, Nuria and her reggaeton music, Harry and his relaxed flow, and Tamara and her curious spirit. A massive thanks to Iñigo Martincorena for hosting me at his group during 3 months at the Sanger Institute. It was great to experience *in situ* the vibes and scientific culture of one of the world-wide epicenters of genome research. It was really inspiring and fun to meet with a horde of talented scientists, including Tim, Alex, Fede and Francesco, among others. I am also greatly indebted to Jan Korbel for initially hosting me for

two months and later giving me the opportunity to continue my research career at his laboratory at EMBL, an unique European organization, spanning multiple countries and pursuing top notch research in life sciences. It is truly inspiring to work with you and also rewarding to feel the support to pursue some of my ideas at your lab. Massive thanks to the great colleagues and friends from Korbel group, which made me feel at home from moment one. Thanks to Esa, Shimin, Hyobin, Patrick, Tania, Wolfy, Lucca, Maja, Karen, “Jefecito’’ Ahmad, Ashley, Nina, Marco and all Korbelers for thrilling scientific discussions, beer sessions, bouldering days, bike trips, parties and great moments in general. I would also like to express my gratitude to all the brilliant scientists I collaborated with during my thesis. In particular to Adrian Baez, Jonas Demeulemeester, Peter Van Loo, Fran Supek, David Torrents, Young Seok, Yilong Li, Peter J. Campbell and all the colleagues from the PCAWG Consortium. Thanks also to David Torrents and his team at the Barcelona Supercomputing Center (BSC). Although unfortunately did not work out, I cannot forget everything I learned and enjoyed working with you.

I would also like to thank all the funding agencies that supported my PhD research. These include the foundation “la Caixa”, who supported me during one year at Barcelona. The European Molecular Biology Organization, which funded my visit to Jan Korbel laboratory at EMBL through their short-term fellowship program. The Galician government, my primary founder during my PhD, which also provided me additional support to visit Martincorena’s laboratory at the Sanger Institute. Now I will turn to Spanish, my native language, to thank friends and family.

Investigar es un arte. Eso implica que un investigador, al igual que un artista, necesita pasar sus primeras etapas de formación en movimiento, exponerse a distintas corrientes de pensamiento, métodos de trabajo, tecnologías y aprender, a poder ser, de los mejores en su campo. Esa vida de investigador errante es dura, pero por otro lado tremendamente gratificante. Estos años me han brindado la posibilidad de conocer a personas increíbles, de diversas culturas, que han dejado su impronta en mi forma de pensar y ver la vida. Gracias a la chavalada de Villaviciosa: Mario, Pablo, Richi, Santi, Pedro, Borja y compañía. A la mayoría de vosotros os conozco desde que éramos unos guajes. A día de hoy como diría Amaral: “Cómo hemos cambiado”, pero sigue siendo increíble poder pasar tiempo con vosotros cuando voy de vacaciones, recordar anécdotas y crear nuevas. Gracias a mis “Otros” amigos de Villaviciosa: Astor, Txiki, Beto y Corta; entre otros. Juntos hemos pasado momentos memorables en la floristería de Txiki, siempre punto de encuentro, y centro neurálgico del frikismo en Villaviciosa. Os debo mi afición por la ciencia ficción, los juegos de mesa, el modelismo, la montaña y muchas otras cosas. Especiales agradecimientos para Txiki, gran artista y mejor persona, por sus fabulosas ilustraciones, las cuales han hecho esta tesis, si cabe, aún más especial para mi.

Durante este tiempo también he sido piragüista errante. Muchas gracias a todos los miembros de los clubs de piragüismo de Villaviciosa, Vigo, Pontecesures, Cambridge y Heidelberg. Siempre me he sentido uno más y he disfrutado cada uno de los entrenamientos con vosotros.

En especial, agradecer a Rodrigo y Lucía, quienes me acogieron con los brazos abiertos en Cambridge, y tuve el placer de apoyar en la Devices to Westminster, una de las regatas más largas del mundo. También a Mario y Aitor, compañeros de K2 y alguna que otra fiesta. A Marcos, con quien pasé muchas horas remando en el Eume. Desde el punto de vista lúdico/festivo agradecer a la gente de “Bancos Bio”: Moreno, Alvarín, Ceci, Paula, Borja y compañia. A Gerard y Adriá, “els meus amics catalans”. Adrià, por muchas más caminatas y montañas que descubrir juntos. A “Losters” Barcelona: Jose, Inés, Tere, Ana y compañía. A los “caca friends”: Reza, Sebastian, Hana, Anna, entre otros. A Toni y Elena, mis amigos españoles en Heidelberg. A la “climbing crew”: Josh, Karin, Terra, Anna, Fidel, Dewi, etc… es increíble descubrir y disfrutar este deporte con vosotros. Finalmente, a Marta, fuiste alguien muy importante para mí y siempre te desearé lo mejor en la vida.

A mi familia, ese reducido núcleo de personas cuya mera compañía despierta la felicidad más absoluta. El mayor desafío al que me he enfrentado desde que decidí iniciar mi camino en el mundo de la investigación ha sido no poder disfrutar de vosotros día a día. Aun así, ya sea desde Madrid, Barcelona, Vigo, Santiago, Cambridge o Heidelberg, siempre os he sentido muy cerca. Agradecer a mi abuela Susa. Para ti siempre he sido el mejor en lo que hago y punto. Probablemente consecuencia del “sesgo abuela”, pero tu fé en mí siempre ha sido un motor y una fuente de inspiración muy grande. A mis padres, Menchu y Javier, a quienes se lo debo todo. Muchas gracias por creer siempre en mí y ser una fuente eterna de cariño y apoyo. A Ana. Mucho más que una madrina, una segunda madre y un faro que ilumina mi día a día con su constante alegría y ganas de vivir. A Paco y Cova, mis abuelos del Puntal, con quienes pasé momentos increíbles en su casa a los pies de la ría de Villaviciosa. A mis tías Emma y Aurorina, estoy deseando veros en unas semanas en Asturias. A mis primas y primos, un beso muy grande.

TABLE OF CONTENTS

ABBREVIATIONS AND ACRONYMS

BFB:	Breakage-fusion-bridge
bp:	Base pairs
cDNA:	Complementary DNA
ChIP-Seq:	Chromatin Immunoprecipitation followed by sequencing
CN:	Copy number
CR:	Clipped-reads
DHS:	DNase hypersensitivity
DNA:	Deoxyribonucleic acid
DP:	Discordant read-pair
EGFR:	Epidermal growth factor receptor
ERK:	Extracellular signal-regulated kinases
ERV-K:	Endogenous retrovirus-K
FDR:	False Discovery Rate
FL-L1:	Full-length L1
FN:	False Negative
FP:	False Positive
FPKM:	Fragments per Kilobase of transcript per Million mapped reads
ICGC:	International Cancer Genome Consortium
IRs:	Indel reads
Kbp:	Kilobase pair
LINE-1/L1:	Long interspersed nuclear element 1
L1-EN:	L1 endonuclease
LRE:	L1 Retrotransposable Element
MAPK:	Mitogen-Activated protein kinases
MAPQ:	Mapping quality
Mbp:	Megabase pair
MEI:	Mobile element insertion
NHEJ:	Non-Homologous End-Joining
MMs:	Mismatches
mRNA:	Messenger RNA
ONT:	Oxford Nanopore Technologies
ORF:	Open reading frame
ORF1p:	ORF1 protein
PCAWG:	Pan-Cancer Analysis of Whole Genomes
PCR:	Polymerase chain reaction
PEM:	Paired-end mapping
PSD:	Processed pseudogene
RNA:	Ribonucleic acid
RNA-seq:	RNA sequencing
RT:	Replication time
SVs:	Structural variants

SVA: SINE-VNTR-Alus
TCGA: The Cancer Genome Atlas
TN: True negative
TP: True positive
TPRT: Target primed reverse transcription
TraFiC: Transposon Finder in Cancer
VCF: Variant Call Format
WGS: Whole Genome Sequencing

ABSTRACT

Retrotransposons are selfish DNA sequences which populate the genome of most eukaryotic forms of life, with approximately one third of the human genome being derived from retroelements. Their remarkable prevalence is a consequence of their ability to propagate using a copy and paste mechanism termed retrotransposition. Although retrotransposons have been traditionally considered as “junk” DNA, their mobilization in the germ cells has profoundly impacted the evolution of the human genome, contributing to the generation of new genes and regulatory sequences. Increasing lines of evidence indicate that retrotransposons can also mobilize beyond the germline, with recent cancer genome surveys reporting extensive somatic LINE-1 (L1) retrotransposition in diverse tumour types. However, somatic retrotransposition in cancer has been so far investigated in a limited number of genomes, with multiple tumour histologies remaining to be explored. Cancer genome studies have also typically focused on the identification of canonical mobile element insertion events, while retrotransposons are able to mediate more complex forms of genomic alterations, which remain uncharted in the context of cancer.

This PhD dissertation aims to investigate the patterns of activity and consequences of somatic retrotransposition in cancer through the analysis of a large cohort of cancer genomes compiled by the Pan-Cancer Analysis of Whole Genomes Consortium. Given the volume of data to be analyzed, which included 2,954 whole genomes from 35 different tumour types, we developed TraFiC-mem, a computational method for the detection of somatically acquired mobile element insertions. We observed particularly high levels of L1 retrotransposition in esophageal, head-and-neck, lung and colorectal cancers, where mobile element insertions represented a predominant class of structural variation. The bulk of somatic L1 retrotransposition in cancer originates from a small subset of 16 L1 copies with “hot” activity. Hot-L1s display two differentiated patterns of activity, which we term “Strombonian” and “Plinian” due to their similarity with volcano eruption patterns. The aberrant integration of L1 sequences can lead to diverse rearrangement classes, including deletions, translocations, inversions and duplications. L1-mediated rearrangements can have oncogenic consequences, leading to the recurrent deletion of tumour suppressor genes, such as CDKN2A, or the amplification of oncogenes, such as CCND1, through the initiation of breakage-fusion-bridge cycles. These observations illuminate a relevant role of L1 retrotransposition in remodeling the cancer genome, with potential implications for the development of human tumours.

RESUMEN

Los retrotransposones son el tipo de secuencia repetitiva de ADN más abundante en el genoma humano, representando aproximadamente un tercio de la secuencia genómica. Su remarcable prevalencia se debe a su habilidad para propagarse por medio de un mecanismo de "copia y pega" denominado retrotransposición. Aunque tradicionalmente se les ha considerado como ADN "basura", la movilización de los retrotransposones en las células germinales ha tenido un profundo impacto en la evolución del genoma humano, contribuyendo a la generación de nuevos genes y secuencias reguladoras. Cada vez más líneas de evidencia indican que los retrotransposones se movilizan más allá de la línea germinal. Estudios recientes sobre el genoma del cáncer han descrito altos niveles de actividad somática para retrotransposones de tipo LINE-1 (L1) en diversos tipos de cáncer. No obstante, el número de genomas investigados hasta la fecha es reducido, con múltiples tipos de cáncer todavía sin explorar. Por otro lado, los estudios sobre el genoma del cáncer se han centrado principalmente en la identificación de inserciones canónicas de elementos móviles. Sin embargo, se sabe que los retrotransposones son capaces de promover alteraciones genómicas más complejas, lo cual se trata de un proceso mutacional que permanece sin investigar en el contexto del cáncer.

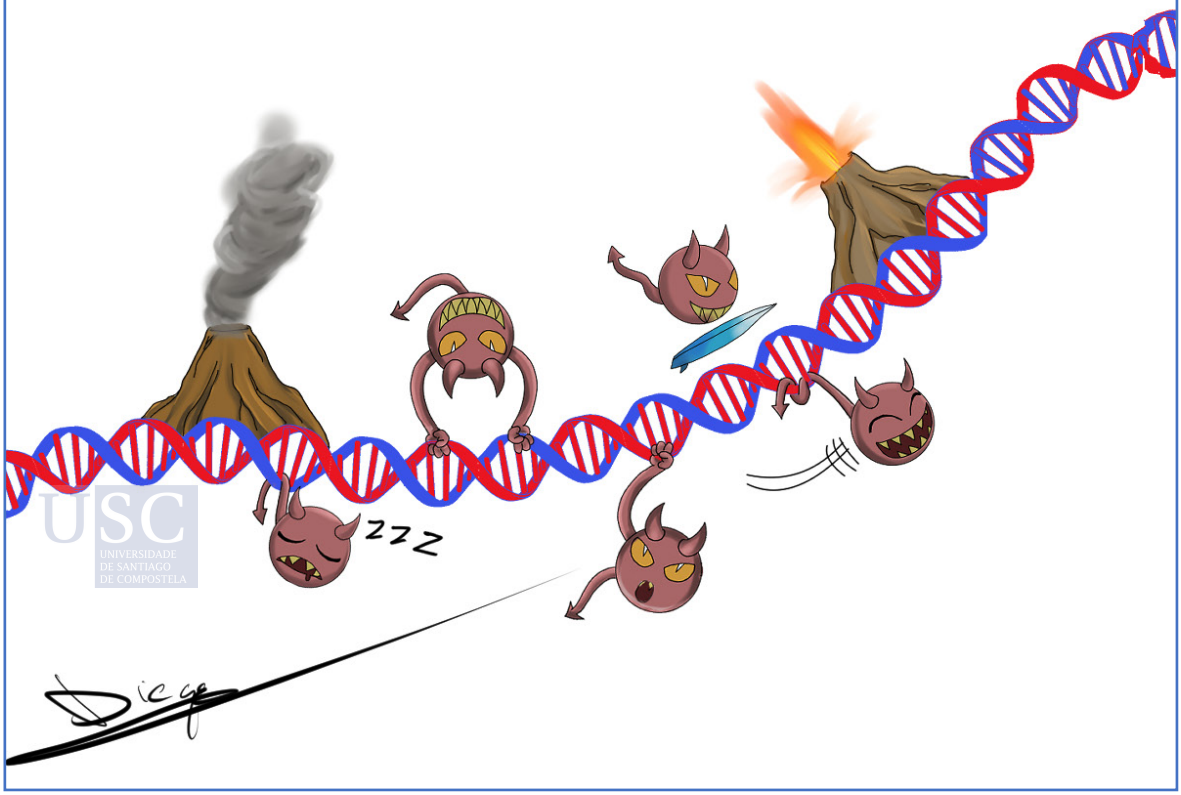
Esta tesis doctoral pretende investigar los patrones de actividad y las consecuencias de la movilización somática de los retrotransposones en el genoma del cáncer. Para ello se analizó una gran cohorte compuesta por 2,954 genomas completos pertenecientes a 35 tipos distintos de cáncer que fueron recopilados por el Consorcio Pan-Cancer. Dado el gran volumen de datos a analizar desarrollamos TraFiC-mem, un método computacional para la detección de inserciones somáticas de elementos móviles. Identificamos niveles particularmente altos de actividad de L1 en cánceres de esófago, cabeza-y-cuello, pulmón y colon, donde la inserción de elementos móviles supuso una clase predominante de variación estructural. El grueso de la actividad somática de L1 deriva de tan sólo 16 copias altamente activas. Estas copias muestran dos patrones de actividad bien diferenciados, los cuales denominamos "Strombolianos" y "Plinianos" debido a su similitud con los patrones de erupción volcánica. La integración aberrante de secuencias L1 puede producir diversos tipos de reordenamientos, incluyendo pérdidas, translocaciones, inversiones y duplicaciones de ADN. Los reordenamientos mediados por L1 pueden tener consecuencias oncogénicas, causando la pérdida recurrente de genes supresores tumorales, tales como CDKN2A, o la amplificación de oncogenes, como CCND1, a través de la iniciación de ciclos de rotura-fusión-puente. Estas observaciones revelan un papel relevante de los retrotransposones de tipo L1 en la reorganización del genoma del cáncer, con posibles implicaciones en el origen y desarrollo de los tumores.

RESUMO

Os retrotransposóns son o tipo de secuencia repetitiva de ADN máis abundante no xenoma humano, representando aproximadamente un tercio da secuencia xenómica. A súa remarcable prevalencia débese á súa habilidade para propagarse por medio dun mecanismo de "copia e pega" denominado retrotransposición. Aínda que tradicionalmente considerouse aos retrotransposóns como ADN "lixo", a súa movilización nas células xerminais tivo un impacto profundo na evolución do xenoma humano, contribuíndo á xeración de novos xenes e secuencias reguladoras. Cada vez máis liñas de evidencia indican que os retrotransposóns móvense máis alá da liña xerminal. Estudos recentes sobre o xenoma do cancro describiron altos niveis de actividade somática para retrotransposóns de tipo LINE-1 (L1) en diversos tipos de cancro. Non obstante, o número de xenomas investigados ate o momento aínda é reducido, con múltiples tipos de cancro aínda sen explorar. Doutra banda, os estudos sobre o xenoma do cancro centráronse principalmente na identificación de inserción canónicas de elementos móbiles. Sen embargo, sábese que os retrotransposóns son capaces de promover alteracións xenómicas máis complexas, as cales permanecen sen investigar no contexto do cancro.

Esta tese de doutoramento pretender investigador os patróns de actividade e as consecuencias da mobilización somática dos retrotransposóns no xenoma do cancro. Para iso analizouse unha gran cohorte composta por 2.954 xenomas completos pertencentes a 35 tipos distintos de cáncro que foron recopilados polo Consorcio Pan-Cancer. Dado o gran volume de datos a analizar desenvolvemos TraFiC-mem, un método computacional para a detección de inserción somáticas de elementos móbiles. Identificamos niveis particularmente altos de actividade L1 en cancros de esófago, cabeza-e-pescozo, pulmón e colon, onde a inserción de elementos móbiles supuxo unha clase predominante de variación estrutural. A meirande parte da actividade somática de L1 deriva de tan só 16 copias altamente activas. Estas copias mostran dous patróns de actividade ben diferenciados, os cales denominamos "Strombolianos" e "Plinianos" debido á súa similitude cos patróns de erupción volcánica. A integración aberrante de retrotransposóns de tipo L1 pode producir diversos tipos de reordenamentos, incluíndo pérdidas, translocacións, inversións e duplicacións de ADN. Os reordenamentos mediados por L1 poden ter consecuencias oncoxénicas, causando a pérdida recurrente de xenes supresores tumorais, tales como CDKN2A, ou a amplificación de oncoxenes, como CCND1, a través da iniciación de ciclos de rotura-fusión-ponte. Estas observacións revelan o papel dos retrotransposóns L1 na reorganización do xenoma do cancro, con posibles implicacións na orixe e no desenvolvemento dos tumores. No anexo ofrécese un resumo ampliado en galego

INTRODUCTION



I.1 The cancer genome

The human genome is a large encyclopedia of approximately 6.4 billion base-pairs of deoxyribonucleic acid (DNA) sequence that encodes all the necessary information for the development of a human being. Each individual can be understood as a large community of approximately 37.2 trillion of cells[1] that become differentiated in multiple cell types and cooperate for the normal function of an organism. Every cell in an individual has its own copy of the genomic DNA sequence, which orchestrates every molecular process occurring within the cell through complex regulatory networks ensuring homeostasis. Deregulation of cellular processes results in different forms of imbalances that can lead to the development of pathological conditions, such as cancer.

Cancer is a major cause of death, accounting for one in eight deaths worldwide[2]. The term 'cancer' is an umbrella which is used to name a plethora of diseases originating from most cell types and organs of the human body, albeit sharing several definitory features. Every cancer is characterized by unscheduled and unrestrained proliferation of cells which eventually invade beyond normal tissue boundaries and metastasize to distant organs. Tumoral cells frequently disrupt human body homeostasis, interfering with the function of organs, eventually causing the death of the individual, with metastases being the primary cause of death from cancer[3].

Early insights into the central role of the genomic alterations in cancer development date back to the late nineteenth and early twentieth centuries. In two seminal studies, David von Hansemann[4] and Theodor Boveri[5] reported the presence of intriguing chromosomal aberrations in cancer cells under cell division using a microscope. This observation led to the hypothesis that cancer arises from abnormal cell clones containing defects on their genetic material. After a century of cancer research since the publication of these studies, most cancers are currently thought to be the consequence of pathogenic DNA mutations (i.e., known as cancer driver mutations) that are acquired during the lifespan of an individual.

Soon after egg fertilization, each cell in the human body is subjected to diverse mutational processes leading to the somatic acquisition of genetic variation (i.e., somatic mutations), ranging from simple base pair substitutions to large-scale chromosomal rearrangements[6]. Although the majority of variants are likely to be neutral, a small subset of them is likely to have an impact on the phenotype of the cell. Selection may act on the resulting phenotypic diversity by removing cells with deleterious somatic mutations or fostering cells containing variants that enhance their proliferative and survival advantages relative to the neighboring cells[6]. This process is likely to operate during the lifespan of any individual, with every adult human having probably thousands of minor winners of this ongoing competition, most of which with reduced abnormal growth capabilities[6]. However, a single cell may eventually acquire a critical number of advantageous mutations which enable its unrestrained proliferation leading to cancer[6].

For the last four decades, one of the central aims of cancer research has been the identification of cancer genes, which are those genes whose mutation causes cancer. This research

direction was initiated in 1982, when the first naturally occurring, cancer-causing somatic single base substitution was identified at codon number 12 of the *HRAS* oncogene in the bladder carcinoma cell line T24/EJ[7]. This and other findings, including the identification of the breast cancer susceptibility gene *BRCA2*[8], launched a new era of cancer genome research, which expanded the catalogue of cancer genes to more than 500 loci[9,10], representing more than 2.5% of the genes in the human genome.

Cancer driver genes are generally classified into oncogenes, which are affected by activating mutations that increase the normal level of activity of the gene, and tumour suppressors genes, affected by inactivating mutations that reduce/remove their activity. A classic example of oncogene is *BRAF*[11], a serine/threonine kinase that regulates the MAP kinase/ERKs pathway, which is pivotal for cell division and differentiation. Mutations in *BRAF* typically promote its constitutive activation, driving uncontrolled cell division and tumorigenesis. On the other hand, the *TP53* gene has a critical role as tumour suppressor[12], as it activates apoptosis to safeguards the maintenance of genomic integrity, among other critical functions. Inactivating mutations in this gene are observed at more than 50% of all human tumours[13], making *TP53* the most frequently mutated cancer gene. Other tumour suppressor genes, such as *BRCA1*, *BRCA2* or *ATM*, are critical components of the cellular DNA repair machinery[14,15], ensuring the stability of genomic sequence.

The development of high-throughput methods for the sequencing of DNA molecules at the turn of the millennium led to the first genome-wide surveys of somatic mutations in cancer[16,17]. The increasing throughput of DNA sequencing platforms enabled the systematic detection and characterization of somatic mutations for large cohorts of cancer genomes. Given the complexities of this endeavor, the International Cancer Genome Consortium[18] (ICGC) and its American counterpart, The Cancer Genome Atlas[19] (TCGA), were created to coordinate this herculean effort, maximizing the use of resources and harmonizing analytical approaches across teams. For the last decade, both consortia have been major powerhouses of cancer genome research, resulting in great advances in our understanding of the cancer genome, including the discovery of novel cancer genes[9,10], and the identification of diverse patterns and mechanisms of somatic mutation[20]. A particularly intriguing mutational process is chromothripsis[21], which involves chromosome shattering, producing extensive chromosomal rearrangements, occasionally affecting multiple cancer genes in a single catastrophic event.

The understanding of the genetic bases of carcinogenesis has already impacted patient prevention, diagnosis and treatment. Breast cancer is a relevant example of how cancer research has led to effective cancer prevention programs. Women carrying inherited pathogenic mutations at either of the breast cancer type 1 or 2 susceptibility genes (*BRCA1* and *BRCA2*) are typically subjected to breast surveillance programs, including annual mammograms and periodical clinical breast examination. On the other hand, multiple targeted therapies have been developed to treat cancers originating as consequence of mutations affecting certain cancer genes. For example, inhibitors of the epidermal growth factor receptor (*EGFR*), such

as gefitinib and erlotinib, are selectively active against lung cancers with hyperactive, mutated versions of the *EGFR* gene[22], but ineffective for tumours with wild-type *EGFR*. Overall, these advances promise to lead into a personalized biomedical era, where cancers are diagnosed, stratified and treated based on their profile of genetic alterations, maximizing the effectiveness of treatments while reducing side effects.

I.2 The Pan-cancer Analysis of Whole Genomes Project

Cancer research and patient care has been traditionally stratified by tumour types, with oncology departments at cancer centers and research groups focused on specific cancer histologies. As a consequence, ICGC and TCGA Consortia were structured in a similar manner, with researchers organized into working groups, each of them devoted to the molecular characterization of a different tumour type.

However, genome analysis has revealed important shared features between tumours derived from different tissue types and organs. For example, mutations at components of the DNA mismatch repair pathway or at the DNA polymerase epsilon led to characteristic hypermutator phenotype in colorectal[23] and endometrial cancers[24]. Similarly, *BRCA1* inactivation in breast and ovarian cancer is related with defective homologous-recombination DNA repair[25], leading to a characteristic short (<10Kb) tandem duplicator phenotype. Mutations at the *CDK12* kinase in ovarian and prostate cancer result in an increased number of mid-to-large sized tandem duplications[26,27]. Commonalities between cancer types can also impact tumour treatment. For example, the oncogene *ERBB2-HER2* is mutated or amplified in subsets of glioblastoma, breast, gastric, bladder, lung and serous endometrial cancers[9,10], which are typically responsive for *HER2*-targeted therapy[28]. These examples highlight the importance of approaching cancer research from a pan-cancer perspective, as shared features between tumour classes will enable the translation of therapeutic and prognostic approaches from one cancer type to another.

The large amount of whole genome sequences (WGS) generated for multiple tumour types from individual ICGC and TCGA working groups represented an unprecedented opportunity to perform a comprehensive meta-analysis across tumour types. With this mission, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the ICGC and TCGA was established in 2016. A technical working group implemented bioinformatic workflows to aggregate raw sequencing data from diverse sources, ensuring that every genome was processed, including read alignment and variant calling, in a harmonized fashion. A total number of 2,658 whole-cancer genomes and their matching normal samples across 38 tumour types were included in the PCAWG cohort[29]. The dataset comprised 1,469 men (55%) and 1,189 women (45%), with a mean age of 56 years (range: 1–90 years).

PCAWG scientists were organized into thematic working groups, covering different aspects of cancer genome biology, including cancer driver genes, mutational signatures, patterns of

structural variation, functional impact of somatic mutations and cancer evolutionary history. This PhD thesis summarizes the work and research findings derived from the mobile element subgroup, which investigated the activity patterns of retrotransposons across this large cohort of cancer genomes[30]. Within this working group, I played a central role, performing the bulk of the computational analysis and participating in the coordination of the group together with my PhD advisor.

I.3 Mobile elements and cancer

Retrotransposons are the most abundant class of DNA repeats in the human genome, representing about one third of the genomic sequence[31]. Their abundance is a consequence of their ability to spread through the genome using a copy and paste mechanism termed "retrotransposition" and their continued mobilization over millions of years of genome evolution[32]. During retrotransposition, active retroelements are transcribed into RNA molecules that are subsequently retrotranscribed into cDNA and integrated elsewhere in the genome.

Evidence indicates that three mobile element families are currently active on the human genome[31]. The most abundant family is the Alu repeat, which accounts for ~11% of the genomic sequence. Alu repeats are short (300 bp on average), but numerous, with more than 1 million copies populating the human genome. In contrast, the long interspersed nuclear element 1 (LINE-1 or L1) family, although less abundant (500.000 L1 copies), is composed of sequences up to 6 Kbp in size, representing 17% of the human genome. The SINE/variable number of tandem repeats/Alu (SVA) family is the youngest and less abundant, with ~3.000 copies scattered through the human genomic sequence. It has been traditionally assumed that retrotransposon mobilization is restricted to the germ cells, enabling their transmission to future generations. However, now we know that the somatic mobilization of L1 retrotransposons frequently occurs in many cancer types, with potential consequences for tumour development.

This research avenue on cancer retrotransposition was initiated by Buzzy Morse and colleagues in 1988, when they reported a somatic L1 insertion into the *myc* locus in a patient diagnosed with a breast carcinoma[33]. The insertion was absent in the normal breast tissue from the patient, being therefore acquired somatically. Four years later, an additional study identified a second somatic L1 integration disrupting the sequence of the tumour suppressor gene *APC* in a patient with colon cancer[34]. Given that *APC* is typically one of the first mutated genes during colon cancer progression, this event highlighted the potential of endogenous retrotransposons to promote cancer-causing mutations. Subsequent research reported L1 hypomethylation[35,36], increased L1 transcription[37] and ORF1p expression[38] in human tumours relative to normal tissues, suggesting that cancer associated epigenetic changes may lead to the activation of L1 copies during the course of cancer.

The first genome-wide survey of somatic retrotransposition in cancer dates back to 2010, when Scott E. Devine's team developed transposon-seq, a high throughput approach for the

identification of retrotransposon insertions from DNA samples[39]. After extensive validation, they applied transposon-seq to a collection of eight cancer cell lines and 30 primary samples from multiple tumour types (i.e., leukemias, breast, lung and brain cancers). A total of nine tumour-specific L1 insertions were identified after extensive filtering in order to remove rare germline polymorphisms, while no evidence for Alu and SVA somatic retrotransposition was found. All somatic events were absent from adjacent normal tissues from the same patient and were restricted to lung tumours. Although the number of samples analyzed and somatic L1 insertions identified was limited, this study served as a harbinger of the research that was about to come.

The standardization of massive parallel short-read sequencing led to the emergence of large-scale cancer genome consortia (e.g., ICGC and TCGA). As these initiatives aimed to sequence hundreds of cancer genomes from diverse tumour types, they provided the ideal environment for a new wave of scientific discoveries regarding somatic retrotransposition in cancer. In 2012, Peter J. Park's laboratory developed TEA, a computational approach for the detection of somatic retrotransposon insertions from paired-end whole genome sequencing data[40]. Using this algorithm, they detected a total of 194 somatic retrotranspositions, the majority L1s, in 43 cancer genomes from five different tumour types, including colon, prostate, ovarian, multiple myeloma and glioblastoma. Notably, all the retrotransposition events identified were restricted to tumours of epithelial cell origin (e.g., colorectal, prostate, and ovarian), with tumours of the colon and rectum having the higher number of somatic L1 insertions. While the number of L1 insertions typically ranged from a handful to a few dozen events, one colorectal cancer case harbored 106 somatic insertions, indicating that high somatic L1 retrotransposition rates occasionally occur in colon cancers.

Additional studies extended the survey of somatic retrotransposition to more than 400 cancer genomes[41–44], including unexplored cancer types at that time, such as head-and-neck, bone, uterus and kidney. This research corroborated the high levels of somatic L1 activity in epithelial tumours, with tumours from the gastrointestinal tract, namely esophageal[45,46], gastric[47], small bowel[47] and pancreatic[48] tumours having particularly high rates of somatically acquired L1 insertions. Outside the gastrointestinal tract, head-and-neck and non-small-cell lung cancers also show frequent somatic L1 insertions[43]. Meanwhile, somatic retrotranspositions are rarely found in hematological malignancies, such as multiple myelomas[40] and acute myeloid leukemia[43], and tumours derived from the central nervous system and brain, such as gliomas[39,40] and glioblastomas[49,50].

While these studies typically focused on primary tumours, Tubio and colleagues investigated somatic L1 activities in multiple cancer biopsies, including metastases, from 11 patients with prostate cancer[44]. They observed a marked heterogeneity in the distribution of somatic L1 insertions, with a subset of L1s being shared between primary and metastatic samples, while others being sample specific. Metastatic samples harbored higher numbers of insertions relative to primary cancers, suggesting that increased levels of retrotransposition may occur

during the later stages of prostate cancer progression. A subsequent study reported similar patterns of intra-tumoral heterogeneity for somatic L1 retrotransposition during the evolution of pancreatic ductal carcinoma[48], with only a small fraction of L1 insertions being shared between primary and metastatic samples. Somatic L1 retrotranspositions have been also identified in esophageal cancer and its precursor, Barret´s esophagus, suggesting that retrotransposon mobilization can also occur in the early stages of cancer[45]. Altogether, these studies indicate that the somatic mobilization of L1 sequences can occur from the early to late stages during the evolution of cancer.

The large majority of somatic L1 insertions described to date in cancer are intronic or intergenic[40,43,44], with a bias in the distribution of somatic insertions towards gene-poor, late replicating and heterochromatic genomic regions[43,44]. Among those integrations occurring within gene boundaries, gene expression changes have been observed for a limited subset[40,43,44], suggesting that the majority of L1 integrations may be passenger events, as described for other mutation classes[6]. As a consequence, only a handful of plausible cancer causing insertions have been reported to date[34,51], all of them targeting exons of tumour suppressor genes. However, mobile elements can alter gene function through other mechanisms[32], which remain to be explored in the context of cancer. These include splicing alterations due to exon skipping or mobile element exonization, the introduction of novel transcription starts sites, premature end of transcription by addition of new polyadenylation signals, deregulation due to gene promoter/enhancer disruption or epigenetic changes.

Since the discovery of the first somatic L1 insertion in 1988[33], the scientific community has gained major insights into the activity of retrotransposons in the cancer genome. However, many questions remained to be explored when I started my PhD. Somatic retrotransposition has been studied in a reduced number of cancer genomes, with multiple tumour types not investigated yet. The factors underlying the differences in the amount of L1 mobilizations across different individuals and cancer types are not well understood. While most cancer genome studies have focused either on somatic retrotransposition or more classical structural variants, such as deletions, duplications or inversions, the interaction between both types of genetic variants and their relative contributions to the mutational landscape in the cancer genome has not been addressed yet. Although several studies have reported gene expression changes associated with the somatic insertion of L1 sequences, the importance of other mechanisms for gene alteration, such as exonization and aberrant splicing, remain to be investigated in the context of cancer.

I.4 Source L1 elements

During the 1980's, the isolation and comparative analysis of full-length L1 (FL-L1) sequences across multiple species led to an intriguing observation. FL-L1 sequences have two open reading frames (ORF1 and ORF2) that have been evolving under selection for their coding potential since the mammalian radiation[52,53]. Now we know that both ORFs encode for the proteins

necessary for L1, Alu and SVA retrotransposition and propagation across eukaryotic genomes. ORF1 encodes for a nucleic acid chaperone that binds to L1 RNA forming a ribonucleoprotein particle[54,55]; and ORF2 encodes for a protein with endonuclease[56] and reverse transcription[57] activities that enables L1 insertion via target-primed reverse transcription[58].

However, the vast majority of L1 copies in the human genome are defective due to internal mutations, in addition to truncation and inversion of their 5' ends[31]. Among this myriad of inactive L1 sequences, a limited set of 90 FL-L1 loci with intact ORFs and, therefore, potentially active, resides in the human reference genome[59]. Recent large-scale sequencing studies of hundreds of human genomes from diverse populations have uncovered thousands of non-reference L1s[60], including FL-L1s, expanding the repertoire of potentially active L1s in the human population. Active L1s are usually termed as source L1 elements, since they are progenitors and source of newly acquired insertions. Retrotransposon insertions derived from source L1 activity are a relevant class of genetic variation, which can lead to certain diseases, such as mendelian disorders[61] and cancer[34]. As a consequence, the identification and characterization of all active L1 copies in the human genome constitute an overarching aim in L1 research.

This research line was initiated by Kazazian's laboratory in 1991. In a seminal study, Dombroski *et al.* isolated the first L1 source element, L1 Retrotransposable Element 1 (LRE-1), as the progenitor of a truncated de-novo L1 insertion into the factor VIII gene causing hemophilia A[62]. LRE-1 was mapped to a locus on chromosome 22q where it has resided for at least 5 million years. Another two disease-causing de-novo L1 insertions led to the discovery of two novel L1 source elements (LRE-2[63] and LRE-3[64]). In both instances, the inserted L1 sequence had a companion non-repetitive sequence corresponding to a 3' transduction.

L1 3'-transductions (from now on, L1 transductions or transductions) originate when the transcription machinery skips the L1 polyadenylation signal and uses an alternative site instead, which is located downstream to the element[65,66]. As a consequence, the L1 is mobilized together with a non-repetitive bit of adjacent DNA sequence to a new genomic position. As L1s usually get truncated on their 5' ends during retrotransposition, the transduced piece of sequence can be integrated with or without a companion L1, leading to the generation of a partnered and an orphan transduction, respectively. Given the non-repetitive nature of the transduced sequence, they can be used to trace back the location of the progenitor copy, which enabled the mapping of LRE-2[63] and LRE-3[64] to their genomic locations at 1q and 2q24.1, respectively. Subsequent studies have revealed that L1 transductions have contributed remarkably to the evolution of the human genome, with approximately 25% of all L1 insertions in the reference harboring transduced sequences[66], which on occasion span coding or regulatory sequences[65].

The next major advance for the detection of active L1s was the development of *in vitro* retrotransposition assays[67], which allowed the quantification of the levels of activity for cloned FL-L1 sequences. The application of this approach for the assessment of 13 newly isolated FL-L1s led to the identification of three additional active elements, and to the estimate

that a human diploid genome bears between 30 and 60 active L1s, on average[68]. Six years later, Brouha and colleagues extended these analyses to the 90 FL-L1s contained within the released human genome sequence[59]. A total of 40 FL-L1s were active *in vitro*, expanding the catalogue of active L1s from six to 46 copies, and doubling previous estimates for the number of active L1s in the human genome. Brouha *et al.* also reported remarkable differences on the levels of retrotransposition activity among source L1 elements[59], with six elements alone being responsible for 84% of the assayed retrotransposition capability. This suggests that most L1 retrotransposition activity in the human genome emerges from a reduced number of extremely active L1s, which they termed as "hot-L1s'', with other elements playing a minor role.

The availability of the reference genome sequence and the development of high-throughput sequencing methods led to the initial efforts to create comprehensive catalogues of genetic variation polymorphisms in humans[69–71]. These include fosmid-based methods, which were used for the detection of polymorphic structural variation[72]. In 2010, Moran's laboratory applied this approach for the identification of polymorphic FL-L1s, detecting and isolating 68 copies, which were absent in the human reference[73]. Most of them (71%; 48/68) displayed *in vitro* activity, indicating higher levels of activity for recently acquired L1 polymorphisms relative to fixed copies at the reference (44% of FL-L1 in the reference are active *in vitro*[59]).

While these findings originated from either the sporadic detection of germline L1 transductions causing mendelian disorders or *in vitro* assays, the patterns of somatic activity for source L1s in human tissues and cancer remained largely unexplored. The emergence of large-scale sequencing studies of cancer (i.e., ICGC and TCGA), provided the ideal environment to tackle these questions. In 2014, Tubio and colleagues developed TraFiC, a computational method for the detection of somatic L1 mobilizations and transductions in cancer whole genomes[44]. After screening a collection of 244 cancer genomes from diverse tumour types, they identified frequent L1 transductions, occasionally disseminating genes, exons, and regulatory elements to new genomic locations. Orphan transduction (i.e., insertions without companion L1) were frequent, comprising half of all transduction events[44]. L1 transductions were traced back to 72 progenitor germline L1s, generating a catalogue of source L1 elements somatically active in cancer. Only 18 of these were previously reported to be active based on *in vitro* assays[59,73], indicating that the repertoire of active L1s still remains far from being completed.

Most transductions arised from a small set of hot-L1s, with four loci accounting for 50% of all transductions[44]. This is consistent with previous data derived from *in vitro* assays[59] and further supports the idea that L1 retrotransposition primarily stems from a reduced set of hot loci. The most active element, accounting for one fourth of all transductions, is located at the cytoband 22q12.1. According to its high levels of activity and its genomic position, this source element is likely to be LRE-1, the first reported source element that was mapped to chromosome 22q as the progenitor of an insertion causing hemophilia A[62].

Analysis of multiple biopsies taken at different stages of cancer development, including metastases, indicated that the activity of individual source elements was dynamic, with different sets of active source L1s along tumour evolution[44]. Active source elements were typically hypomethylated at their promoters, suggesting that methylation changes occurring during tumorigenesis dictate source L1 activation in certain tumours[44]. In a separate study, Scott and colleagues reported that the mobilization of a hot-L1, which evaded somatic repression via promoter demethylation, inactivated the tumour suppressor gene *APC*, initiating colon cancer[51]. This hot-L1 is a polymorphism restricted to individuals of African ancestry, highlighting that differences in the composition of the inherited set of germline source L1s for each human are likely to influence his predisposition to acquire mutations derived from somatic retrotransposition.

A quarter of century after the discovery of LRE-1, the scientific community has gained major insights into the repertoire of active L1 copies and their patterns of activity in cancer genomes and in the germline. Nonetheless, the catalogue of active L1s is likely to be largely incomplete, as it derives from laborious *in vitro* retrotransposition assays or from the identification of transductions in a few genomes. On the other hand, the patterns of source L1 activity in cancer have been investigated in a still reduced collection of cancer genomes and tumour histologies. As a consequence, the relevance of source L1s in cancer, which behave as mutagenic loci, is still poorly understood.

I.5 L1-mediated structural variation

Endogenous retroelements shape and mutate the human genome through diverse mechanisms[32,74]. The most evident is insertional mutagenesis, which results from the mobilization of retrotransposons during their life cycle. L1 encoded proteins can also act in trans to reverse transcribe and integrate cellular RNAs generating processed pseudogenes (PSD)[75]. In addition, L1 and SVA repeats can transduce flanking non-repetitive sequences which are copied along with the retroelement and integrated at new genomic positions[65,76]. Homologous transposable element sequences can mispair and ectopically recombine via unequal crossing-over to produce deletions or duplications of genomic DNA[32,74]. The aberrant integration of retrotransposons can also lead to chromosomal rearrangements[77,78], a mechanism of mutation that has remained particularly unexplored despite its major potential to alter the human genome.

The first evidence of retrotransposon-mediated rearrangements date back to 2002, when Boeke and Moran's laboratories implemented an enhanced system for *in vitro* retrotransposition assays, which enabled the isolation of *de novo* acquired insertions[77,78]. Through this approach, they isolated and cloned a total of 79 L1 insertions generated *in vitro* and performed a comprehensive characterization of their integration patterns. Most insertions contained the classical hallmarks for retrotransposition, such as target site duplications and deletions, L1 endonuclease (L1-EN) cleavage sites at insertion breakpoints, 5' truncation or inversion, frequent transductions and poly(A) stretches at their 3' ends. However, 6% of the insertions were associated with large-

scale chromosomal rearrangements, including four deletions up to 71 Kbp in size and a ~120 Kbp inversion. The L1 inserts found at the rearrangement breakpoints were 5' truncated and had intact poly(A) tails, indicative of passage through a mRNA intermediate. In addition, L1-EN cleavage motifs were detected at the rearrangement breakpoint positions. Alogether, the observed sequence features (i.e., truncation, endonuclease motifs and poly(A) tails) suggest that the described rearrangements occurred during target primed reverse transcription[58] (TPRT), the canonical mechanism for L1 integration.

L1 retrotransposition via TPRT is initiated by the cleavage of the first DNA strand by the L1-encoded endonuclease at the target motif of 3'-AA|TTTT-5'[79,80]. The resulting T-rich single-stranded DNA serves for the attachment of L1 mRNA poly(A) tail and provides a 3'-hydroxyl that primes reverse transcription of the L1 transcript. L1 integration proceeds after the cleavage of the second DNA strand by a currently unknown mechanism, finalizing once the nascent L1 cDNA is joined to the second-strand nick. The cleavage of the second-strand typically occurs a few base pairs downstream or upstream with respect to the initial cleavage site, leading to small duplications or deletions of target site nucleotides[81], which are typically between one and 25 bp in size. The sequence features observed for the L1-mediated rearrangements identified (i.e., truncation, endonuclease motifs and poly(A) tails), indicate that they were likely generated through a variant of the canonical TPRT reaction. For L1-mediated deletions, the second-strand cleavage may have occurred several Kbp upstream with respect to the first cleavage site, resulting in the loss of the DNA sequence between both cleavage positions once L1 retrotransposition is completed. Similarly, the reported L1-mediated inversion may result from a first and second nick occurring in the same DNA strand, which are located ~120 Kbp apart.

As these observations were performed *in vitro*, it was not clear if L1-mediated rearrangements can occur in living cells. In 2005, through the comparative analysis of human and chimpanzee genome sequences, Batzer's laboratory identified 50 L1-mediated deletions that occurred after the divergence of both species[82]. These resulted in the loss of ~18 Kbp and ~15 Kbp of the human and chimpanzee genomes, respectively, and suggested that during primate radiation, L1s may have deleted up to 7.5 Mbp of genomic sequences. These findings confirmed that L1-mediated rearrangements occur *in vivo* and showcases their impact on the evolution of the human and other primates' genomes. Similarly, the comparison of human and mouse genomes revealed 13 SVA-mediated deletions[83], which confirms that other mobile element families, such as SVAs, can mediate rearrangements during their retrotransposition. The identification of two somatic SVA-mediated deletions in patients with neurofibromatosis type 1[84] indicated that somatic retrotransposition events can also lead to chromosomal rearrangements. These events removed the *NF1* gene, which is mutated in 5-10% of neurofibromatosis type 1 cases, highlighting the potential of SVA-mediated rearrangements to cause genetic diseases.

These research studies have greatly expanded our understanding about the different mechanisms through which retrotransposon mobilization can impact the human genome. However, the number of rearrangements identified is still reduced, with most of them detected

in the germline or through *in vitro* assays. Although L1-mediated rearrangements occur frequently *in vitro*, with about 10% of L1 integrations being associated with DNA losses[77,78], their frequency *in vivo* is unknown. Comparative analysis of human, mouse and chimpanzee genomes, reveal a few dozens of events. However, the frequency of retrotransposon-mediated rearrangements in living cells may be higher than observed in the germline, as these are likely to be deleterious, therefore being subjected to purifying selection. In addition, beyond two anecdotal SVA-mediated deletions causing neurofibromatosis type 1[84], the importance of retrotransposon-mediated rearrangements on the etiology of genetic diseases, including cancer, has not been addressed yet. To conclude, aberrant integration of mobile elements can potentially lead to other rearrangement classes, such as duplications or translocations, which may remain to be discovered.

I.6 Computational methods for mobile element insertion detection

Multiple forms of genetic variation, including single nucleotide changes, small indels and structural variants (SVs), can alter the genomic sequence of any living organism. As genomic alterations can drive the development of multiple pathologies, ranging from mendelian disorders[85] to cancer[86], the research community has devoted enormous efforts to implement sensitive and accurate methods for the identification of genetic variation. From the dideoxy chain-termination method, developed by Sanger in 1977[87], to current high-throughput DNA sequencing approaches, technological advances for quantifying and reading DNA molecules have been milestones for the improvement of methods for the detection of genetic variants. These include mobile elements, which have always been a particularly challenging class of genetic variant to identify owing to their repetitive nature, with thousands of nearly identical copies interspersed through the genome.

The first observations that eukaryotic genomes are composed by a large amount of repeated sequences date back to the sixties[88]. One decade later, the development of DNA hybridization methods led the first estimates on the retrotransposon content of the human genome[89–91]. Together, Alu and L1 elements were estimated to account for 10% of the human genome. The development of molecular biology methods enabled the isolation and sequencing of DNA molecules, including retrotransposons[92]. These technological advances led to the initial efforts to assemble the genome of multiple species[93–95], which culminated with release of the first draft of the human genome in 2004[31]. Retrotransposons accounted for ~28% of the human reference genome sequence[31], which doubled previous estimates regarding the retrotransposon content of the human genome.

The availability of genome assemblies for multiple humans and other species facilitated comparative genome analysis, generating the first genome-wide catalogues for species-specific[96] and polymorphic genetic variation[97,98]. For example, Wang and colleagues[97] compared the public[31] (i.e. generated by the Human Genome Project) and Celera[99] human genome sequences, detecting 800 Alu insertions, which represented the largest dataset of

Alu polymorphisms up to that time. In parallel, the development of targeted high-throughput sequencing assays[100–102] enabled the identification of polymorphic retrotransposons absent in the human reference. These approaches rely on primers specific for the mobile element families known to be currently active in the human genome to capture DNA fragments containing recently acquired insertion polymorphisms. Fragments are typically sequenced after PCR amplification or cloning to determine the genomic position of the mobile element insertion.

Through this targeted approach, multiple groups screened DNA samples from geographically diverse human populations, extending the collection of known mobile element polymorphisms. For example, Ewing *et al.*[103] applied an Illumina-based targeted approach to identify 367 polymorphic L1s from a cohort of 25 humans, while Witherspoon *et al.*[104] used a similar strategy to identify 487 Alu insertions in 4 unrelated individuals. In 2005, Eichler's team[72] devised a WGS method for the detection of SVs greater than 8 Kbp in size, including mobile element insertions (MEIs), providing a more comprehensive view of the SV landscape than targeted approaches. This method relies on the generation of fosmid libraries for DNA fragments, which are then sequenced via capillary end-sequencing.

In 2007, Korbel and colleagues[105] developed paired-end mapping (PEM), an enhanced WGS method for the detection of SVs of at least 3 Kbp in size. Instead of using fosmids, PEM couples DNA sharing, fragments circularization and 454 sequencing to obtain paired-end reads from 3Kb-long inserts. Then, they devised a computational approach that aligns the paired-end reads into the reference genome to search for SVs, including deletions, inversions and insertions relative to the reference, based on the identification of four distinctive alignment signatures (a to d): (a) In the case of deletions, paired-ends span a genomic interval longer than expected based on the fragment size distribution of the sequencing library; (b) Genomic inversions are denoted by paired-ends aligning with different relative orientations; (c) Insertions of non-repetitive sequences are detected by read-pairs spanning an interval shorter than the expected span distribution; (d) Distal insertions are characterized by a set of reads aligning to the reference genome at the insertion breakpoint boundaries (known as anchors) while their respective mate reads align to a distal locus. MEIs correspond to the latter category, with the mates mapping to a transposable element annotated elsewhere on the reference genome[105,106]. After extensive validation assays, they applied PEM for SV discovery in two females of African and European ancestries, uncovering more than 1000 polymorphic SVs, most of them insertions and deletions. This study served to set the methodological framework for the implementation of current high-throughput paired-end sequencing approaches.

One year later, Campbell *et al.*[107] applied a similar paired-end mapping approach for the identification of somatically acquired SVs in cancer cell lines. As opposed to Korbel *et al.* method, they used Illumina sequencing technology, which produces shorter paired-end reads (i.e., 29–36bp) for DNA inserts of 300bp in length. Besides the SV classes described above, they detected translocations based on the search for read-pairs aligning in different chromosomes. In addition, they used reads spanning the rearrangement breakpoint junctions,

and therefore clipped during their alignment, to characterize the rearrangement breakpoints at base pair resolution, which provided hints of their mechanisms of origin. Approximately half of the acquired rearrangements exhibited short microhomologies between both break-ends, suggesting a non-homologous end-joining (NHEJ) mechanism of DNA break repair. They also inferred copy number (CN) based on counting reads for 15 Kbp genomic bins and the application of a segmentation algorithm, which were highly consistent with those derived from SNP array data. This study was instrumental for the subsequent adoption of Illumina paired-end sequencing, which is currently the most frequently used high-throughput approach for genetic variation detection at current genetics and cancer genome studies.

The widespread adoption of paired-end sequencing, fostered by its use at large-scale population genetics (e.g. 1000 Genomes Project[108]) and cancer genome consortia (e.g. TCGA[19] and ICGC[18]), led to the implementation of a large suite of computational methods for the discovery of SVs from whole genome sequencing data. Early algorithms, such as BreakDancer[109], relied exclusively on the search for discordant read-pairs, which provide the approximate location for SVs and are not sensitive enough for short SVs with sizes within the span range distribution. In contrast, DELLY[110] integrates discordant read-pair with clipped-read analysis to increase breakpoint resolution and detect deletions as short as 20bp. However, large SVs (e.g., deletions longer than 1Mb) remained to be undetectable for discordant and clipped-read based approaches, leading to the development of methods, such as LUMPY[111], which also incorporates read-depth information. Other tools, such as SvABA[112], attempt to assemble the sequence for candidate loci identified based on discordant read-pair analysis. The resulting contigs are then realigned into the reference in order to accurately resolve the rearrangement breakpoint junctions. All these methods use a matched-normal whole genome sample, typically derived from blood or adjacent healthy tissue, to differentiate between SVs somatically acquired during tumour development and inherited polymorphisms.

Although the approaches described above are able to detect most forms of SVs, MEIs remained particularly underexplored at that time as they imply specific challenges[106], which are inherent to the repetitive nature of retrotransposons, with thousands of copies interspersed through the human genome. As short-reads are smaller in size than retroelements, they will align to multiple annotated retrotransposons in the reference genome, generating mapping artefacts around them. To avoid false positives, MEI detection algorithms typically filter out candidate insertions if they overlap an annotated retrotransposon belonging to the same subfamily as the insertion. MEI callers also search for multiple hallmarks that define genuine retrotransposition insertion events: (1) polyadenylation at their 3' ends, (2) target-site duplications and deletions ranging from 1bp to 20bp in size, (3) frequent 5' truncation or inversion for L1 insertions and (4) transduced sequences for L1 and SVA insertions.

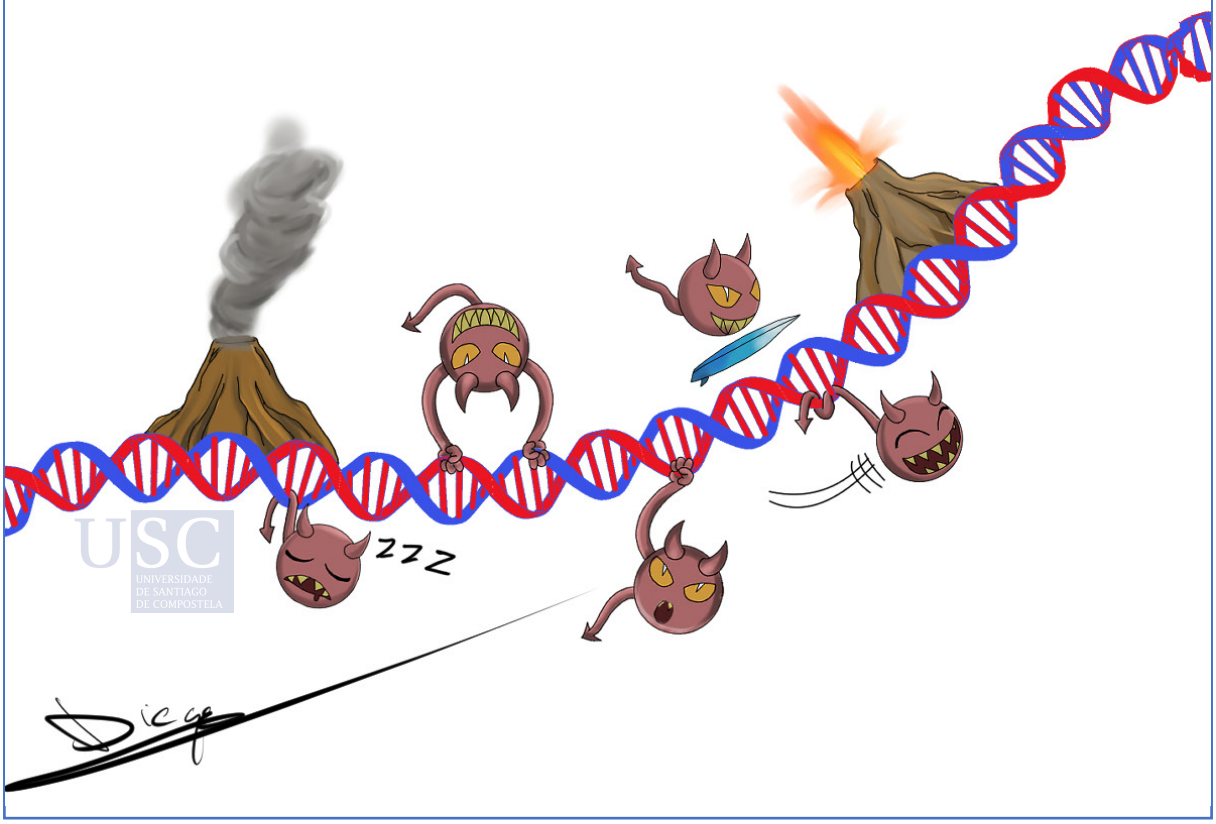
Most algorithms for retroelement insertion detection initially search for discordant read-pair clusters supporting candidate insertion events. These are defined as read-pairs with one end aligning non-ambiguously at the insertion position (anchored reads), while the other

end mapping to multiple genomic positions (non-anchored reads), as it corresponds to the integrated retroelement sequence. The most widely used approach for the identification of discordant read-pairs supporting a MEI is to realign non-anchored reads to a library of consensus sequences for active retrotransposon subfamilies. Methods like Tea[40], RetroSeq[43], Mobster[113] or TraFiC[44] use this strategy. MELT[114] uses an alternative approach, by searching for non-anchored reads mapping into annotated retrotransposons in the reference genome. Discordant read-pairs supporting a retrotransposon insertion from the same family are grouped into clusters, based on reciprocal overlap for anchored reads alignment positions and mapping orientation. As a result, two clusters of discordant read-pairs composed by anchors aligning in forward and reverse orientations, respectively, are detected, each demarcating one end of the retrotransposition insertion event. Algorithms such as TraFiC[44], use a specialized approach for the detection of transductions, which can be detected based on non-anchor reads aligning uniquely downstream to a somatic or germline FL-L1 insertion.

After discordant read-pair clustering, methods typically search for reads spanning the junction between the mobile element and the genomic sequence at the integration point. These reads typically get clipped during their alignment to the reference, with one piece mapping uniquely and demarcating the insertion breakpoint, while the other corresponding to the poly(A) tail or the 5' end of the mobile element. Clipped-reads are clustered based on their clipping positions and are used to determine multiple insertion features, such as the presence of poly(A) tails, target site duplications, the insertion length and the occurrence of 5' inversions. As a last step, candidate MEI are filtered based on multiple criteria[106], including the number and quality of discordant and clipped-reads supporting the event, the mappability or the presence of repetitive sequences of the same family at the insertion genomic position, and the presence of retrotransposition hallmarks described above. Somatic MEI callers like Tea[40], RetroSeq[43] or TraFiC[44] distinguish between somatic and germline insertions by tumour and matched-normal cross comparison.

Despite all these advances, currently available methods for the detection of MEIs typically require long computation times to process cancer whole genome data, making them not well suited for current large-scale surveys, involving hundreds of cancer whole genomes. In addition, while most methods are able to detect L1, Alu and SVA retrotransposition events, they have not been designed to search for non-canonical integration patterns. For example, only TraFiC[44] is able to identify L1-mediated transductions, while none of the methods able to detect processed pseudogenes and rearrangements associated with the integration of retrotransposon copies. This limitates their application for the investigation of novel mutational mechanisms drived by retrotransposons, creating the need for the development of tailored approaches, which will provide a more comprehensive view of the impact of retrotransposons in the cancer genome.

OBJECTIVES



This PhD project intends to investigate the activity of mobile elements in a large collection of cancer genomes from 38 tumour types within the context of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium. This will provide a comprehensive portrait of the patterns of somatic retrotransposition across diverse tumour histologies, and novel insights into the general mechanisms of mutation promoted by retrotransposons in cancer and their consequences on tumour development. The objectives and specific research questions (RQs) of my PhD thesis are:

OBJECTIVE 1: Characterization of the pan-cancer landscape of somatic retrotransposition

RQ1: What is the rate of somatic retrotransposition across cancer types?

RQ2: How somatic retrotransposition insertions alter gene function?

RQ3: What factors influence the distribution of somatic L1 integrations?

OBJECTIVE 2: Identification of source L1 elements and characterization of their patterns of activity in cancer

RQ4: What is the repertoire of active source elements in cancer?

RQ5: What are the pan-cancer patterns of source L1 activity?

OBJECTIVE 3: Identification of novel mutational processes mediated by L1 in cancer

RQ6: What are the mechanisms of L1-mediated rearrangement formation?

RQ7: Can L1-mediated rearrangements have oncogenic consequences?

OBJECTIVE 4: Development of algorithms for the detection of somatic retrotranspositions

RQ8: New methods for the detection of somatic mobile element insertions

RQ9: Method validation through orthogonal approaches

The characterization of the pan-cancer landscape of somatic retrotransposition (Objective 1) will tell us about the rates of somatic retrotransposition across diverse tumour types, the impact of mobile element insertions, including PSD, in gene function and the influence of distinct genomic features on the distribution of L1 integrations in the cancer genome. In Objective 2, we will catalogue germline full-length L1s active in cancer and investigate their patterns of activity. In Objective 3, we will search for novel mutational mechanisms of genomic rearrangement mediated by the integration of L1 sequences and investigate their impact on tumour development. The development of computational methods for the detection of somatic retrotransposition events in this large cohort of cancer whole genomes (Objective 4) will be instrumental for the achievement of Objectives 1-3. Research questions pertaining to objective 1 are covered in the Chapter 1 of this thesis, while those relative to objectives 2-4 are described in Chapters 2-4, respectively.

Objective 1: Characterization of the pan-cancer landscape of somatic retrotransposition

RQ1: What is the rate of somatic retrotransposition across cancer types?

Our participation in the PCAWG initiative will provide access to 2,954 cancer whole genomes from 38 different tumour types. Besides the large number of cancer genomes available, consortium members have collaborated on their comprehensive characterization, generating catalogues of somatic mutations, cancer driver events and gene expression expression quantifications, among others. We will leverage this unprecedented resource to perform a comprehensive survey of the activity of retroelements in cancer and to investigate the potential factors associated with high rates of somatic retrotransposition, including mutation in cancer genes and levels of genomic instability.

RQ2: How somatic retrotransposition insertions alter gene function?

We will assess the functional impact of somatic retrotransposition insertions - identified in RQ1 - through the analysis of the RNA-seq data available for 35% (1,043/2,954) of the PCAWG tumours. In particular, we will search for changes in gene expression associated with the insertion of retrotransposons at gene bodies. We will also identify splicing alterations driven by the integration of retrotransposons and processed pseudogenes at exons and introns. This will provide new perspectives regarding the mechanisms through which somatic retrotransposition can alter gene function.

RQ3: What factors influence the distribution of somatic L1 integrations?

We will investigate the genome-wide distribution of somatic L1 retrotranspositions - identified in RQ1 - across the cancer genome. Then, we will apply a statistical framework[115] to deconvolute the influence of multiple genomic covariates, including replication timing, histone marks and DNA accessibility, on the observed distribution. These analyses will provide further insights on the factors that determine the rate of somatic retrotransposition insertion along the cancer genome.

Objective 2: Identification of source L1 elements and characterization of their patterns of activity in cancer

RQ4: What is the repertoire of active source elements in cancer?

We will systematically use somatically acquired L1 transductions - detected in RQ1 - as barcodes to identify the source L1 elements whence they derive, generating a catalogue of germline L1s active in the PCAWG dataset. Given the large amount of cancer whole genomes to be analyzed, this catalogue may largely expand the collection of known active L1s in humans. We will also look into the patterns of activity for highly active (i.e., hot) source L1s and relate these with their allele frequencies in the human population.

RQ5: What are the pan-cancer patterns of source L1 activity?

We will study the activity profiles for active source L1s - identified in RQ4 - across multiple cancer types and the impact of source L1 activation in the rates of somatic retrotransposition.

Objective 3: Identification of novel mutational processes mediated by L1 in cancer

RQ6: What are the mechanisms of L1-mediated rearrangement formation?

The availability of CN and SV calls for all PCAWG cancer whole genomes represented an excellent opportunity to investigate the relevance of L1-mediated rearrangements in cancer. We will search for L1 insertions associated with CN alterations and SVs. Each L1 insert will be further inspected for target primed reverse transcription hallmarks (i.e., polyA tails and L1-EN motifs), which will serve to confirm that the associated rearrangement was originated by aberrant L1 integration. L1-mediated rearrangements will be classified as deletions, duplications, inversions and translocations, based on the analysis of the sequencing data and the nature of the CN or SV associated. This work will provide new insights regarding the impact of L1 retrotransposition in the cancer genome.

RQ7: Can L1-mediated rearrangements have oncogenic consequences?

We will investigate the impact of L1-mediated rearrangements in oncogenesis by searching for events - detected in RQ6 - leading to the recurrent deletion of tumour suppressor genes or the amplification of oncogenes.

Objective 4: Development of algorithms for the detection of somatic retrotransposition

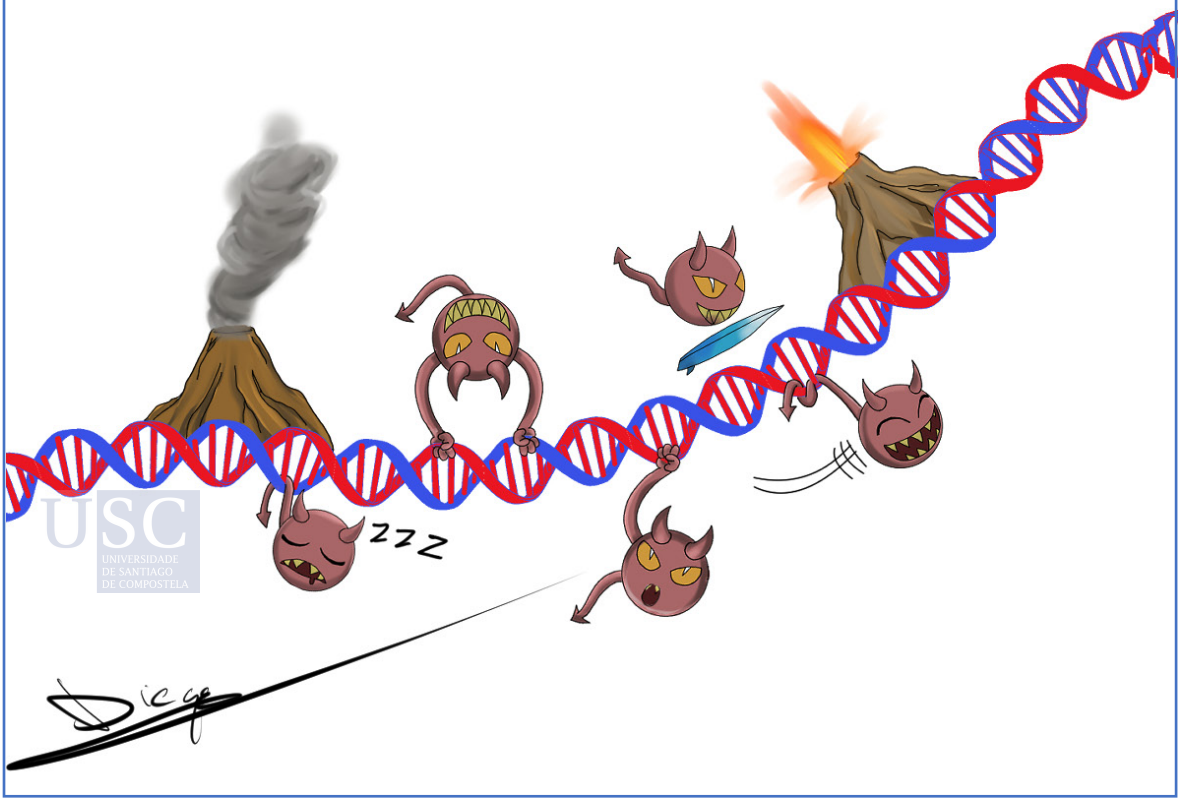
RQ8: New methods for the detection of somatic mobile element insertions

The detection and analysis of somatic retrotransposition insertions in 2,954 cancer whole genomes represents a great computational challenge. Given the massive volume of sequencing data to be processed, we will improve TraFiC[44] to detect somatic mobile element insertions more efficiently in terms of time and memory usage. We will also develop new computational methods for the identification of PSD and L1-mediated rearrangements, which are not detectable with currently available algorithms.

RQ9: Method validation through orthogonal approaches

The computational pipelines developed will be evaluated using multiple orthogonal approaches, which include simulations, long-read sequencing and PCR. This will provide estimates of the sensitivity and false discovery rates of our pipelines.

METHODS



MT.1 Pan-cancer datasets

Whole genome sequencing data

We analyzed Illumina paired-end WGS data for 2,954 tumours and matched-normal samples across 38 cancer types. On the basis of the robustness of the retrotransposition calls (false discovery rate of <5%; see C4.4), we opted to retain all samples that were originally blacklisted by the PCAWG Consortium[29]. These were excluded from SNV and SV analyses due to poor sequencing quality or biases on the read directionality, but were not problematic for the detection of somatic retrotransposition insertions. For the majority of donors, the tumour specimens were fresh frozen samples, whereas blood samples were used as matched-normals. Most of the tumour samples were biopsies taken from treatment-free primary cancers, although there was also a small number of donors with multiple samples derived from primary, metastatic and/or recurrent tumours. The average coverage for normal samples was 30 reads per bp, whereas tumours had a bimodal coverage distribution with maxima at 38 and 60 reads per bp (Figure S1). BWA-mem[116] v.0.7.8-r455 was used to align each tumour and normal sample to the human reference (build GRCh37) as described in the PCAWG flagship manuscript[29]. The TCGA Program Office and the Ethics and Governance Committee of the ICGC managed all the ethical aspects underlying the collection and use of human tissue and genomic data.

Transcriptome data

About half of the tumours (1,188) were also subjected to paired-end whole transcriptome sequencing (RNA-seq). RNA-seq reads were aligned onto the reference genome (build GRCh37) using two different alignment pipelines, one using STAR[117] v.2.4.0 while the other using TopHat2[118] v.2.0.12 for read mapping. Gene expression was quantified in fragments per kilobase of transcript per million mapped reads (FPKM) with HTSeq[119] v.0.6.1p1. Consensus normalized expression values were produced by averaging the FPKM derived from STAR and TopHat2 alignments. A more detailed description of the RNA-seq data processing workflow is provided by the “Integration of Transcriptome and Genome Working Group” in their companion manuscript[120].

Copy number and structural variation calls

CN profiles for all PCAWG tumours were generated by the “Evolution and Heterogeneity Working Group” through a consensus approach that combined six different state-of-the-art CN calling algorithms[121]. Consensus SV calls were generated by the “Somatic Structural Variation Working Group” by integrating SV calls from four SV calling pipelines[26]. Only SVs detected by at least two methods and with consistent change in CN were included in the consensus callset.

MT.2 Mobile element insertion detection and genotyping

Detection of somatic retrotranspositions in the PCAWG dataset

BAM files for 2,954 tumour and matched-normal WGS were processed with TraFiC-mem v1.1.0 (https://gitlab.com/mobilegenomes/TraFiC) to identify somatic MEIs, including solo-L1, L1-mediated transductions, Alu, SVA and ERV-K. Some donors (i.e., cancer patients) had multiple tumour samples available, which derived from biopsies taken from different tumour locations, including metastases. In those cases, each sample was independently processed with TraFiC-mem. Then, the list of MEI detected across all the samples from the same donor was merged into a non-redundant callset, using a breakpoint offset of ±15 bp to cluster together MEIs belonging to the same retrotransposon family. For each cluster, the MEI supported by the highest number of discordant read-pairs was selected as representative during the merging process.

Genotyping of germline source L1 elements

Source L1s absent on the reference genome were genotyped across the 2,822 matched-normal WGS using TraFiC-genotyper v1.1.0 (https://gitlab.com/mobilegenomes/TraFiC-genotyper). For each source locus and matched-normal sample, the algorithm counted the number of reads supporting the reference (i.e., L1 absence) and alternative (i.e., L1 presence) alleles. Reads supporting the reference allele spanned the source L1 insertion breakpoint with a minimum overhang of 20 bp, while those supporting the alternative allele were clipped at a maximum distance of 3 bp with respect to the insertion breakpoint position. Reads aligning to multiple genomic positions or clipped at both ends were filtered, as they usually represent mapping artefacts, leading to spurious genotype signals. The allele frequency (AF) for each source L1 was computed as the ratio between the number of reads supporting the alternative allele and the total number of reads (i.e., alternative plus reference). A heterozygous genotype call was made for source L1s with an AF between 0.1 and 0.9, a homozygous call for AF higher or equal than 0.9 and the genotype was set as 'missing' if the AF was lower than 0.1. Genotypes supported by less than 4 supporting reads were also set as 'missing'. In order to prevent sample-specific genotyping errors due to the accumulation of artefactual clipped alignments around the insertion breakpoints, heterozygous and homozygous alternative genotypes were also set as 'missing' if they were supported by at least 5-fold more clipped-reads than the median among all the analyzed samples. A single multi-sample Variant Call Format[122] (VCF) v4.2 file, containing genotypes for source L1s across the complete set of normal samples, was produced as output.

Source L1s contained in the human reference genome sequence were genotyped through a different strategy. For each source L1, we counted the number of reads supporting the deletion of the element (alternative allele) and the number of reads supporting its presence (reference allele). The AF and genotype assignment for each source L1 was computed as described for non-reference insertions.

MT.3 Analysis of somatic retrotransposition

Enrichment and depletion in the rate of somatic retrotransposition across tumour types

For each tumour type with a minimum sample size of 15, we assessed whether it was enriched or depleted in the number of retrotranspositions relative to the rates observed at pan-cancer level using zero-inflated negative binomial regression, as implemented in the zeroinfl function of the pscl R package. This type of model takes into account the excess of zeros and the overdispersion that is present in this dataset. The MEI counts per sample were regressed on a binary factor that expressed whether they belonged to that particular type of cancer or to any other cancer type. For each regression, the magnitude and sign of the z-score indicates the effect size and directionality of the association. More specifically, positive z-scores indicate that a higher number of counts was observed in a cancer type compared with the rest (enrichment), whereas negative scores indicate a lower number of counts (depletion). Each z-score has a P-value associated to indicate the level of statistical significance.

Association between the rate of somatic L1 retrotransposition and mutation in tumour suppressor genes

To assess whether the disruption of a particular tumour suppressor gene was associated with an increased level of somatic L1 retrotransposition, we used the whole genome panorama of cancer driver events per sample produced by the "Drivers and Functional Interpretation Working Group"[123]. This panorama includes coding and non-coding SNVs, insertions and deletions, CN alterations, SVs and potentially predisposing germline variants. For each tumour suppressor gene in the COSMIC database[124] with mutational data, we stratified the samples in two groups - tumour suppressor gene mutated and non-mutated. Then, we compared the L1 counts distribution between both groups using a Mann-Whitney U-test to assess for significance. P-values were corrected for multiple testing using Benjamini-Hochberg. Adjusted P-values lower than 0.05 were considered significant. This analysis was done both at pan-cancer level and per tumour type.

Correlation between the rates of L1 retrotransposition and other SV classes

For each sample, we computed the total number of somatically acquired SVs, the number of L1 insertions and the number of events per each of 5 SV classes: deletions, duplications, translocations, head-to-head inversions and tail-to-tail inversions. Then, we used Spearman's rank test to assess the correlation between the number of somatic L1s and each SV class at both pan-cancer and tumour type level.

Impact of retrotransposition insertions on gene expression

For each somatic L1 insertion occurring within a cancer gene or within the promoter of a non-cancer gene, we compared the expression of the gene in the sample having the insertion, measured in FPKM, with the remaining samples of the same tumour type through a Student's t-test. P-values were corrected for multiple testing using Benjamini-Hochberg. Adjusted P-values lower than 0.1 were considered significant.

Analysis of fusion transcripts involving processed pseudogenes

For each PSD somatic insertion detected in tumour samples with RNA-seq data available, we extracted all the sequencing reads aligning either into the source gene or in the insertion locus. Unmapped reads were also extracted with samtools[125] v1.7. All the collected reads were aligned with BLASTn[126] v.2.7.1 to the isoforms of the source gene registered in RefSeq[127], in addition to the genomic sequence at the integration site (± 5 Kbp relative to the PSD insertion breakpoint). Then, we searched for read-pairs which had one mate aligning on the insertion site sequence while the other into the source gene transcript, as these are indicative of the expression of fusion transcripts. Only read-pairs aligning with > 98% of identity were considered. All fusion transcript candidates were confirmed through manual inspection of sequencing data with the Integrative Genomics Viewer[128] v.2.4.10.

Association between L1 insertion rate and genomic features

The L1 insertion rate was calculated as the total number of somatic L1 insertions, identified across the complete PCAWG cohort, per 1 Mbp window. L1-EN motif density was computed as the number of canonical endonuclease motifs, here defined as TTTT|R (where R is A or G) or Y|AAAA (where Y is C or T), per 1-Mb. Bivariate correlations between L1 insertion rate, endonuclease motif density and replication timing were assessed using Spearman's rank. To study the association between the L1 insertion rate and multiple predictor variables at single-nucleotide resolution we used a statistical framework based on negative binomial regression, as described in detail previously[115]. This method was adapted herein such that originally the regression adjusted for content of trinucleotides in each genomic bin, while in this case we instead adjusted for the content of the L1-EN motif. More specifically, we stratified the genome into four bins (0-3) by the closeness of match to the canonical L1 motif, here defined as TTTT|R (where R is A or G). The bin 0 contains dissimilar DNA motifs, which have 4 or more (out of 5) mismatches (MMs), encompassing 1149.7 Mbp of the genome. Bin 1, 2 and 3 contain genome segments with exactly 3, exactly 2 and at most 1 MM, encompassing 749.4 Mb, 380.2 Mbp and 114.1 Mbp of the GRCh37 assembly, respectively. The closest match of either of the two DNA strands was considered.

Histone mark data (ChIP-Seq for H3K9me3, H3K4me3, H3K36me3, H3K27ac) and DNase hypersensitivity (DHS) data for the regional analyses was collected from Roadmap Epigenomics Consortium by averaging fold-enrichment signal over 8 cell types (E017, E114, E117, E118, E119, E122, E125 and E127) and processed by stratifying into four genomic bins, as described previously[115]. For histone marks and DHS, bin 0 are the areas of the genome with below-baseline signal (Roadmap fold-enrichment compared to input < 1), while bins 1-3 are approximately equal-sized bins covering the remaining parts of the genome with above-average fold-enrichment score. In particular, DHS bins 1-3 encompass 122.8-123.0 Mbp each; for H3K36me3 129.1-136.0 Mbp each; for H3K4me3 43.2-43.7 Mbp each; for H3K27ac 73.6-75.1 Mbp each. RNA-Seq data was also collected from Roadmap and processed as previously[115] by averaging over 8 cell types (E071, E096, E114, E117, E118,

E119, E122, E127): bin 0 consisted of non-expressed genes (FPKM=0) and intergenic DNA that was not explicitly listed as expressed (total 1076.6 Mb), while bin 1 (up to 0.59 FPKM), 2 (up to 5.68 FPKM) and 3 (above 5.68 FPKM) spanned 389.9, 462.1 and 473.8 Mbp of the genome, respectively. Replication time (RT) data was processed similarly as histone marks, but collected from ENCODE and processed by averaging the wavelet-smoothed signal over 8 cell types (HeLa S3, HEP G2, HUVEC, NHEK, BJ, IMR-90, MCF-7 and SK-N-SH) and then dividing into four equal-sized genomic bins (quartiles), where bin 0 is the latest-replicating and bin 3 is the earliest-replicating. Essential genes were determined by CERES score based on CRISPR essentiality screens, ordering by median score across all 342 cell lines tested[129] and then stratifying genes into equal-frequency bins, from less negative to more negative median CERES score (implying commonly essential genes). For the purposes of finding L1 rates in CERES essential genes an additional 1 Kbp flanking the transcript was also considered together with the gene. All enrichment scores shown in plots compare bins 1-3 for a particular feature (RT, histone marks, gene expression, L1 motif) versus bin 0 of the same feature, which therefore always has log enrichment=0 by definition and is not shown on enrichment plots. The regional analyses were restricted to parts of the genome with perfect mappability scores, according to the CRG Alignability track of the UCSC browser[115].

MT.4 Experimental validations

Generation of long-read data for cancer cell lines

Due to the unavailability of tissue specimens for PCAWG tumours, we used for validation purposes a lung cancer cell line (NCI-H2087) previously reported to have high retrotransposition rates[44], and its corresponding matched-normal cell line (NCI-BL2087) derived from blood. Both cell lines were subjected to long-read sequencing with the MinION device from Oxford Nanopore Technologies (ONT) as follows.

Genomic DNA was sheared to 10 Kbp fragments using Covaris g-TUBEs (Covaris), cleaned with 0.4x Ampure XP Beads (Beckman Coulter Inc). After end-repairing and dA-tailing using the NEBNext End Repair/dA-tailing module (NEB), whole genome libraries were constructed with the ONT 1D ligation library prep kit (SQK-LSK108, ONT Ltd). We obtained four and five libraries for NCI-H2087 and NCI-BL2087, respectively. Genomic libraries were loaded on MinION R9.4 flowcells (FLO-MIN106, ONT Ltd), and sequencing runs were controlled using the ONT MinKNOW software v18.01.6. We used the ONT basecaller Albacore v2.0.1 to identify DNA sequences directly from raw data and generate fastq files. Files with quality score values below 7 were excluded at this point. Minion adapter sequences were trimmed using Porechop v0.2.3 (https://github.com/rrwick/Porechop). Then, we used minimap2[130] (v2.10-r764-dirty) to map sequencing reads onto the hs37d5 human reference genome, and the SAM files were converted to BAM format, sorted and indexed with samtools[125] v1.7 for each one of sequencing runs. BAM files were merged, sorted and indexed. After this process, sequencing coverage were 8.2x (NCI-BL2087) and 9.17X (NCI-H2087), and average read size of mapped reads were ~4.5 Kbp (NCI-BL2087) and ~11 Kbp (NCI-H2087).

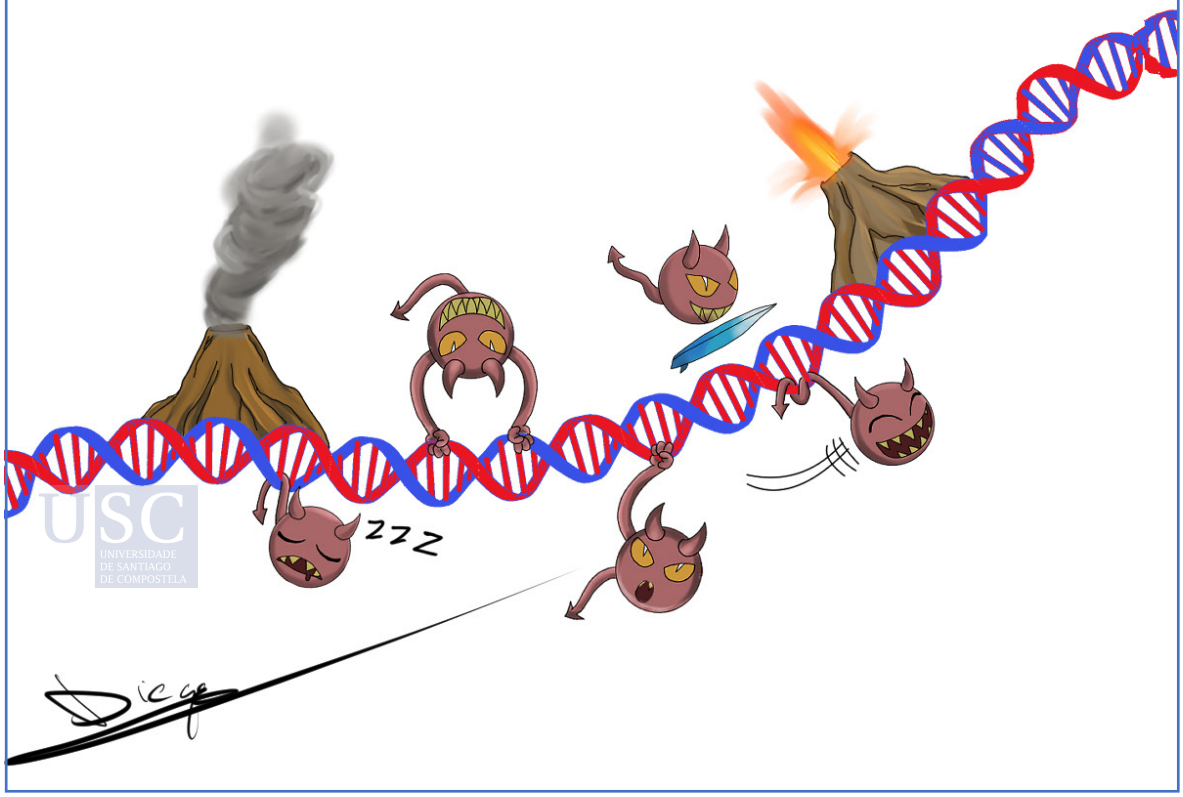
Validation of L1-mediated rearrangements through PCR

Due to the unavailability of tissue specimens for PCAWG tumours, we validated the algorithms for the detection of L1-mediated rearrangements in two cancer cell lines known to have high levels of somatic retrotransposition[44] (NCI-H2009 and NCI-H2087). We designed primers for PCR validation of the 20 somatic L1-mediated rearrangements, mostly deletions, identified in these cell lines. PCR primers were designed with Primer3[131] v0.4.0, to amplify both the insertion breakpoints and the sequence at insertion target site as follows.

For the 5' breakpoint, a forward primer targeted the DNA sequence upstream relative to the insertion point while the reverse targeted the 5' end of the L1 insertion (L). On the other hand, 3' breakpoint amplification relied on a forward primer that was specific for the end of human-specific L1 elements (5'-GGGAGATATACCTAATGCTAGATGACAC-3'103) and a reverse primer targeting the genomic sequence immediately adjacent to the 3' insertion breakpoint (R). In the case of orphan and partnered transduction, reverse and forward primers were designed for the ends of the non-repetitive transduced insert. A third pair of primers flanking the insertion event were designed to amplify complete insertion and reference alleles (T).

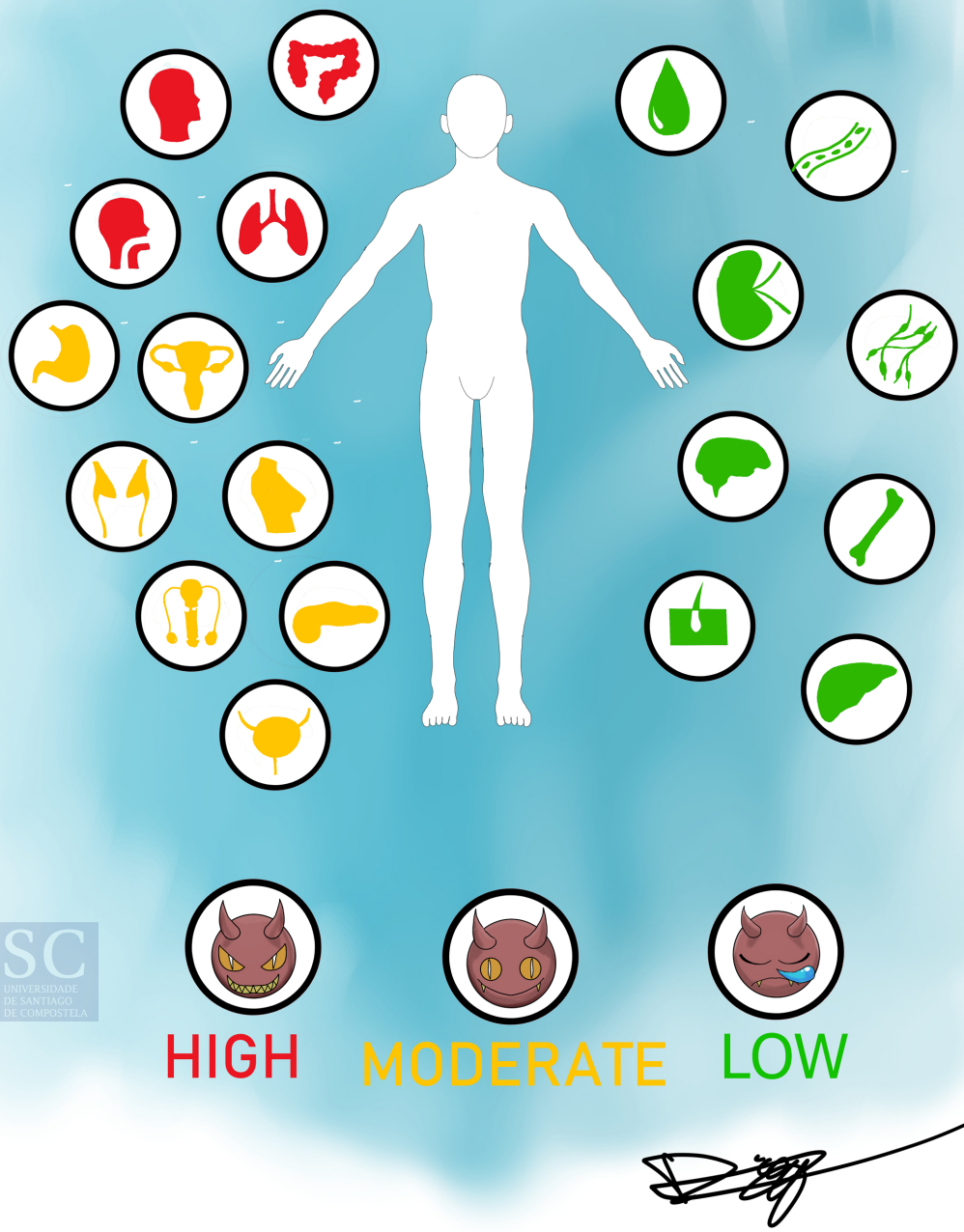
Each PCR mixture contained 10ng of DNA, 5pmol of each primer, 5U Taq DNA polymerase (Sigma-Aldrich, catalog number D1806) with 1x Buffer containing MgCl2, 0.2mM of each dNTPs, and water to a final volume of 25µl. PCR conditions were as follows: initial denaturation at 95ºC for 7 minutes; then 30-35 cycles of 95ºC for 30 seconds, 60ºC for 30 seconds, 72ºC for 45 seconds; and a final extension of 72ºC for 7 minutes. In some cases, when amplification failed, we used Platinum Taq High-fidelity, with a 94ºC denaturation and a 68ºC extension. PCR amplicons were sequenced with single-molecule sequencing using a MinION from ONT. Amplicons were pooled and total DNA was cleaned with 0.4x AMPure XP Beads (Beckman Coulter Inc). After end-repairing and dA-tailing using the NEBNext End Repair/dA-tailing module (NEB), the sequencing library was constructed with the ONT 1D ligation library prep kit (SQK-LSK108, ONT Ltd) and loaded on a MinION R9.4 flow cell (FLO-MIN106, ONT Ltd). Mapping to the human reference genome was performed as described above, with minor modifications.

RESULTS



CHAPTER 1:

“Pan-cancer landscape of somatic retrotransposition”



C1.1 Pan-cancer analysis of somatic retrotransposition

We used TraFiC-mem to detect somatic mobile element insertions in the 2,954 cancer whole genomes included in the PCAWG dataset (see sections MT.2 and C4.1). A total of 19,166 somatic retrotranspositions insertions were identified across all cancer genomes (Figure 1a). Consistently with previous reports[41–44], L1 insertions overwhelmingly dominated the retrotransposition landscape in the PCAWG cohort, while only 130 Alu and 23 SVA somatic insertions were found. Although both Alu and SVA repeats rely on L1-encoded proteins for their mobilization, the number of somatic Alu correlated weakly with the number of L1s (Spearman's rho = 0.21, $P < 0.05$), while no correlation was found for SVA (Spearman's rho = 0.07, $P < 0.05$).

The majority (97%; 10,177/10,544) of L1 integrations belonged to the Ta subfamily (Figure S2), while a minority of insertions (n=367) had diagnostic nucleotides indicative for pre-Ta (i.e., the oldest subfamily of human-specific L1s). Similar patterns were observed for Alu, with young AluYa5, AluYb8 and AluYb9 subfamilies dominating the Alu retrotransposition landscape in cancer (Figure S2). Consistently, 74% (17/23) of all SVA insertions belonged to the human-specific SVA_E and SVA_F lineages. These analyses indicate that somatic retrotransposition in cancer is dominated by evolutionary young L1, Alu and SVA subfamilies (Figure S2), which is consistent with previous reports for polymorphic MEI[114,132].

Somatic L1 insertions frequently (46%; 8,558/18,739) displayed severe truncation of their 5' ends, which ranged from 5,895 to 30 bp in size (median = 341 bp), potentially as a consequence of the poor processivity of L1 retrotranscriptase[133]. Inversions of internal L1 sequences due to twin-priming[134] were also common (38%; 7087/18,739), with only 311 insertions being full-length. On the contrary, most (73%; 73/100) somatic Alu insertions in the PCAWG dataset were full-length, which is consistent with the insertion patterns previously described for germline polymorphisms[114]. Due to the intrinsic complexity of SVA sequences, which contain an internal domain composed by a variable number of tandem repeats, their length could not be reliably estimated from short-read data. A majority (92%; 17,280/18,892) of L1, Alu and SVA insertions had poly(A) tails at their 3' ends, indicative of passage through a mRNA intermediate prior integration. L1, Alu and SVA integration was typically associated with small target site duplications (median = 14 bp) and deletions (median = -5 bp). Overall, truncation, inversion, poly(A) tails and target site alterations are hallmarks of TPRT[58], the canonical mechanism for retrotransposon integration. While most (85%; 15,956/18,739) L1 insertions had features consistent with TPRT, a minority (2%; 352/18,739) were heavily truncated at both ends and lacked poly(A) tails. This integration pattern resembles an alternative endonuclease independent insertion mechanism used by L1s to integrate into preexisting double strand breaks, which has been described in cell lines deficient in NHEJ repair[135].

We further searched for somatically acquired processed pseudogenes (PSD), identifying 274 insertions in 105 cancer samples, a number that exceeds by far previous estimation for

PSDs in cancer genomes[136]. Although few events were typically detected per sample, certain cancer genomes had particularly high numbers of PSD insertions. For example, a pancreatic adenocarcinoma tumour (SA533710) harbored ~26% (70/274) of all PSD identified in the PCAWG cohort (Figure 1b). PSDs originated from a total of 234 template genes, with 16 genes leading to recurrent PSD formation, including *LYZ* and *CEACAM6*, which had 6 and 5 somatically acquired retrocopies, respectively. The majority (91%; 248/274) of PSD events displayed TPRT hallmarks, including poly(A) tails, target site alterations and inversion or truncation of their 5' ends, which confirmed the reverse transcription and integration of cellular mRNAs by L1-encoded enzymatic machinery.

C1.2 Somatic retrotransposition activities across tumour types

Somatic retrotransposition is a common mutational process in cancer, with 35% (1,046/2,954) of all cancer genomes in the PCAWG dataset having at least one somatic mobile element insertion. The total number of retrotranspositions identified in a tumour was heterogeneous, ranging from a single insertion up to 638 events in SA494351, a remarkable head-and-neck cancer (Figure 1b). Samples with high retrotransposition rates were relatively frequent in colorectal, head-and-neck, lung squamous and esophageal cancers (Figure 1c, Figure S3), where 3-27% of tumours had more than 100 somatic insertions. As consequence, these four tumour types alone accounted for more than 70% (13,373/19,166) of all somatic events found in PCAWG tumours, while they only represented 9% (266/2,954) of PCAWG samples (Figure 1d). Indeed, somatic retrotransposition was the second most frequent type of SV in esophageal adenocarcinoma, and the second in head-and-neck and colorectal adenocarcinomas (Figure 1e), what highlights the relevance of retrotransposition as a predominant SV class in these tumour types

In general, apart from the tumour types with high retrotransposition rates described above, the majority of adenocarcinomas, such as those arising from the stomach, pancreas, breast, uterus, ovary, cervix and prostate, had moderate levels of retrotransposition. Meanwhile, skin, bone, brain and blood cancers had low levels of somatic retrotransposition, with no insertion identified for 93% (764/826) of those tumours. Although these findings are consistent with previous cancer genome surveys[41–44], it is particularly intriguing to observe low retrotransposition rates at brain tumours, as previous single-cell based studies have reported high levels of somatic L1 mosaicisms in healthy brain tissues[137]. A potential explanation for these discrepancies is that brain somatic mosaicism has been reported in neurons, while brain tumours typically derive from either cerebellar stem cells (medulloblastoma), or glial cells, including astrocytes (pilocytic astrocytoma) and oligodendrocytes (oligodendroglioma), where mobile elements may be subjected to cell type specific repression. In addition, within the framework of a collaborative work with Dr Francesco Maura (Sanger Institute), we explored somatic retrotransposition in a tumour cohort of 67 whole genomes sequences from 30 patients with multiple myeloma[138], and five nodal peripheral T-cell lymphoma tumours[139]. No somatic insertions were detected in these blood tumours, suggesting the absence of somatic retrotransposition during the course of these blood malignancies.

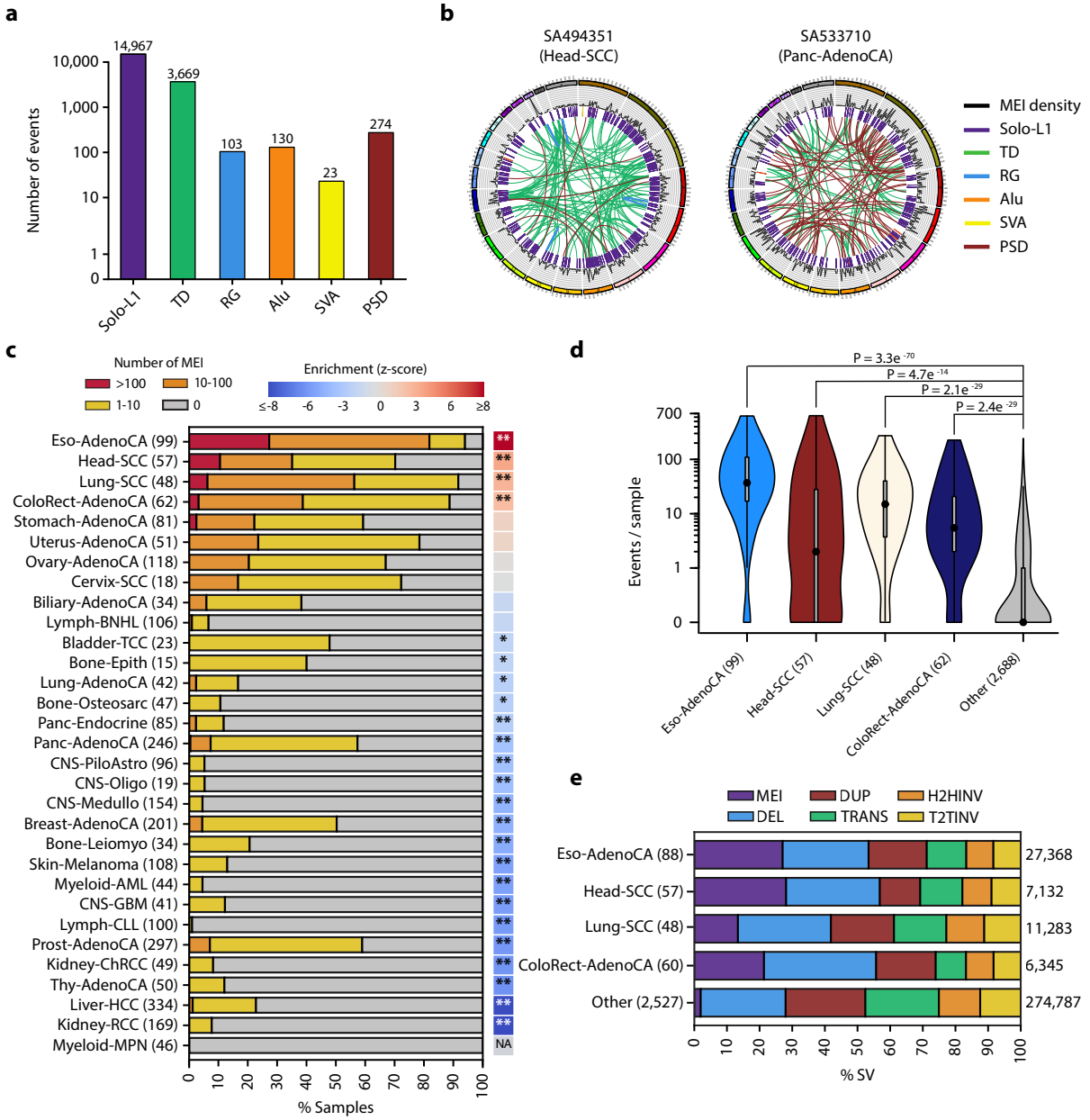


Figure 1. Landscape of somatic retrotransposition across human cancers. **(a)** Number of somatic retrotransposition events identified across six categories: solo-L1, L1-mediated transductions (TD), L1-mediated rearrangements (RG), Alu, SVA and pseudogenes (PSD). **(b)** Circos plots showing two cancer genomes with extremely high numbers of somatic retrotranspositions (left) and PSDs (right). **(c)** Proportion of tumour samples with >100 (red), 10–100 (orange), 1–10 (yellow) and 0 (gray) somatic retrotranspositions. The number of samples analyzed for each tumour type is shown in parentheses. Only cancer types with sample size n ≥ 15 are included. Retrotransposition enrichment or depletion for each tumour type together with the level of significance (zero-inflated negative binomial regression) is shown. *P < 0.05, **P < 0.01. NA, not applicable. **(d)** Retrotransposition events per sample across the four tumour types enriched in somatic retrotranspositions. Remaining tumours grouped into 'Other'. Number of samples from each group shown in parentheses; point, median; box, 25th to 75th percentiles (interquartile range); whiskers, data within 1.5 the interquartile range. P-values indicate significance from a two-tailed Mann–Whitney U-test. **(e)** Contribution of each of six SV classes to the total number of SVs detected in the four tumour types enriched in somatic retrotransposition. Remaining tumours grouped into 'Other'. SVs classes: mobile element insertions (MEI), deletions (DEL), duplications (DUP), translocations (TRANS), head-to-head inversions (H2HINV) and tail-to-tail inversions (T2TINV). The total number of SVs per cancer type is indicated on the right side of the panel.

Another remarkable observation is that somatic L1 retrotransposition rates were markedly heterogeneous across patients within a cancer type. For example, 30% (17/57) of head-and-neck tumours had no somatic insertions, while 11% (6/57) had more than 100 events. Similarly, the number of somatic retrotranspositions ranged between 0 (41% of the samples) and 141 in stomach adenocarcinomas, with an average of 11 events per sample. In order to investigate the potential factors driving these differences, we searched for associations between increased retrotransposition rates and mutations affecting genes catalogued as cancer drivers[9] (see section MT.3). This analysis revealed a pan-cancer association between *TP53* mutation and increased levels of somatic L1 retrotransposition (Mann–Whitney U test, $P < 0.05$, Figure 2a).

At tumour type level, we observed significant differences between patients with and without *TP53* mutations for head-and-neck, biliary, stomach, pancreatic and breast cancers, in which mutated tumours had a 2-15 fold enrichment on the number of L1 retrotranspositions (Figure 2b). Furthermore, *TP53* driver mutations were frequent (52%; 535/1035) among adenocarcinomas, which are characterized by high L1 retrotransposition rates, while they were only found in 18% (62/341) of all blood, brain and skin cancers. Overall, these findings are consistent with previous lines of evidence suggesting that *TP53* restrains mobile elements at various levels[140,141], including cell cycle arrest, senescence and apoptosis in response to retrotransposon induced DNA damage and transposon silencing. Nonetheless, 27% (379/1401) of *TP53* wild-type tumours had at least one somatic L1 insertion, including 4 with more than 100 events, and 43% (357/826) of *TP53* mutated tumours had no insertion events. Both observations suggest that additional factors, such as tissue type specific epigenetic alterations or environmental exposures, are likely to contribute to the activation of L1 sequences.

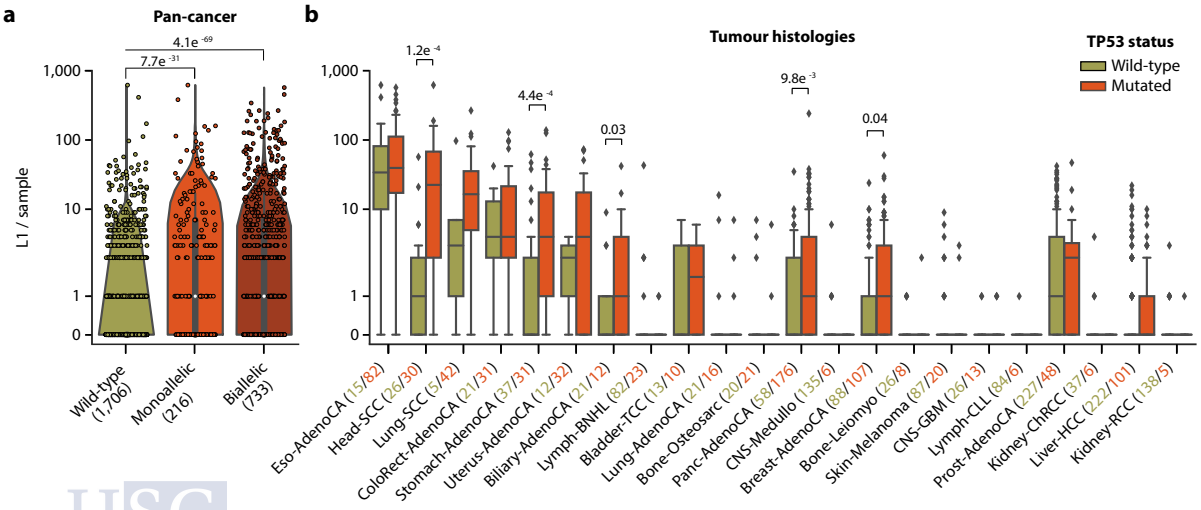


Figure 2. *TP53* mutation is associated with high rates of L1 retrotransposition. (a) Distribution of L1 counts for three groups of samples according to their *TP53* mutational status: wild-type, monoallelic and biallelic driver mutations. Each data point corresponds to one tumour sample. Groups are compared though Mann–Whitney U. **(b)** Distribution of L1 counts across tumour types with samples grouped in two categories: *TP53* wild-type and *TP53*-mutated (monoallelic or biallelic). Groups are compared though Mann–Whitney U.

We also observed moderate to strong correlation between the burden of L1 somatic retrotranspositions and other SV classes across a wide variety of tumour types (Figure 3). More precisely, the number of L1 insertions correlated with most rearrangement classes (Spearman´s, $P < 0.05$) for 39% (12/31) of tumour histologies, including esophageal and head-and-neck cancers. In contrast, the correlation was only significant for translocations (Spearman's rho = 0.298, P = 0.02) in colorectal cancers. These correlations may be driven by increased levels of genomic instability in *TP53* mutated tumours, as *TP53* inactivation correlated both with increased levels of somatic L1 retrotranspositions and SVs in PCAWG tumours (Figure 2; Figure S4).

In collaboration with Stratton´s group (Sanger Institute), we also investigated if somatic retrotransposition could trigger APOBEC mutagenesis[142], which is a frequent mutational source in multiple cancer types, including esophageal, head-and-neck and lung tumours. As APOBEC cytidine deaminases restricts retrotransposons through DNA-editing-dependent and

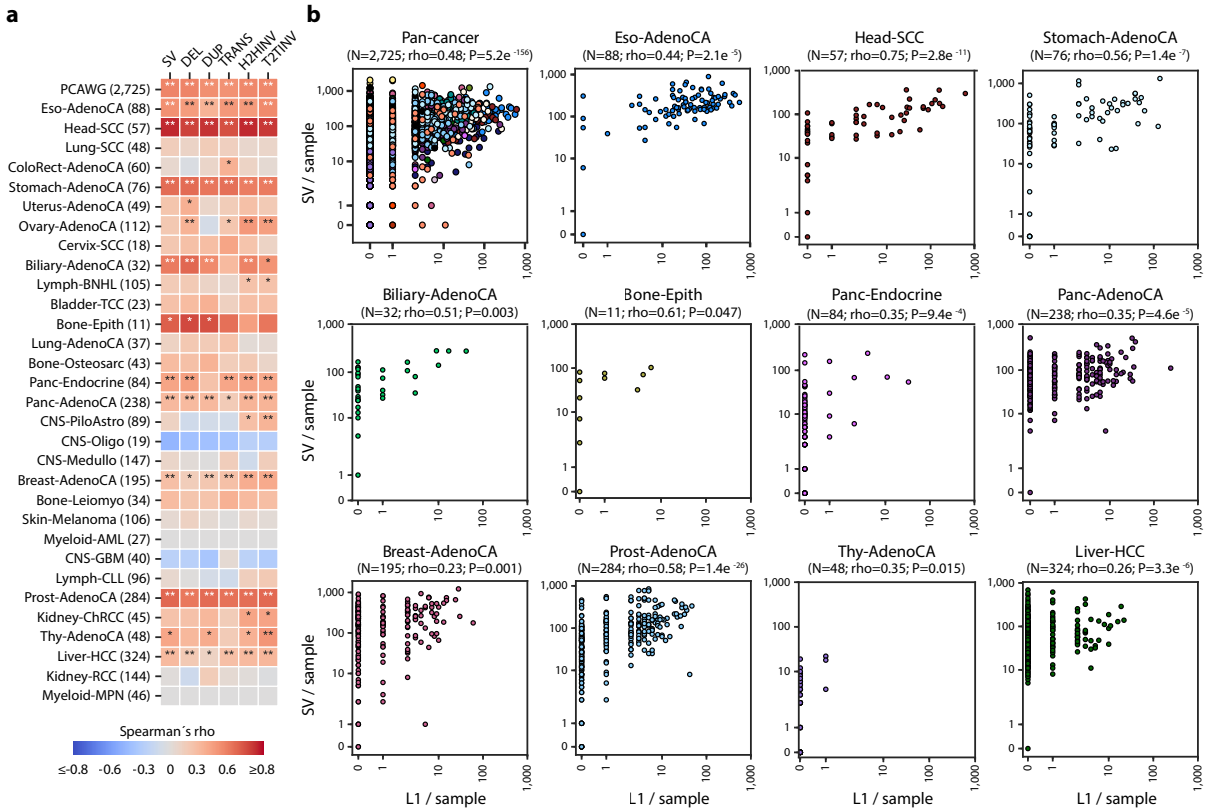


Figure 3. Correlation between L1 retrotransposition and structural variation burden. (a) Heatmap showing the correlation between the number of L1 events per sample, the total number of SVs and the number of SVs for each of 5 different SV classes: deletions (DEL), duplications (DUP), translocations (TRANS), head-2-head inversions (H2HINV) and tail-2-tail inversions (T2TINV). Correlations assessed both at pan-cancer and tumour type levels. Spearman's correlation coefficients are shown in a blue (negative) to a red (positive) coloured gradient. P-values lower than 0.05 and 0.01 are represented as single and double asterisks, respectively. **(b)** Correlations between the number of L1 events and the total number of SVs per sample at both pan-cancer and tumour type levels. Only tumour types with significant correlations are displayed. Each dot represents an individual sample and is coloured according to tumour histology. Both axes are displayed on a symlog scale.

independent mechanisms[143], the activation of endogenous L1 retrotransposons in cancer may trigger APOBEC response, leading to APOBEC-mediated mutagenesis. In this study, single-cell derived subclones for 28 cell lines, spanning a wide range of cancer types, were cultured for extended periods and subjected to WGS at multiple time points. Cells under cell culture exhibited substantial fluctuations in the accumulation of APOBEC-mediated mutations over time, with episodic bursts of APOBEC activity. Interestingly, the number of *in vitro*-acquired L1 insertions significantly correlated with the burden of APOBEC-mediated mutational signature 13 ($p < 0.001$; Bonferroni-adjusted). This association was particularly pronounced in breast and lung adenocarcinomas, while no correlation was found for lymphomas and colorectal cancers, suggesting that additional currently unknown factors may trigger APOBEC response in the absence of retrotransposition. As mutational signatures annotation was available for the primary tumours included in PCAWG, we extended these analyses to the PCAWG dataset. However, there was no evidence of a correlation between somatic L1 retrotransposition and APOBEC-mediated signature 13 ($p = 1.0$). Although the cell line data suggest a possible relationship between APOBEC mutagenesis and retrotransposition activities, more direct experimental testing is required to establish this. The variable strengths of the effects observed across the analyzed cell lines and the negative results in the PCAWG dataset suggest that other factors may be involved.

C1.3 Functional impact of somatic retrotransposition

We also investigated the functional impact of the large collection of somatically acquired retrotranspositions detected in the PCAWG dataset (see section MT.3). L1 integrations frequently (43%; 7,979/18,636) occurred within gene boundaries, including promoters and introns, with 66 events targeting genes catalogued as cancer drivers[144]. A total of 1,330 genes had recurrent L1 retrotransposition insertions, with *LRP1B* (n=49), *DLG2* (n=41) and *EYS* (n=36) ranking as the most frequently mutated genes. However, these are long-sized genes located at common fragile sites, which are known to be subjected to particularly high mutational rates[145,146]. Hence, after correcting for replication timing via binomial regression no gene was significantly recurrently mutated by L1s.

The absence of significance does not exclude the possibility that a small fraction of somatic events may affect gene function. Therefore, we used the RNA-seq data available for 35% (1,043/2,954) of PCAWG tumours to search for genes differentially expressed after an L1 insertion on their promoter. Among the 83 L1 insertions targeting promoters, four genes were overexpressed (Student's t-test, $q < 0.1$). This includes a 6-fold increase in gene expression of the *ABL2* oncogene in a head-and-neck squamous carcinoma sample, SA494343, relative to the remaining head-and-neck tumours without the L1 insertion (Figure 4a,b). We further extended the transcriptomic analysis to insertions affecting any component of cancer genes (i.e., introns or exons), which revealed the overexpression of the tumour suppressor gene *RB1* in a bladder tumour (Student's t-test, $q < 0.10$) (Figure 4c). Additional analysis indicated that

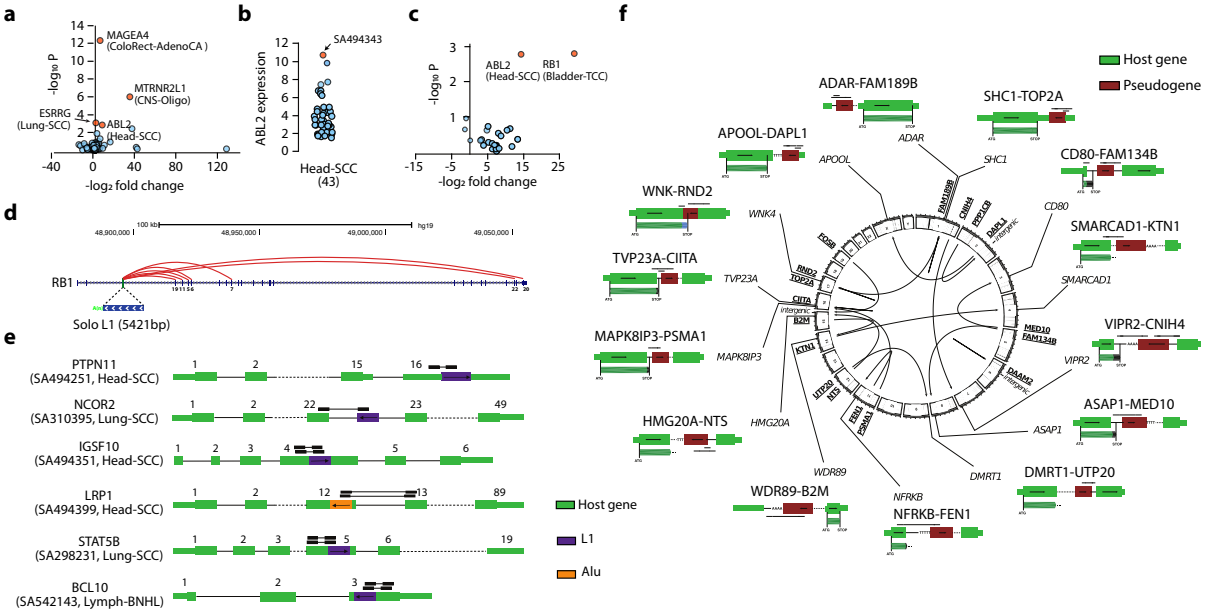


Figure 4. Gene expression alterations associated with somatic retrotransposition. (a) Gene expression fold-change (x axis) and significance (y-axis) for genes with a somatic L1 insertion at their promoters compared to the remaining tumours from the same cancer type (Student's t-test, q < 0.1). **(b)** Expression of the *ABL2* oncogene in head-and-neck squamous carcinoma (Head-SCC) samples. Sample with L1 insertion at *ABL2* coloured in red, while others in blue. **(c)** Differential gene expression for cancer genes affected by L1 events (same format as in a). **(d)** Aberrant splicing at *RB1* due to the exonization of an intronic L1. Chimeric splice junctions connecting exons with the L1 insert are represented as red lines with the number of supporting read-pairs underneath. **(e)** Examples of aberrant splicing due to somatic L1 and Alu exonization. Split and discordant read-pairs supporting a fusion transcript are shown above each gene model. **(f)** Processed pseudogene (PSD) insertions leading to fusion transcript expression. Arcs with arrows within the circos indicate the PSD retrotransposition event, connecting the source PSD (underlined and bold) with the corresponding integration region. The predicted fusion transcripts and their putative coding potential are displayed for every event, including start and stop codons. Uncertain termination sites represented as dots.

the observed change in gene expression is likely caused by an L1 integration at the second intron of *RB1*. More precisely, we identified discordant read-pairs supporting the aberrant splicing between the L1 insert and 7 different *RB1* exons, which revealed the existence of multiple *L1-RB1* isoforms generated via L1 exonization (Figure 4d). Aberrant splicing involving exons 3, 25 and 27 (ENST00000650461.1_1) were supported by more than 20 read-pairs, suggesting that these fusion transcripts are highly expressed. It is plausible that the wild-type allele may increase its expression as a compensatory mechanism in response to a dysfunctional *RB1* copy, resulting in the observed *RB1* overexpression. It is also possible that the fusion transcripts or protein products derived from the altered allele may compete with the wild-type, representing a potentially novel mechanism of tumour suppressor gene disruption via L1 exonization. Further molecular biology essays will be necessary to investigate these hypotheses. We detected six additional instances of exonization (Figure 4e), including 5 insertions within exons and one additional L1 insertion in the intron of the gene *NCOR2*.

As 39% (106/274) of somatic PSDs in the PCAWG dataset are inserted within gene bodies, we also searched for evidence of PSD expression. The analysis of the RNA-seq data available

for 144 samples containing at least one PSDs revealed discordant read-pairs supporting the expression of 17 PSD (Figure 4f). While in three instances the PSD was expressed without involving the host gene, the majority of the events (82%, 14/17) were fusion transcripts. Four fusions involved PSDs inserted at UTR sequences, nine at introns and one within one exon. The predicted sequences for the fusion transcripts typically had their open reading frames disrupted, either due to intron inclusion or because the PSD is integrated in antisense orientation (Figure 4f). Overall, these data demonstrate that retrotransposition of L1 elements and PSD insertions can occasionally impact gene function through diverse mechanisms in the cancer genome, including expression and splicing alterations, with potential oncogenic consequences.

C1.4 Multiple genomic features shape L1 insertion distribution

The 18,739 somatically acquired L1 insertions detected in the PCAWG dataset provided an excellent opportunity to investigate the patterns of L1 insertion distribution across the cancer genome. The genome-wide distribution of somatic L1 retrotranspositions was markedly heterogeneous along the cancer genome (Figure 5a). As L1 integration relies on a self-encoded endonuclease that targets a degenerate consensus target sequence (5-TTTT/A-3), we first investigated whether the distribution of somatic L1s across the cancer genome could be determined by the occurrence of L1-EN target motifs. We used a statistical approach based on negative binomial regression to deconvolute the influence of multiple genomic features[115], including replication timing, diverse epigenetic marks, chromatin state and gene expression (see section MT.3). This analysis revealed a 244-fold enrichment of L1 insertions in sequences closely resembling L1-EN motifs (Figure 5b, Figure S5a). As replication timing is known to have a major impact on the local mutational rates in cancer genomes[145], we investigated its association with somatic L1 retrotransposition. Adjusting for the potentially confounding effect of L1-EN motifs, we observed a strong association between somatic L1 retrotransposition and DNA replication time, with the latest-replicating quarter of the genome being 8.9-fold enriched in L1s (95% confidence interval, 8.25–9.71) with respect to the earliest-replicating quarter (Figure 5b-c, Figure S5b).

This data resembles the patterns described for L1 polymorphisms[31], which have been traditionally explained by the effect of purifying selection upon L1 sequences at gene-rich early replicating regions, as retrotransposons can induce genomic rearrangements via ectopic recombination[147]. Therefore, the association between L1 retrotransposition and late replication timing may be a consequence of the extraordinary selective pressures operating during tumour development. However, as described in the previous section, we only observed a limited number of insertions with clear functional consequences and, in addition, there was not a significant association between gene essentiality and L1 rates (1.03-fold decrease in essential genes) (Figure S5c), suggesting that only a minor fraction of the somatic insertions is under negative selection. Furthermore, cancer genome analysis indicates that tumours

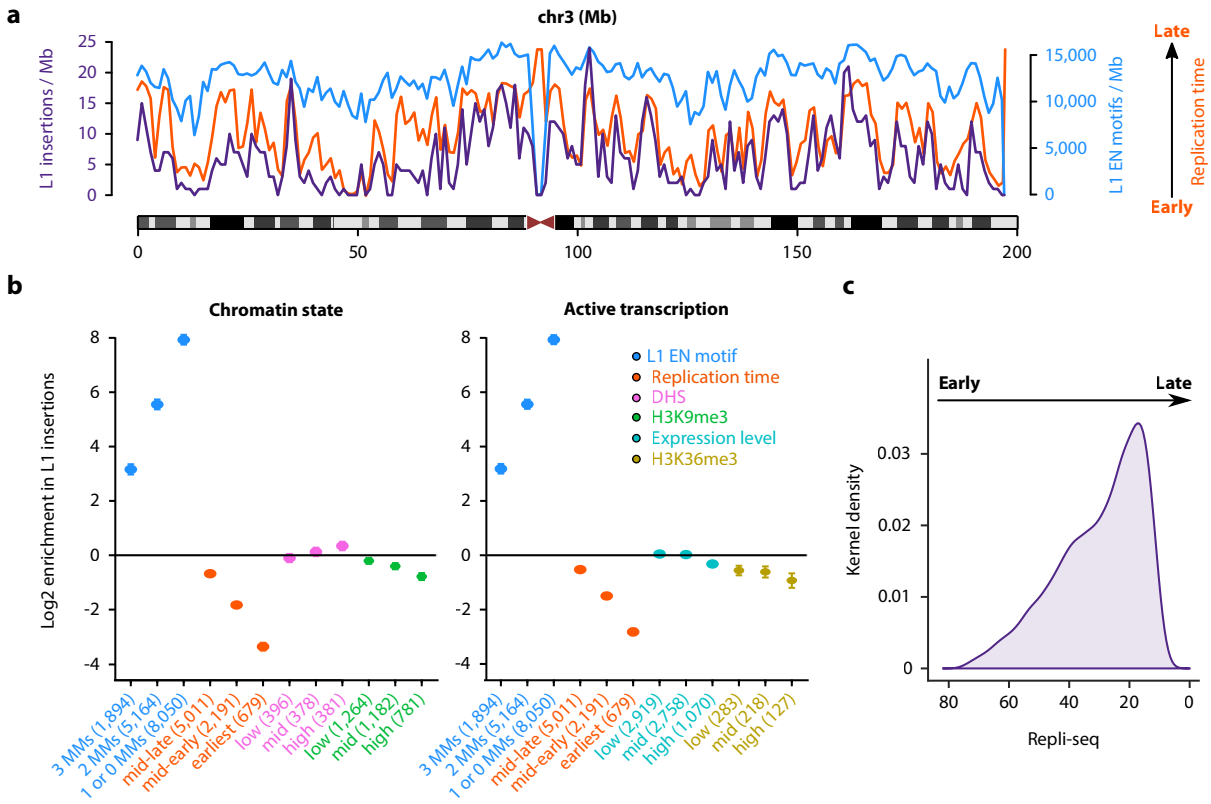


Figure 5. Distribution of L1 somatic insertions across the cancer genome and its association with genome organization features. (a) L1 insertion rate (purple), L1-EN motif density (blue) and replication timing (orange) per 1 Mbp window. For illustrative purposes, only chromosome 3 is shown. **(b)** Association between L1 insertion rate and multiple predictor variables at single-nucleotide resolution. Enrichment scores (thick dots) are adjusted for multiple covariates and compare the L1 insertion rate in bins 1–3 for a particular genomic feature (L1-EN motif, replication timing, open chromatin, histone marks and expression level) versus bin 0 of the same feature, which therefore always has log-transformed enrichment = 0 by definition and is not shown. The error bars represent 95% confidence intervals. The number of observations per bin is provided in parentheses. MMs, the number of mismatches with respect to the consensus L1-EN motif (see section MT.3). Heterochromatic regions and transcription elongation are defined based on H3K9me3 and H3K36me3 histone marks. Accessible chromatin is measured through DNase hypersensitivity. **(c)** L1 insertion density, using kernel density estimate (KDE), along the replication timing spectrum. DNA replication timing is expressed on a scale from 80 (early) to 0 (late).

generally evolve via the positive selection of mutations that increase cell fitness[148], with negative selection having a much limited effect, which suggests that other factors may explain the associations we describe above.

An alternative hypothesis is that the observed association between somatic L1 retrotransposition and replication timing is the consequence of an insertional bias towards late replicating DNA. A recent study characterized the abundance and cellular location of L1-encoded proteins, ORF1p and ORF2p, during the cell cycle, finding that L1 retrotransposition has a strong cell cycle bias and preferentially occurs during S phase[149]. Our results are in agreement with these findings, suggesting that L1 retrotransposition may peak in the later stages of nuclear DNA synthesis, leading to the enrichment of L1 insertions in late replicating DNA. Recent *in vitro* retrotransposition assays have generated a large collection of engineered L1 insertions in human

cultured cell lines[150,151]. Remarkably, *de novo* L1 insertions were enriched into late-replicating DNA in PA-1, NPC and Hela cell lines, while they were depleted for the hESC cell line. This may suggest that the directionality of the association between L1 retrotransposition rates and replication time may be cell type dependent. Based on these findings, we investigated if there were also differences between tumour histologies by repeating the analysis for those cancer types with at least 100 L1 insertions detected in total. We found no significant differences, with L1 insertions being consistently enriched in late replicating DNA for all the tumour types assessed (Figure S5d). In addition, we did not observe differences between samples with at least 100 L1 insertion events (Figure S5e), supporting a prevalent association between late replication and retrotransposition in cancer. Further research will be required to elucidate the nature of the discrepancies between the patterns observed *in vitro* and in cancer genomes.

As somatic L1 retrotransposition has been previously reported to be enriched into heterochromatic regions[44], we also examined L1 rates in closed heterochromatic regions by analysing K9-trimethylated histone H3152 (H3K9me3). When adjusting for the confounding effects of L1-EN motif content and replication time, we found that somatic insertions are depleted in heterochromatin (1.72-fold, 95% confidence interval, 1.57–1.99; Figure 5b) and enriched in open chromatin (1.27-fold in the highest tertile relative to the lowest; 95% confidence interval, 1.14–1.41; Figure 5b). This discrepancy with previous analyses[44] is likely the consequence of the confounding effect between heterochromatin and late-replicating DNA regions, which was not previously addressed. We also found a negative association between the rate of L1 insertions and genomic features related with active transcription of chromatin, characterized by fewer L1 events at active promoters (1.63-fold; Figure S5c), a slight but significant reduction in L1 rates in highly expressed genes (1.25-fold lower; 95% confidence interval, 1.16– 1.34; Figure 5b) and a depletion at H3K36me3 (1.90-fold reduction in the highest tertile; 95% confidence interval, 1.59–2.29; Figure 5b), a mark of actively transcribed regions deposited in the body and at the 3′ end of active genes[152].

Collectively, our data suggest that the genomic features with a major influence in the distribution of somatically acquired L1 insertions in cancer are L1-EN motifs and replication timing, with other genomic variables, such as chromatin state and epigenetic features related with transcription, having a moderate effect.

C1.5 Contributors

José M.C. Tubio contributed to the processing of PCAWG whole genomes with TraFiC-mem. Eva G. Alvarez and Adrián Baez-Ortega participated in the analysis of somatic retrotransposition, generating several figures. Young Seok performed the differential gene expression analysis. Ana Dueso-Barroso and David Torrents identified PSD fusion transcripts and developed the figure caption 4f. Fran Supek had major contributions into the development of the section pertaining to the association between L1 insertion and genome organization features. Iñigo Martincorena contributed to the analysis of genes recurrently affected by L1 integrations, the

association between *TP53* inactivation and increased retrotransposition, and provided valuable input on the associations with replication timing and heterochromatin. Mia Petljak performed the correlations between APOBEC and L1 retrotransposition.

C1.6 Publications

The core of this chapter's content has been published as an Article in Nature Genetics [1]. Contents were completely rearranged, rewritten and extended to include additional analysis and details that were not covered in the original manuscript due to space constraints. All figures included derive either from the main or supplementary figures of the aforementioned publication. Analysis pertaining to multiple myeloma and nodal peripheral T-cell lymphoma not otherwise specified result from collaborations led by Francesco Maura at Sanger Institute [2-3]. The described associations between APOBEC mutational signatures were investigated through a collaboration with Mia Petjak from the Sanger Institute [4]. The complete list of authors and their affiliations is provided in the appendix.

[1] Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat Genet 52, 306–319 (2020). DOI: https://doi.org/10.1038/s41588-019-0562-0, ISSN: 1061-4036

[2] Maura, F., Bolli, N., Angelopoulos, N. *et al.* Genomic landscape and chronological reconstruction of driver events in multiple myeloma. Nat Commun 10, 3835 (2019). DOI: https://doi.org/10.1038/s41467-019-11680-1, ISSN: 2041-1723

[3] Maura, F., Dodero, A., Carniti, C. *et al.* CDKN2A deletion is a frequent event associated with poor outcome in patients with peripheral T-cell lymphoma not otherwise specified (PTCL-NOS). Haematologica 106, 11 (2021). DOI: https://doi.org/10.3324/haematol.2020.262659, ISSN: 0390-6078

[4] Petljak, M., Alexandrov, L. B., Brammeld, J. S. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell 176, 1282–1294. e20 (2019). DOI: https://doi.org/10.1016/j.cell.2019.02.012, ISSN: 0092-8674

CHAPTER 2:

“Hot L1 elements are drivers of somatic retrotransposition”

C2.1 Patterns of L1 activity in cancer

Using our algorithm TraFiC-mem, we detected a total of 3,696 somatic L1-mediated transductions across the 2,954 tumours included in the PCAWG dataset (see sections C4.1 and MT.2). Orphan transductions, in which a downstream sequence of an active L1 is retrotransposed without the cognate L1, represented 64% (2367/3,696) of all the events, with the remaining being partnered transductions (i.e., with a companion L1). The median size of transduced sequences was 333 bp, although long transductions reaching sizes up to 1.5 Kbp were occasionally detected. Consistently with the patterns described above for solo-L1 insertions, transductions were particularly abundant in esophageal, head-and-neck, lung and colon cancers; with these four tumour types alone encompassing 70% (2451/3541) of all transductions.

Since transductions are defined by the retrotransposition of a non-repetitive genomic sequence, they can be used to unambiguously identify the L1 loci whence they derive. We found that 114 germline source L1s were responsible for all transductions identified in the PCAWG cohort (Figure 6a). While 60 source L1s were previously reported to be active[44,59,73,114], 54 elements are indeed novel active copies, expanding the catalogue of active L1s in humans. We genotyped source L1 elements across the 2,641 matched-normal genomes (see section MT.2), finding that 22 are fixed alleles in the human population, with the remaining 92 being polymorphic. A majority (71%, 65/92) of these polymorphic source L1s are common variants with minor allele frequencies (MAF) over 5%, while 27 represent rare (MAF = 1-5%) and very rare (MAF < 1%) polymorphisms.

In theory, when a FL-L1 retrotransposition occurs, it takes with it all the machinery required to catalyse further retrotranspositions. By using transductions as a marker of retrotransposition competency, we searched for examples of somatically acquired L1 retrotranspositions that led to further dissemination from the new insertion site. Notably, this analysis revealed 198 somatic retrotranspositions derived 100 FL-L1 loci that were themselves somatically acquired insertions. For example, in a remarkable head-and-neck tumour sample, SA197656, one somatic FL-L1 integration at 4p16.1 mediated 18 transductions, with the next most active element being a germline source element at 22q12.1, which accounted for 15 transductions. Hence, although contributing to a minority of the total number of transductions we identified in PCAWG (5%; 198/3,696), somatically acquired FL-L1s are a relevant source of further L1 retrotranspositions, on occasions. We observed considerable variability in activity among the source L1s, with a reduced set of 16 highly active (i.e., hot) L1 loci being responsible for 67% (2,440/3,669) of all detected transductions (Figure 6b). This is consistent with previous studies in naturally-occurring human tumours[44] and *in vitro* assays[59], further supporting the idea that most retrotransposition mobilizations in the germline and the soma originate from a reduced number of L1 copies with hot activity.

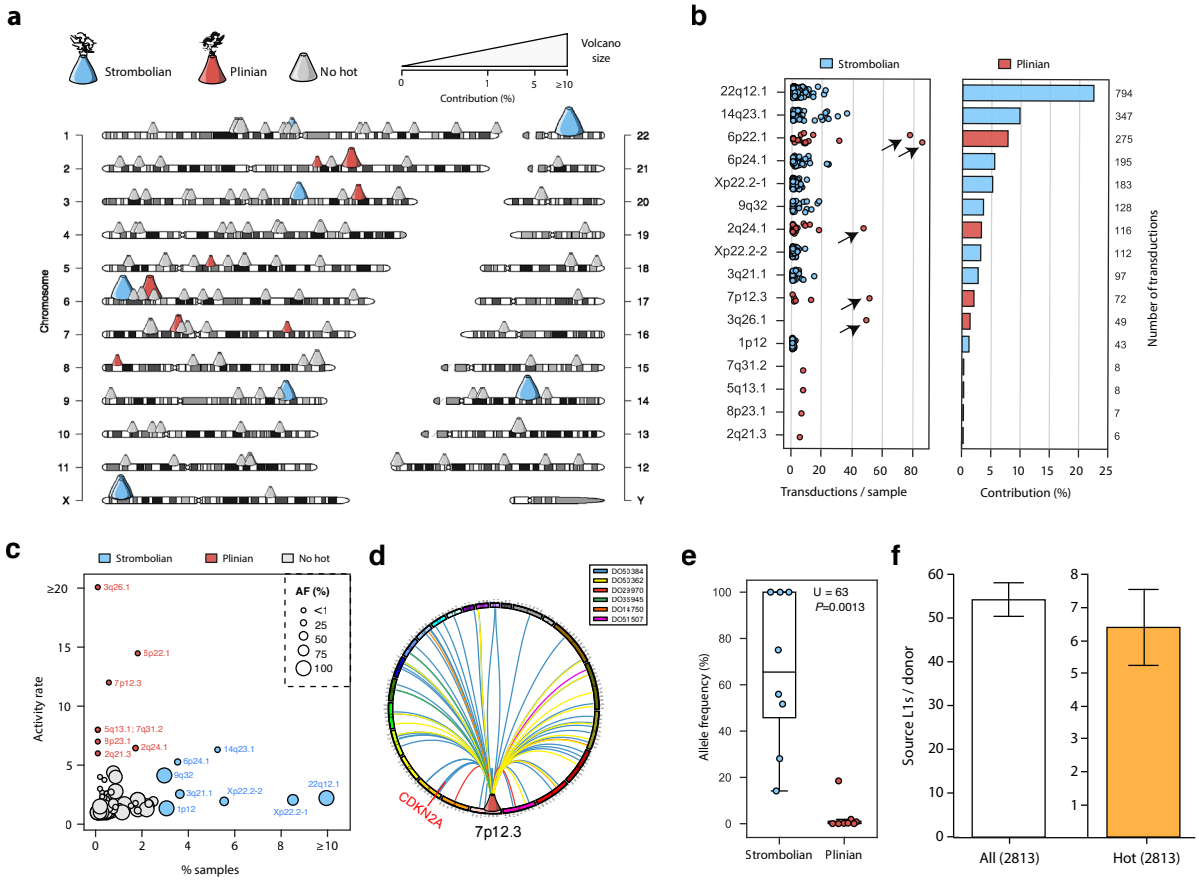


Figure 6. A catalogue of germline L1 source loci operating in cancer. **(a)** Chromosomal map where each source L1 is displayed as a volcano. Each volcano is coloured according to the type of source L1 activity. The contribution of each source loci, expressed as percentage, to the total number of transductions identified in PCAWG tumours is represented in a size gradient, with top contributing elements exhibiting larger sizes. **(b)** On the left, dots show the number of transductions promoted by each hot element in individual samples. Arrows highlight retrotransposition burst. On the right, the contribution of each hot locus is represented. The total number of transductions mediated by each source element is shown in the right side of the panel. **(c)** Source L1 activity rate (i.e., measured as the average number of transductions mediated by the element) versus the percentage of samples with retrotransposition activity where the element is active. Source L1 clustering with DBSCAN reveals a well-defined grey cluster composed by L1s without hot activity and two clearly differentiated groups of outliers, corresponding to Plinian and Strombolian hot loci, respectively. Source L1s allele frequencies are illustrated through a point size gradient, with common elements exhibiting larger sizes than rare. Extreme points observed for a source L1 with an activity rate of 49 and for an L1 active in 31% of the samples are shown at "≥20" and "≥10", respectively (for visualization purposes). **(d)** Novel Plinian germline source element in 7p12.3 mediates 72 transductions amongst only 6 cancer samples. This includes a transduction that induces the deletion of the tumour suppressor gene *CDKN2A*. **(e)** Contrasting allele frequencies for Strombolian and Plinian source loci (P=0.0013; Mann-Whitney U test). **(f)** Number of germline source and hot-L1s per PCAWG donor.

To investigate the pan-cancer patterns of source L1 activity, we performed a clustering analysis based on two metrics: (1) the proportion of samples where a given source L1 element displayed transduction activity, and (2) the activity rate of each element, estimated as the average number of transductions mediated by a given source L1 across all the samples where the element was active. Source L1s with hot activity differentiated into two (a-b) groups of outliers (Figure 6c): (a) Hot loci active in a small set of samples (< 2%) but with high activity

rates (>= 5 transductions per sample on average), leading to the accumulation of up to 86 somatic transductions in a single cancer genome; and (b) Hot source L1s frequently active (>= 3% of samples) at low activity rates.

These two distinct patterns of hot activity resemble volcano eruption types, namely Plinian and Strombolian, as follows. Plinian volcanoes, such as Mt. Vesuvius at Pompeii, are characterized by sporadic but particularly intense volcanic activity. Meanwhile, Mt. Stromboli, in Italy, has been in almost continuous activity for 2,000 years, producing mildly explosive eruptions. In analogy, Plinian L1s were rarely active across tumours (Figure 6d), but produced intense bursts of somatic retrotransposition, while Strombolian L1s were frequently active in cancer, but mediated only small-to-modest eruptions of somatic L1 activity.

We found that, whereas Strombolian elements are relatively common (MAF > 2%) and sometimes even fixed alleles in the human population, all Plinian elements are rare polymorphisms (MAF ≤ 2%) (Figure 6e). This remarkable dichotomous pattern of activity and allele frequency may be the consequence of differences in their age and in the selective pressure acting upon these L1 loci, with Plinian elements likely representing recently acquired hot-L1s, which have not yet reached an equilibrium with our species (see section D.3).

Each donor (i.e., patient) in pan-cancer bears on average between 50 and 60 L1 source copies and between five and seven hot-L1 elements, but only 38% (1075/2814) of all donors carries ≥1 Plinian element (Figure 6f). Given the mutagenic potential of these copies, we searched for SNPs in linkage disequilibrium, finding tagging SNPs for 78% (53/68) of the polymorphic source L1 with MAF≥0.1%, which will enable the incorporation of these copies in future genome wide association studies of cancer susceptibility (see section D.3).

C2.2 Source L1 activity across cancer types

We observed that the number of source L1 with transduction activity in a tumour was highly heterogeneous, ranging between zero and 22 active L1s per sample (Figure 7a). Remarkably, the number of active L1s in a cancer genome strongly correlated with the number of somatic retrotranspositions (Spearman's rho = 0.76, P < 0.05, Figure 7b). Similarly, cancer types with high retrotransposition rates (i.e., colon, head-and-neck, lung and esophageal) had 2-4 active source elements on average per sample, which represented a 4-8 fold enrichment with respect to the average across the whole pan-cancer dataset (mean = 0.5). This enrichment was particularly pronounced in a remarkable head-and-neck squamous cell carcinoma, SA494351, which had the highest number of somatic L1 insertions (n=638) and active source elements (n=23) identified in a single tumour in the PCAWG cohort. Altogether, the data indicates that the cumulative contribution of multiple active copies largely determines the retrotransposition load in a given cancer genome, with tumours displaying high retrotransposition rates being characterized by a high number of active source L1.

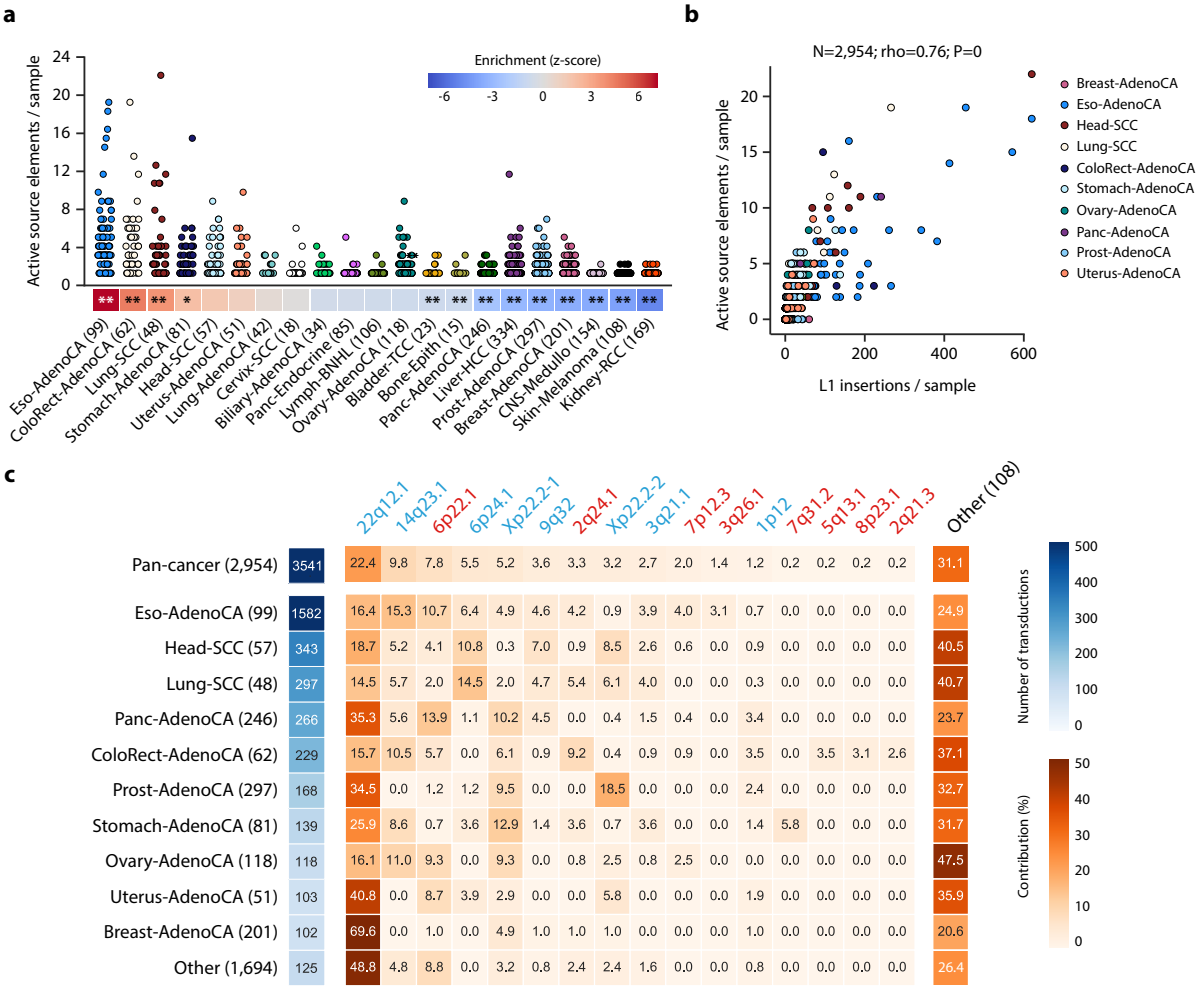


Figure 7. The dynamics of L1 source elements activity in human cancer. (a) Number of active germline L1 source elements per sample, across cancer types with source element activity. A source element is considered to be active in a given sample if it promotes at least one transduction. Enrichment or depletion in the number of active source elements for each tumour type together with level of significance (Zero-inflated negative binomial regression) are shown. P-values lower than 0.05 and 0.01 are indicated with one and two asterisks, respectively. The number of samples analyzed for each tumour type is shown in parenthesis. **(b)** Correlation between the number of somatic L1 insertions and the number of active germline L1 source elements in PCAWG samples. Each dot represents a tumour sample and colours match cancer types. Sample size (N) together, Spearman's rho and P-value are shown above the panel. **(c)** The total number of transductions identified for each cancer type is shown in a blue coloured scale. Sample size for each tumour type is shown in parenthesis. Contribution of each source element is defined as the proportion of the total number of transductions from each cancer type that is explained by each source loci. Only hot-L1s are shown while the remaining are grouped into the category 'Other'. Cytoband identifiers for hot-L1s are coloured in blue and red for Strombolian and Plinian elements, respectively.

Notably, we observed that the activity of individual hot-L1s was frequently associated with specific tumour types (Figure 7c). For example, the source element at 6p24.1 was the second most active copy in lung (14.5%) and head-and-neck (10.8%) cancers, while it remained silent at colon cancers. 6p24.1 is a common polymorphism in the human population (MAF = 28% at pan-cancer), which displayed similar allele frequencies among lung (27%), head-and-neck (25%) and colon (22%) cancer patients. This indicates that the observed activity patterns were

not the result of differences in the prevalence of 6p24.1 alleles across patients from different tumour types. Similarly, the hot-L1 at the cytogenetic band Xp22.2-2 was the second more active copy in prostate tumours, contributing to 18.5% of all transductions. However, Xp22.2-2 displayed more modest levels of activity in head-and-neck (8.5%), lung (6.1%) and uterus tumours (5.8%), while being a minor player in the retrotransposition landscape at the remaining tumour types. As described for 6p24.1, Xp22.2-2 is a common polymorphism (MAF = 56% at pan-cancer) that displayed similar allele frequencies at prostate (67%), head-and-neck (52%), lung (57%) and uterus (56%) cancer patients. Nonetheless, the hot-L1 at 22qq12.1 was a clear exception to these patterns, as it was consistently the most active source L1 for all the cancer types investigated.

Overall, the marked levels of heterogeneity observed in L1 source activity across tumour types may be the consequence of differences in the genomic context where hot-L1 copies are allocated. Given that epigenetic silencing is known to dictate source L1 activation[44,51,153], we believe that tissue-specific differences on the epigenetic programs for distinct tumour histologies may drive these patterns (see section D.3).

C2.3 Contributors

This study was initiated by Alicia L. Bruzos, who generated the initial catalogue of 114 source L1 elements as part of her master thesis. Javier Temes and Eva G. Alvarez created the karyoplots with active L1 elements represented as volcanoes. Sebastian Waszak, Jan Korbel and all the members from the “Germline Working Group” provided valuable feedback and ideas.

C2.4 Publications

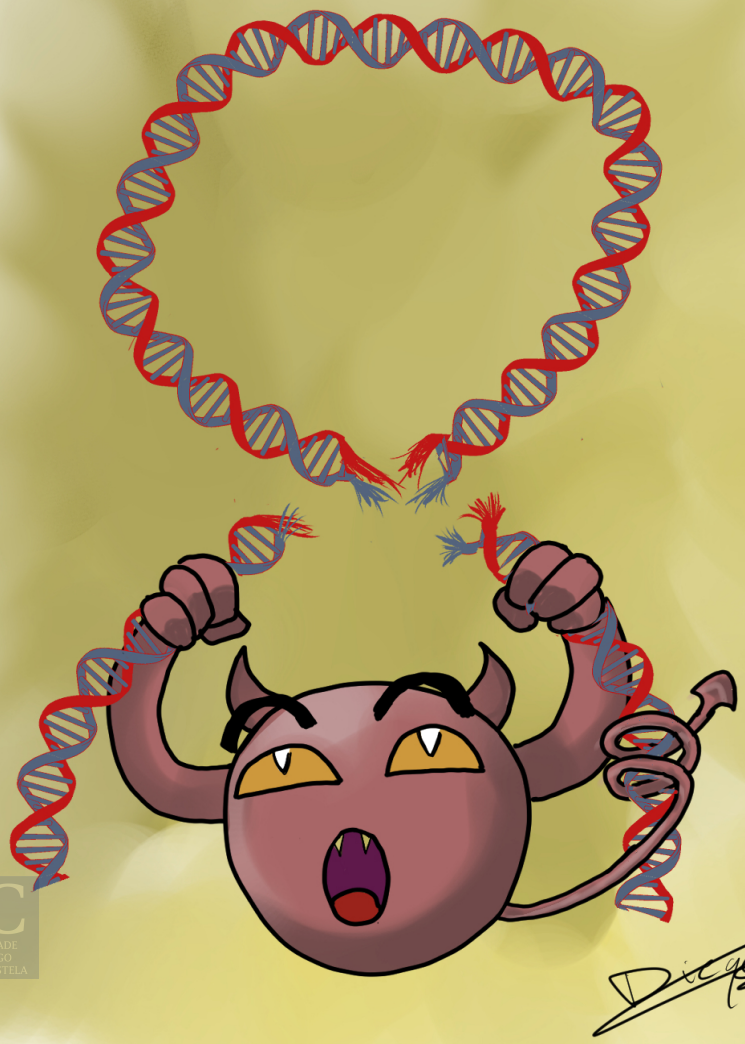
This chapter’s content derives from two manuscripts published in Nature [1] and Nature Genetics [2], respectively. The text was considerably rewritten, extended and adjusted to fit the flow of this thesis. All figures included derive either from the main or supplementary figures of the aforementioned publications. The complete list of authors and their affiliations is provided in the appendix.

[1] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020). DOI: https://doi.org/10.1038/ s41586-020-1969-6, ISSN: 0028-0836

[2] Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat Genet 52, 306–319 (2020). DOI: https://doi.org/10.1038/s41588-019-0562-0, ISSN: 1061-4036

CHAPTER 3:

“L1-mediated structural variation in the cancer genome”



C3.1 Genomic deletions induced by aberrant L1 integration

During the analysis of somatic retrotransposition, we noticed a very intriguing pattern in some cancer genomes with high levels of somatic L1 activity. A single cluster of reads supporting one of the ends of an L1 integration was associated with the beginning of a CN loss (Figure 8a). Further analysis of the CN change revealed the missing reciprocal cluster, which supported the second end for the L1 integration, at the far end of the CN loss (Figure 8b), suggesting that the deletion has occurred in conjunction with the integration of an L1. A poly(A) stretch was present at one of the breakpoints of the CN loss together with the L1-EN target motif. These hallmarks resemble those previously reported for L1-mediated rearrangements[77,78], suggesting that the aberrant integration of an L1 sequence led to the loss of DNA.

We developed a computational method to systematically search for L1-mediated deletions across the complete set of PCAWG cancer genomes. We detected 90 somatic events that matched the patterns described above, spanning deletions of different sizes, ranging from 0.5 Kbp to 53.4 Mbp (Figure 8c). To rule out the possibility that these deletions were mediated by the homologous recombination of two different L1s on the reference, we searched for candidate L1-mediated deletions containing unequivocally a single L1 insert at their breakpoints. These include small deletions and L1 inserts that are shorter than the library size, allowing sequencing read-pairs to overlay the entire structure. For example, in a lung tumour sample (SA313800), we identified a 1 Kbp deletion containing two distinct clusters of discordant read-pairs at its breakpoints (Figure 8d). While a fraction of the discordant read-pairs clearly indicated the existence of an L1 insertion, the others were able to overpass the inserted L1 fragment unambiguously supporting the deletion. Another type of L1-mediated deletion that can unequivocally be assigned to a single L1 insertion event are deletions generated by the integration of orphan L1 transductions. For example, in one esophageal tumour sample (SA528932), an orphan transduction mobilized by a source L1 at chromosome 7 caused a deletion of 2.5 Kbp in size (Figure 8e), confirming that is derived from a single somatic retrotransposition event.

We successfully reconstructed the breakpoint junctions for the 90 deletions mentioned above, finding an L1-derived sequence in all of them. In addition, 82% (74/90) of the reported deletions contained a sequence resembling L1-EN consensus cleavage sites at their 3' breakpoints (5′-TTTT/A-3′ degenerated motif). This confirms that L1 machinery, through TPRT, was responsible for the integration of most of the L1 sequences that caused neighbouring DNA loss. For 16% (14/90) of the events, the cleavage occurred at the phosphodiester bond between a T and G instead of the standard T and A site. Additionally, all the deletions associated with L1-EN motifs also contained a polyadenylate tract at their 3' breakpoints, indicative of passage through an RNA intermediate and further supporting a TPRT related origin. Overall, these features are consistent with the mechanistic model proposed two decades ago based on *in vitro* assays[77,78], where a variant of the canonical TPRT reaction leads to the loss of DNA.

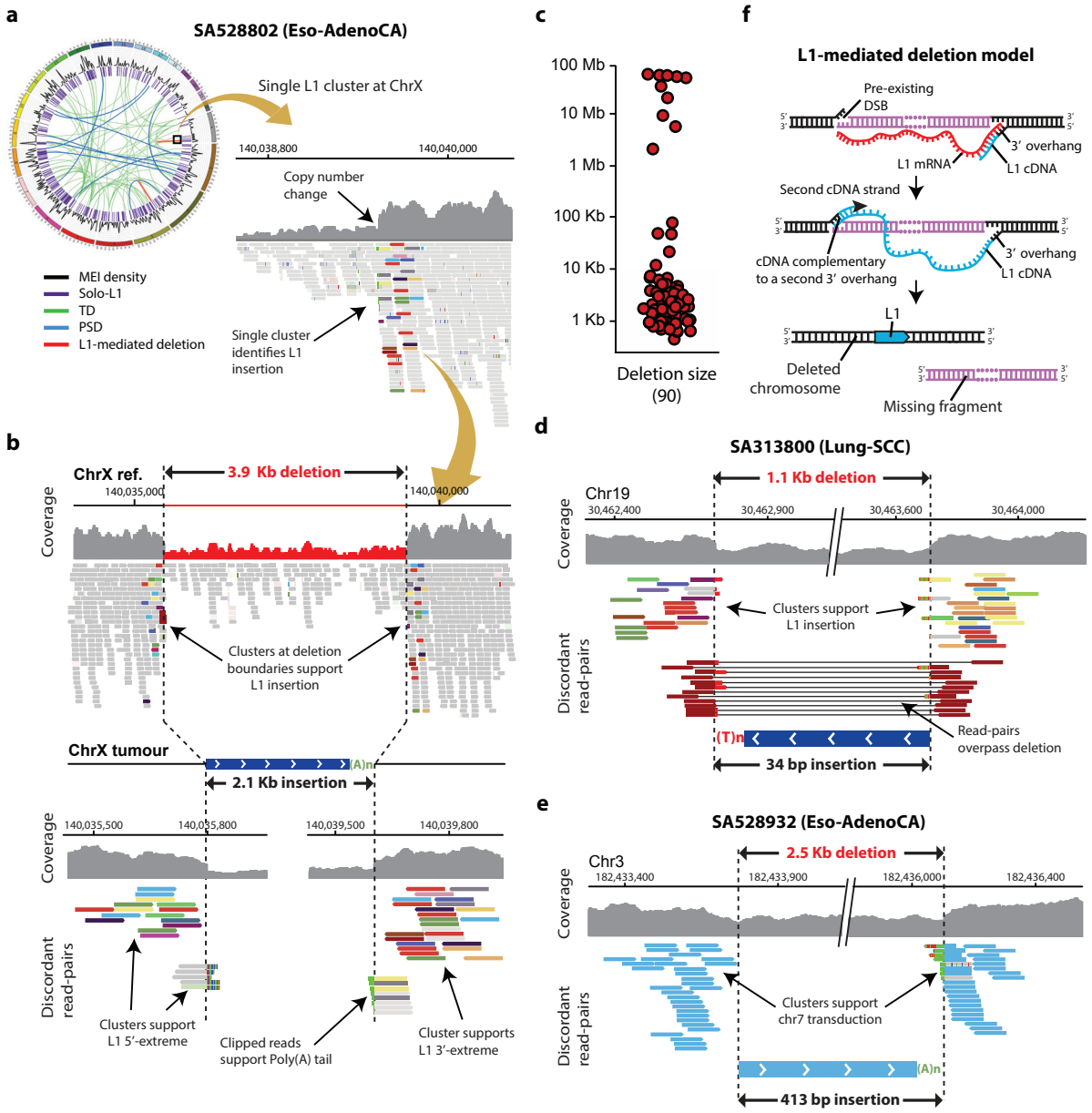


Figure 8. The hallmarks of somatic L1-mediated deletions revealed by copy number and paired-end mapping analysis. (a) A discordant read-pair cluster supporting the 5' end for an L1 integration is associated with the beginning of a CN loss. Paired-end reads are coloured by the chromosome on which their mates can be found, with multi-coloured clusters being composed by reads with mates aligning over multiple L1 loci along the genome. **(b)** The missing discordant cluster supporting the 3' end for an L1 integration is located at the end of the CN loss, indicating the existence of a 3.9 Kbp deletion occurring in conjunction with the integration of a 2.1 Kbp L1 somatic insertion. (A)n and (T)n represent poly(A) and poly(T) tails, respectively. **(c)** Distribution of the sizes for the 90 L1-mediated deletions identified in the PCAWG dataset. **(d)** L1-mediated deletion with a short L1 sequence bridging the deletion breakpoints. Multi-coloured read-pair clusters support the L1 bridge, while discordant reads spanning the L1 insert and therefore supporting the deletion are displayed in red. **(e)** L1-mediated deletion caused by an orphan transduction. Blue coloured discordant read-pairs clusters at the deletion breakpoints support the integration of a 413-bp transduced sequence from a source element located in chromosome 7. **(f)** Model for L1-mediated deletion formation.

Remarkably, 75% (47/63) of the deletions with their breakpoints characterized to base-pair resolution had short (1–5 bp long, median=3 bp) microhomologies between the pre-integration site and the 5′ end for the L1 sequence integrated right there. In addition, for 14% (9/63) of the instances short insertions (1–33 bp long, median=9 bp) were found at the breakpoint junctions. Both signatures suggests that the 5' end of an L1 cDNA molecule is attached to a distal 3' overhang derived from a pre-existing double-strand break upstream of the initial integration site via "error-prone" NHEJ or other type of microhomology mediated repair[154]. As a consequence, the L1 insert bridges the L1-EN cleavage site and the double strand break, resulting in the loss of the DNA interval between both breakpoints (Figure 8f).

In contrast, 8% (7/90) of the L1-mediated deletions exhibited a different insertion pattern. Here, the L1-EN motif was not found at the deletion breakpoints, the L1 integrant lacked a poly(A) tail and it was truncated at both ends. This integration pattern was previously reported to be indicative of endonuclease independent retrotransposition, an alternative mechanism for L1 integration[135]. In 2002, Moran's laboratory observed that L1 sequences were able to mobilize *in vitro*, even after inactivation of L1-EN, in NHEJ deficient cells[135]. Under NHEJ deficiency, DNA breaks provide 3′-hydroxyl residues that can serve as primers for the reverse transcription of L1, leading to its insertion and concomitant repair of the DNA lesion. The identification of endonuclease independent retrotransposition events in the PCAWG dataset indicates that this alternative mechanism for L1 integration can also naturally occur in the context of cancer, leading to DNA losses.

C3.2 Megabase-size L1-mediated deletions cause loss of tumour suppressor genes

Although most of the reported L1-mediated deletions ranged from a few hundred to thousands of base pairs, on occasion, The aberrant integration of L1 sequences caused the loss of megabase-sized chromosomal regions. For example, in one esophageal tumour sample (SA528901) we detected a 45.5 Mbp interstitial deletion affecting chromosome 1 (Figure 9a) that displayed the described hallmarks for TPRT. Here, the L1 element was heavily truncated, allowing a fraction of the sequencing read-pairs to span the complete L1 insert, unequivocally supporting that the observed CN change was indeed a deletion mediated by L1 retrotransposition. Similarly, in a remarkable lung tumour (SA313800), the insertion of an L1 promoted a 51.1 Mb-long deletion at chromosome X that included the centromere (Figure 9b).

L1-mediated deletions can be occasional cancer driver events through the removal of tumour suppressor genes. For example, in an esophageal tumour sample (SA528932), a partnered L1 transduction from a source L1 loci at 7p12.3 caused a 5.3 Mbp loss affecting the short arm of chromosome 9 that removed *CDKN2A* (Figure 9c), a relevant tumour suppressor gene that is frequently mutated in many cancer types[9], including esophageal tumours. Notably, the integrated sequence was a FL-L1, which mobilized itself into chromosome X producing a second 3' transduction (Figure S6). This observation showcases that L1 integrants involving

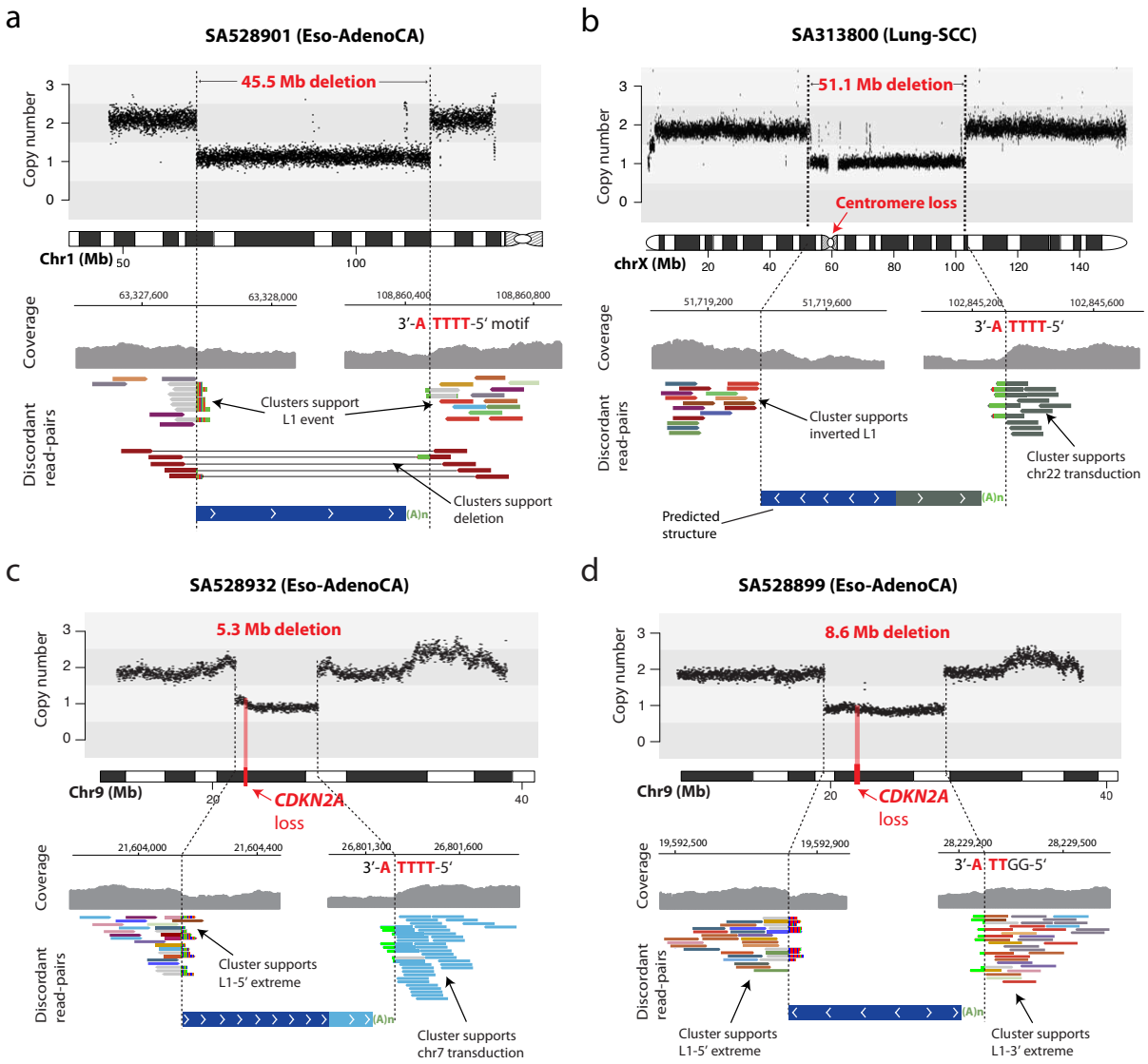


Figure 9. Somatic integration of L1 causes loss of megabase-size interstitial chromosomal regions in cancer. **(a)** Large-scale interstitial deletion on chromosome 1 associated with the integration of a short L1 sequence. The L1-EN cleavage motif (5′-TTTT/A-3′) is found at one deletion breakpoint, confirming that aberrant L1 integration via TPRT caused the observed rearrangement. **(b)** Partnered transduction from a source element at chromosome 22 promotes a 51.1 Mbp deletion removing the centromere. **(c)** Partnered transduction causing a 5.3 Mbp deletion leading to the loss of one copy of the tumour suppressor gene *CDKN2A*. **(d)** Integration of an L1 retrotransposon generates an 8.6 Mbp deletion that affects *CDKN2A* in a second esophageal adenocarcinoma patient.

large deletions can be competent for retrotransposition. In a second esophageal tumour sample (SA528899), an L1 integration at chromosome 9 promoted a deletion of 8.6 Mbp in size that, again, removed *CDKN2A* (Figure 9d). Analysis of the variant allele frequencies revealed that both deletions were clonal, suggesting that they may have occurred early during the evolution of these tumours. These findings highlight the potential of aberrant L1 integration to promote DNA losses with oncogenic roles.

C3.3 L1 elements can generate a wide variety of structural variation classes

While searching for L1-mediated deletions, we noticed that aberrant L1 retrotransposition can be implicated in the generation of other forms of structural variation. In one esophageal tumour sample (SA528896), two separate L1-mediated translocations were observed within the context of a complex cluster of rearrangements (Figure 10a). First, an L1 transduction from a source element at 14q23.1 was associated with an unbalanced translocation involving 1p and 5q. Second, another L1 sequence was found at the junction between 5p and an unknown genomic locus, completing a large interstitial CN loss on chromosome 5 that affected the centromere.

These observations suggest that L1 retroelements can bridge double strand breaks located in different chromosomes. To further investigate this question, we mined publicly available sequencing data for a lung cancer cell line (NCI-H2087) previously reported to have high levels of somatic retrotransposition[44]. This search uncovered a translocation connecting 1q31.1 with 8q24.12 in NCI-H2087. Interestingly, both translocation breakpoints were flanked by discordant read-pair clusters supporting an orphan L1 transduction derived from an L1 source element located at chromosome 6p24 (Figure 10b). Both L1-EN motif and poly(A) stretches were found, suggesting that the interchromosomal rearrangement likely originated through the aberrant operation of the canonical TPRT reaction. More precisely, the transduced cDNA sequence may pair with a 3′ overhang derived from a pre-existing double-strand break in a second chromosome, resulting in an L1-mediated translocation event (Figure 10c).

We also found evidence that aberrant L1 integration can also cause duplications of large genomic regions in cancer. For example, in another relevant esophageal tumour sample (SA528848), we identified two discordant read-pair clusters that supported the integration of a truncated L1 element, coupled with an increase of coverage delimited by both L1 insertion breakpoints (Figure S7a). CN analysis indicated that the two L1 clusters demarcated the boundaries of a 22.6 Mbp duplication, suggesting that the L1 insertion could be the cause of such rearrangement by bridging sister chromatids during or after DNA replication (Figure S7b). Detailed analysis of the sequencing data revealed the presence of additional discordant read clusters that supported both a tandem duplication and an L1-EN motif at the 3' insertion breakpoint, confirming a single L1 event as the cause of that tandem duplication. Notably, this rearrangement increases the CN of the cyclin C gene, *CCNC*, which is dysregulated in some tumours[155].

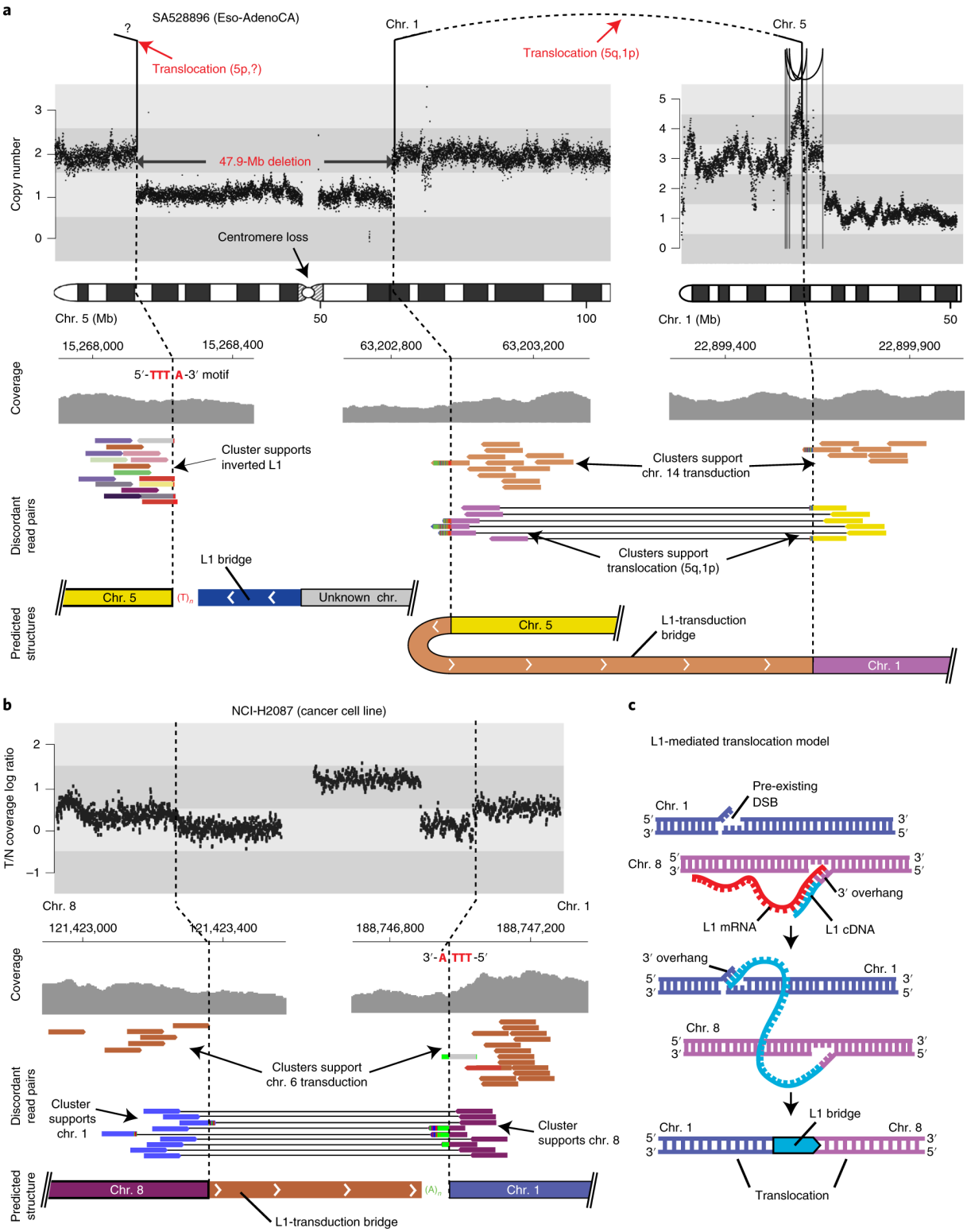


Figure 10. Somatic L1 integration promotes translocations in human cancers. (a) Complex rearrangement, including two translocations containing discordant read-pair clusters supporting L1 insertion events at their breakpoints. In the first, an L1 transduction from a source element on chromosome 14q23.1 bridged an unbalanced translocation between 1p and 5q. A second somatic retrotransposition event bridged 5p with an unknown locus, completing a 47.9 Mbp interstitial CN loss. **(b)** An orphan transduction from a source element at chromosome 6 bridges chromosomes 1 and 8, generating a translocation. **(c)** Model for L1-mediated translocation formation. .

C3.4 Breakage-fusion-bridge cycles initiation by L1 retrotransposition

Breakage-fusion-bridge (BFB) cycles are a mechanism of genomic instability that is initiated with the end-to-end fusion of broken chromatids, either from the same or two different chromosomes, generating a dicentric chromosome[156]. During mitosis, the two centromeres of a dicentric chromosome are pulled to opposite poles of the dividing cell, creating an anaphase bridge, which is resolved by double strand DNA breakage at an arbitrary position between both centromeres. As the resulting chromosomal products typically exhibit further telomere deficiencies, the chromosome is likely to undergo multiple BFB rounds till it gets finally stabilized through the repair of its ends. BFBs are particularly relevant in the context of cancer[157,158], as successive BFB rounds frequently lead to the amplification of oncogenes, contributing the necessary genomic alterations for malignant transformation.

Although initially discovered by McClintoc in the late 1930s[159,160], when she observed frequent chromosome breakage and fusion at mitotic maize cells after X-ray exposure, end-to-end fusions are currently known to originate through multiple mechanisms[161], including telomere attrition and chromothripsis. During the analysis of somatic retrotransposition, we found that aberrant L1 retrotransposition can be an alternative mechanism for the end-to-end fusion of sister chromatids and dicentric chromosome formation. In a lung cancer tumour (SA313800), we identified two discordant read-pair clusters with the same orientation and located 5.5 Kbp apart that supported the presence of an L1 insertion along the long arm of chromosome 14 (Figure S7c). Both clusters colocalize with two discordant read-pair clusters in head-to-head orientation, which represent the classical read signature for a fold-back inversion[162], suggesting that the L1 insertion event was involved in the generation of the inverted rearrangement. In addition, the fold-back inversion breakpoints are exactly at the boundary between a large-scale deletion affecting the first 27 Mbp of chromosome 14 and a 79.6 Mbp duplication of the 14q arm. The only genomic structure that adjusts to these patterns is a fold-back inversion in which the two sister chromatids are bridged by an L1 insertion, generating an isochromosome (14q). Again, this can be explained through a variant of TPRT reaction (Figure S7d), where the L1 cDNA uses a pre-existing double strand break to invade the sister chromatid during DNA replication, resulting in the described genomic conformation.

In the example described above, no further breaks occurred and the isochromosome remained stable. However, we found examples in which the fusion of two chromatids by an L1 bridge induced further cycles of BFB repair. In an esophageal tumour sample, SA528848, we identified a cluster of reads on the long arm of chromosome 11 that had the typical hallmarks of an L1-mediated rearrangement (Figure 11a). CN data analysis showed that the L1 insertion breakpoints demarcated a 53 Mbp deletion, which involved the loss of the telomeric region, and a massive amplification on chromosome 11. The amplified region on chromosome 11 contains the *CCND1* oncogene, which is amplified in many human cancers[163]. The other end of this amplification was bound by a conventional fold-back inversion rearrangement (Figure 11a), which is indicative of BFB repair[162,164].

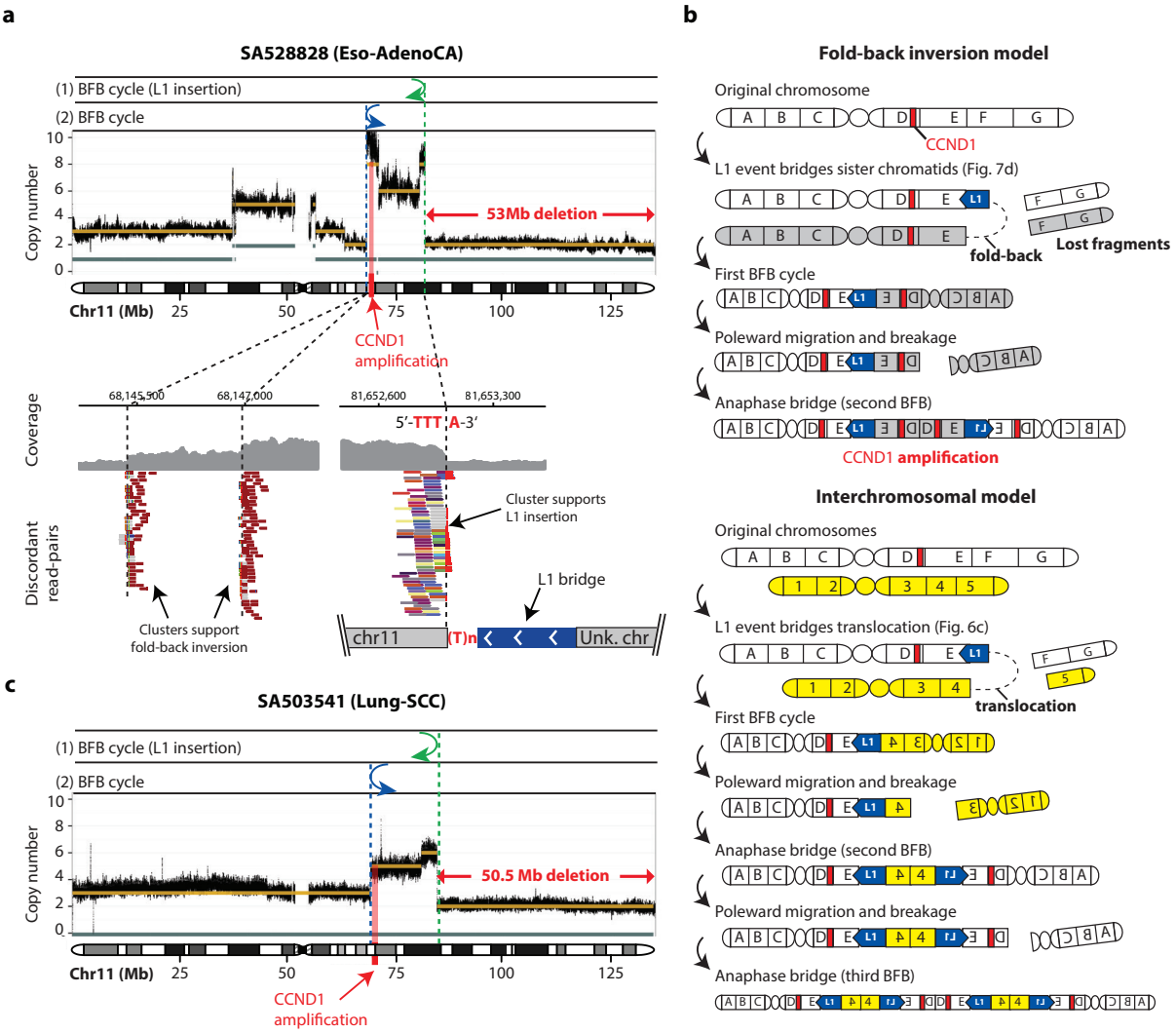


Figure 11. Somatic integration of L1 can trigger breakage-fusion-bridge cycles that lead to oncogene amplification. (a) A multi-coloured cluster supporting the 3' end for an L1 insertion is associated with a telomeric deletion and adjacent large-scale amplification involving the oncogene *CCND1*. The L1-endonuclease cleavage site motif (5′-TTT/A-3′) is found at the junction between the insertion and deletion breakpoints. Two additional discordant read-pair clusters (brown reads) in head-2-head orientation demarcate a CN change within the amplicon. These patterns are consistent with two BFB rounds, marked with (1) and (2), associated with L1 integration. The CN plot shows the consensus total CN (gold band) and the minor allele CN (Gray band). **(b)** Two potential mechanistic models that explain the patterns described above. Both models differ on the type of event triggering the first BFB-cycle: fold-back L1-mediated inversion and translocation, respectively. **(c)** A second BFB leading to the amplification of *CCND1* oncogene that is associated with an L1 insertion.

These patterns suggest the following sequence of events. During or soon after S phase, a somatic L1 retrotransposition bridges across sister chromatids in inverted orientation, breaking off the telomeric ends of 11q, which are then lost during the subsequent cell division (fold-back inversion model, Figure 11b). The chromatids bridged by the L1 insertion now produce a dicentric chromosome. During mitosis, the two centromeres are pulled to opposite poles of the dividing cell, creating an anaphase bridge, which is resolved by further dsDNA breakage. This induces a second cycle of BFB repair, albeit not mediated by L1 retrotransposition. Then,

successive BFB-cycles lead to rapid-fire amplification of the *CCND1* oncogene. Alternatively, an interchromosomal rearrangement mediated by L1 retrotransposition (interchromosomal rearrangement model, Figure 11b) followed by multiple BFB rounds could generate similar CN patterns with telomere loss and amplification of *CCND1*.

We identified four additional instances of BFB initiated by L1 retrotransposition in the PCAWG dataset (Figure S8). Remarkably, in a lung adenocarcinoma, SA503541, we found an L1-mediated rearrangement that clearly resembles the patterns described above (Figure 11c), including telomere loss, *CCND1* amplification and read-clusters supporting a fold-back inversion within the amplicon. In this case, the data is consistent with two BFB rounds, leading to the acquisition of two extra copies of *CCND1*. The independent occurrence of similar rearrangements, involving the amplification of the same oncogene, in two different tumour samples (SA528848 and SA503541) showcase the relevance of this novel L1-mediated mutational process. Overall, these data indicate that L1 retrotransposition is an alternative mechanism for the formation of dicentric chromosomes, leading to the initiation of BFB-cycles. If this occurs near an oncogene, such as *CCND1*, the resulting amplification can provide a powerful selective advantage to the clone and potentially lead to cancer development.

C3.5 Contributors

José M.C. Tubio led the computational analysis for the detection of L1-mediated rearrangements. Adrian Baez-Ortega implemented components of the computational approach for L1-mediated rearrangements detection. Jonas Demeulemeester and Peter Van Loo performed the clonality analysis for L1-mediated rearrangements and generated the CN plots. Yilong Lee generated the plots for the joint visualization of rearrangement junctions and CN data. Peter J. Campbell provided essential intellectual input in the interpretation of sequencing data, which led to the identification of BFB-cycles initiated by L1 retrotransposition.

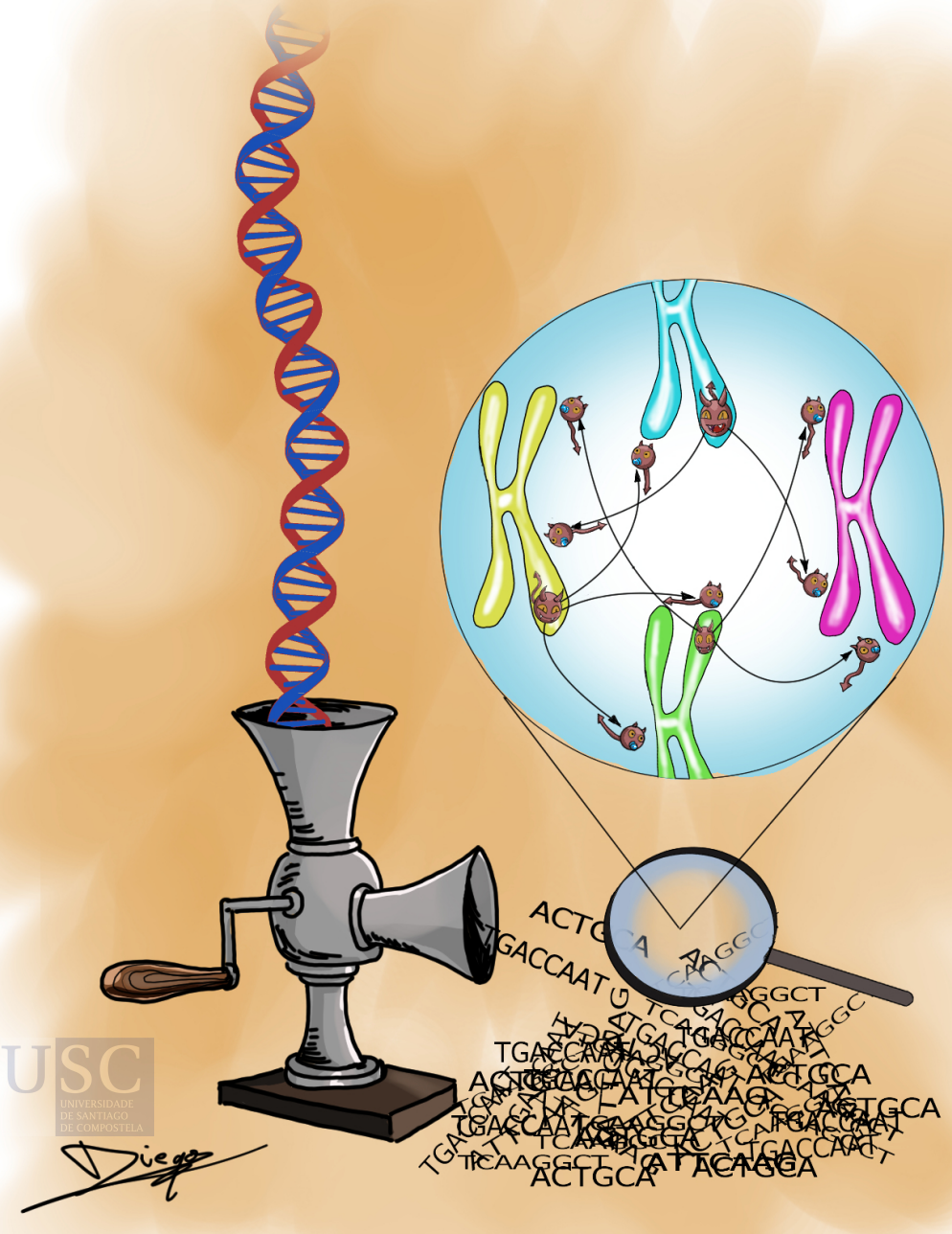
C3.6 Publications

This chapter's content originates from multiple sections of a published manuscript [1]. The text was considerably rewritten and adjusted to fit the writing style and flow of this thesis. All figures included derive either from the main or supplementary figures of the aforementioned publication. The complete list of authors and their affiliations is provided in the appendix.

[1] Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat Genet 52, 306–319 (2020). DOI: https://doi.org/10.1038/s41588-019-0562-0, ISSN: 1061-4036

CHAPTER 4:

“Computational methods for mobile element insertion detection”



C4.1 The TraFiC-mem algorithm

TraFiC-mem represents an updated version of the algorithm TraFiC[44] (Transposon Finder in Cancer), which was specifically designed for the identification of somatic retrotransposition events in cancer genomes by means of the analysis of Illumina paired-end sequencing data. Contrary to the previous version of the algorithm, the new pipeline uses BWA-mem instead of RepeatMasker as a search engine for the identification of retrotransposon-like sequences in the sequencing reads. This enables a more efficient processing of the cancer genome data, reducing computing time and memory usage.

TraFiC-mem takes as input tumour and matched-normal bam files containing pre-aligned reads onto the human reference genome. Although it should be compatible with bams derived from any short-read mapper, it has been extensively tested with BWA-mem alignments, which are, therefore, the preferred choice. TraFiC-mem leverages discordant read-pair and clipped-read information to identify somatic MEIs including solo-L1, L1-mediated transductions, Alu, SVA and ERV-K. TraFiC-mem is implemented using Snakemake[165], a flexible Python-based workflow language, that allows to execute the pipeline from single-core workstations to computing clusters, without the need to modify the workflow. In order to enhance reproducible research and its portability to cloud computing platforms, TraFiC-mem and its third party dependencies are also distributed as a Docker container (https://hub.docker.com/r/mobilegenomes/trafic). TraFiC-mem is available together with complete documentation and tutorial (https:// gitlab.com/mobilegenomes/TraFiC). TraFiC-mem pipeline consists of 6 major steps which are described below (Figure 12).

Identification of MEI candidates via discordant read-pair analysis

Discordant read-pairs are collected from the tumour and normal compressed alignments in BAM format. Read-pairs are considered discordant if, based on the alignment bitwise FLAG, they are not mapping with the expected insert size, align onto different chromosomes or one of the mates fails to align. The read-end with the highest mapping quality is considered the anchor. Discordant read-pairs are filtered out if the anchor has a mapping quality (MAPQ) of zero, if both reads are supplementary alignments or are flagged as PCR or optical duplicates. Then, discordant read-pairs are subjected to two different approaches for the detection of solo retroelement insertions and transductions.

For solo insertions, non-anchor reads are realigned with BWA-mem[116] v0.7.17 onto a library of human mobile element consensus sequences, including L1, Alu, SVA and ERV-K. After BWA-mem search, all anchored reads with mates aligning over the same mobile element class are clustered together if (a) they have the same mapping orientation, forward or reverse, and (b) they reciprocally overlap after extending their mapping intervals by the library read size. Two types of discordant clusters are generated, namely forward and reverse, according to the orientation of anchor reads. Clusters are filtered by requiring at least 4 supporting read-pairs and those with enough support go into the meta-clustering stage. For each cluster a range is

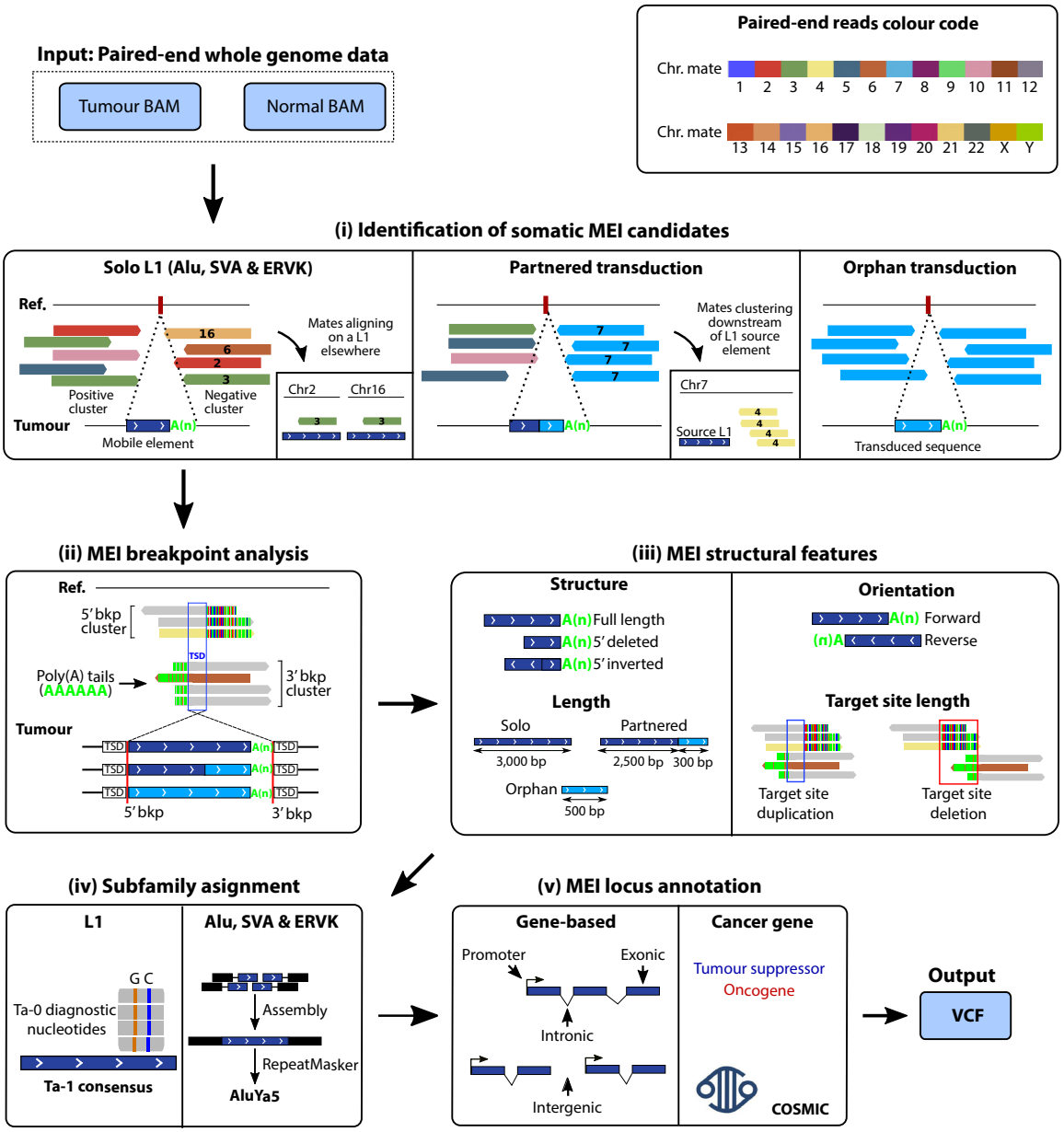


Figure 12. Overview of TraFiC-mem. TraFiC-mem detects somatic mobile element insertions (MEIs) from paired-end mapping data through 5 consecutive steps: (i) Identification of candidate MEI. Solo-retrotransposon insertions are detected by the identification of two reciprocal clusters (positive and negative, or head-to-head) of interchromosomal reads whose mates map onto retrotransposons of the same type located elsewhere in the genome. Partnered transductions are detected by the identification of one cluster of interchromosomal reads whose mates map onto L1 retrotransposons of the same family elsewhere in the genome, and one single reciprocal cluster of reads whose mates are clustered at a unique region adjacent to a donor source L1 element (the example illustrates a transduction from chromosome 7). Orphan transductions are detected by the identification of two reciprocal clusters whose mates map downstream to a source element as described for partnered transductions. **(ii)** MEI breakpoint analysis. TraFiC-mem seeks for two additional clusters (5' breakpoint cluster and 3' breakpoint cluster) of clipped-reads in the candidate insertion region, in order to reveal the 5' and 3' insertion breakpoint coordinates to base-pair resolution. **(iii)** MEI structural features annotation. **(iv)** Subfamily assignment. Subfamily specific diagnostic nucleotides are used to determine the subfamily for L1 events. **(v)** MEI locus annotation: The target genomic region is annotated and MEIs inserted within cancer genes, according to the COSMIC database, are flagged. Output is a VCF file.

defined based on its left-most and right-most mapping positions after considering all reads composing the cluster. Forward and reverse cluster ranges are extended at their ends by the library size and then paired together by reciprocal overlap into meta-clusters supporting both ends of candidate MEIs. Unpaired solo-like clusters supporting only a single end of a candidate MEI are further considered in the transduction and L1-mediated rearrangement detection modules.

For L1 transductions, including orphan and partnered events, discordant read-pairs are further filtered by requiring anchors and mates to have a MAPQ ≥ 37. Anchored reads are clustered together based on the same criteria as for solo plus one additional condition, the mate alignment positions must reciprocally overlap too after the extension of their mapping intervals. Forward and reverse clusters supporting a candidate transduction event plus unpaired clusters generated during solo search are grouped based on reciprocal overlap into meta-clusters. Meta-clusters composed by two transduction supporting clusters are classified as orphan transduction candidates, while those composed by two different clusters, one supporting a solo insertion and the other a transduction, indicate a candidate partnered transduction insertion event. Remaining unpaired clusters supporting a single end of a candidate transduction are further considered in the L1-mediated rearrangement detection module.

Reconstruction of MEI breakpoints via clipped-reads analysis

Once MEI candidates have been identified via discordant read-pair analysis, TraFic-mem seeks for two additional clusters of clipped-reads (CRs) that would indicate the two exact breakpoint coordinates for each candidate insertion. Soft and hard clipped-reads are extracted within a range of ±50 relative to the end and beginning position of the positive and negative discordant clusters, respectively. Reads marked as duplicates and reads clipped both at their begin and end are filtered out, as they usually constitute mapping artefacts. After read filtering, clipped-reads are organized into clusters supporting the same breakpoint position, taking into account their clipping orientation and using a maximum offset of 3bp. Clusters supported by a single CRs, more than 500 CRs or detected in the matched-normal genome are excluded. Then, for each breakpoint cluster, supporting CRs are assembled through a multiple sequence alignment approach that uses MUSCLE[166] v3.8.31 and 'Cons' from the EMBOSS suite[167] v6.6.0. The resulting contigs are realigned on the reference with BLAT[168] v34.0 to resolve the predicted breakpoint positions and to determine if they support the genome and mobile element junction (5'breakpoint) or the genome and poly(A) tail junction (3' breakpoint). In the case that multiple 5' or 3' breakpoints are detected, the one supported by the highest number of CR is selected.

MEI structural features annotation

MEI structural features including insertion length, structure (full-length, partial, inverted), DNA strand, and size of the target-site duplication and deletion, are determined for the insertions with both breakpoints successfully reconstructed. To determine insertion size and structure, the contig supporting the 5' breakpoint is realigned to the corresponding L1, Alu, SVA or ERV-K consensus sequence using BLAT[168] v34.0. As retrotransposons only get truncated at their 5' end, the insertion length is computed as the distance between the beginning of the alignment and the end of the consensus sequence. Insertions supported by contigs spanning at least 98% of the consensus sequence are considered full-length, and 5' truncated or inverted otherwise. Truncation and inversion status are determined based on if the contig alignment orientation is the same as the insertion DNA strand (5' truncation) or different (5' inversion). Insertion orientation is inferred taking into account the relative clipping orientation for 5' and 3' breakpoint clusters. Insertions at the plus strand are characterized by 5' and 3' breakpoints supported by reads clipped at their end (end-clipped) and begin (beg-clipped), respectively, while insertions in the minus strand follow opposite clipping orientations. Target-site duplication and deletion sizes are estimated based on the distance between both breakpoints. The relative position of end-clipped and beg-clipped breakpoints is used to differentiate target-site duplications and deletions. While target-site duplications are characterized by end-clipped clusters located downstream with respect to the beg-clipped cluster, the opposite pattern is characteristic of target-site deletions.

MEI subfamily assignment

Two different strategies are applied to infer the subfamily for the L1, Alu, SVA and ERV-K inserts. For L1 insertions, non-anchor reads are realigned with BWA-mem[116] v0.7.17 on the L1 consensus sequence (GenBank identifier: L19088.1). The resulting SAM is converted into a binary sorted BAM file using samtools[125] v1.7. Genotype likelihoods at each genomic position are computed with samtools mpileup and subsequently used for variant calling with bcftools consensus caller[169] v1.7. Single nucleotide variants are filtered by requesting a quality score higher than 20 and a minimum number of 2 supporting reads. Subfamily inference is done based on the identification of subfamily diagnostic nucleotide positions[170]: L1 integrations bearing the diagnostic "ACG" or "ACA" triplet at 5,929-5,931 position are classified as "pre-Ta" and "Ta", respectively. Ta elements are subclassified into "Ta-0" or "Ta-1" according to diagnostic bases at 5,535 and 5,538 positions (Ta-0: G and C; Ta-1: T and G). For those L1 insertions without sequencing reads covering the diagnostic nucleotides the subfamily is set as undetermined.

Alu, SVA and ERV-K subfamily inference is based on Repeatmasker v4.0.7 (http://www.repeatmasker.org) annotation of contigs derived from the assembly, using Velvet[171] v1.2.10, of non-anchor reads. If multiple annotation hits are obtained, the one with the highest Smith-Waterman score is selected as representative.

MEI filtering

Candidate MEI events are filtered according to multiple criteria to produce a high-confidence set of insertions. Insertions failing to pass at least one of these conditions are discarded.

A. Insertion supported by one positive and negative cluster, with both clusters composed by at least 4 discordant read-pairs.

B. At least one insertion breakpoint is resolved at base pair resolution.

C. Consistency in the insertion features derived from 5' and 3' breakpoints.

D. MEIs without a reference element of the same family, and with ≥85% of nucleotide identity relative to the consensus sequence of the family, within a range of ±150 bp.

E. Preliminary subfamily assignment during discordant read-pair analysis consistent with final subfamily classification.

F. Insertion located outside a range 200 bp of a cluster from the same retrotransposon class detected in the matched-normal or polymorphic insertion from any of these databases: TraFiC-ip[44], dbRIP[172], 1000 Genomes Phase 3[60] and PCAWG[29].

Annotation and VCF generation

Insertion positions are annotated using the software ANNOVAR[173] v2016-02-01 and the gene annotation resource GENCODE[174] v19. In addition, MEIs inserted within cancer genes are flagged, based on the Cancer Gene Census COSMIC database[9]. The primary TraFiC-mem output is a standard VCF[122] file containing the coordinates for all somatic MEI calls plus comprehensive annotations, including family, subfamily, insertion length, conformation, orientation, size of the target site duplication or deletion, gene annotation, number of supporting reads, and consensus sequences spanning the breakpoint junctions. Additional information is provided for L1-mediated transductions, which includes the transduced sequence length, the genomic position of the source element, and source element identifier. Filtered MEI candidates are also reported together with the list of failed filters.

C4.2 L1-mediated deletion search algorithm

As no computational method was available for the detection of L1-mediated deletions (see section C3.1) from sequencing data, we had to implement a new specialized algorithm. The method takes as input unpaired clusters derived from TraFiC-mem discordant read-pair module, CN calls and tumour plus matched-normal BAM files. Then, it integrates discordant read-pair with read depth information to search for CN losses connecting two distal unpaired discordant clusters, supporting a potential L1 insertion associated with a loss of DNA. Two complementary approaches are used. The first strategy exploits CN calls to find instances where a CN loss connects two unpaired clusters with opposite orientations located at its ends. While this method is particularly sensitive for detecting large L1-mediated deletions, it typically fails for deletions shorter than 100Kb. Therefore, a more sensitive approach for short L1-

mediated deletions was implemented.

Positive and negative unpaired clusters located less than 100 Kbp apart are inspected for drops of coverage connecting them. As MEI insertions are typically associated with increases in the coverage due to target site duplications, the adjacent 300 bp around each cluster are excluded from all read-depth calculations. Tumour and matched-normal read-depth ratios are computed for non-overlapping 500bp sliding windows spanning the interval between both clusters. Then, the observed read-depth ratio distribution is compared to a null distribution to search for significant drops in coverage. Null distributions are obtained per individual tumour sample by randomly sampling 100,000 genomic windows of 500bp in size, drawn from CN segments that have the predominant CN in that particular sample. Then, non-parametric P-values are calculated by comparing the observed read depth ratios with the expectation according to the null distribution, which are adjusted via Benjamini–Hochberg multiple-testing correction. Unpaired clusters in opposite orientations linked by a CN drop with an adjusted P-value under 0.1 are selected as candidate L1-mediated deletions.

Candidate events are subjected to a second filtering round where each cluster is assessed for coverage drops by comparing the two 2,500 bp windows located upstream and downstream relative to the cluster position. Again, non-parametric P-values are corrected for multiple-testing by Benjamini–Hochberg. Candidate L1-mediated deletions are classified at decreasing order of confidence in three tiers: 1) candidate events with both clusters associated with significant internal drop of coverage (P-value < 0.1); 2) only one cluster with a significant coverage drop; and 3) no cluster with a significant drop. L1-mediated deletion candidates from tier 1 and 2 are selected and further manually inspected using the Integrative Genomics Viewer[128]. As described for standard L1 insertions, L1-mediated deletions are classified as solo-L1, orphan or partnered transductions based on the read composition of the two clusters flanking the genomic DNA loss. In a last step, TraFiC-mem breakpoint analysis module is used to reconstruct both breakpoints for each L1-mediated deletion event, resolve the conformation of the L1 or transduction bridge and identify retrotransposition insertion hallmarks, including the presence of Poly(A) tails and inversion at their ends.

C4.3 Processed pseudogene detection

PSD were detected through another tailored approach that relies on the same principles used by TraFiC-mem for the identification of somatic MEI events. In addition to the criteria described at the discordant read-pair module of TraFiC-mem, the algorithm for pseudogene detection requires the non-anchor reads to align uniquely to annotated exons of the same protein-coding gene based on GENCODE[174] v19 annotation. To distinguish between genuine PSD insertions and genomic rearrangements involving coding regions, the method uses TraFiC-mem breakpoint analysis module to reconstruct insertion breakpoints and search for retrotransposition hallmarks, including poly(A) tails and target site duplications. Candidate PSD insertions without companion poly(A) tracts are discarded.

C4.4 Computational methods validation

TraFiC-mem evaluation in synthetic cancer genomes

We evaluated the precision and recall of TraFiC-mem through the reanalysis of the synthetic cancer genome data previously generated for the evaluation of TraFiC[44]. A total of 10,000 previously detected L1 insertions[44], including solo, partnered and orphan transductions were embedded in the human reference (build hg19) at random locations using BedTools[175] v.2.25.0, of which 773 were excluded as they were sampled within assembly gaps. Then, ART[176] (version MountRainier-2016-06-05) was used to generate paired-end sequencing reads for the human reference and the synthetic genome containing L1 insertions at a read-depth of 38x. The two generated FASTQ files were aligned into the reference using BWA-mem[116] v.0.7.17 and the resulting BAM files were subsampled and merged at different proportions to generate 4 BAM files containing somatic L1 insertions at 25%, 50%, 75% and 100% levels of clonality. These were processed with TraFiC-mem to call MEIs, using the BAM file derived from the reference as matched-normal in every case. TraFiC-mem insertion calls were intersected with the list of simulated MEIs (i.e., ground truth) using a breakpoint offset of 50bp to determine the number of true-positive (TP), false-positive (FP), true-negative (TN) and false-negative (FN) events for every clonality. Precision and recall were computed for each VAF as follows: Precision = TP/ (TP + FP); Recall = TP/ (TP + FN).

TraFiC-mem precision and recall were higher than 95% and 90% for all the assessed VAF and L1 insertion classes (Figure 13a), respectively. While no differences were observed among insertion classes in terms of recall, TraFiC-mem average precision was 95% for partnered transductions, increasing to 99% and 100% for orphan transductions and solo insertions, respectively. We further assessed the accuracy of TraFiC-mem annotations through the comparison of the predicted insertion lengths and orientations with the expectations based on the simulations. Inferred and expected lengths strongly correlated (Spearman rho = 0.93; P-value = 0.0), while the insertion orientation was consistent in 99% of the cases (Figure 13b-c).

TraFiC-mem evaluation in cancer cell lines

Due to the unavailability of DNA for PCAWG specimens, we used NCI-H2087, a lung-cancer cell line known to have high numbers of L1 insertions[44], to evaluate TraFiC-mem in real data. We employed short-read data[44] previously generated for NCI-H2087 and NCI-BL2087, a matched-normal lymphoblastoid cell line from the same patient, and generated ONT data for both cell lines (see section MT.4). Short-read alignments were processed with TraFiC-mem to call somatic retrotranspositions present at the cancer cell line but that were absent in the matched-normal. Then, ONT data was used as an orthogonal line of evidence to validate the resulting 308 candidate retrotranspositions identified by TrafiC-mem.

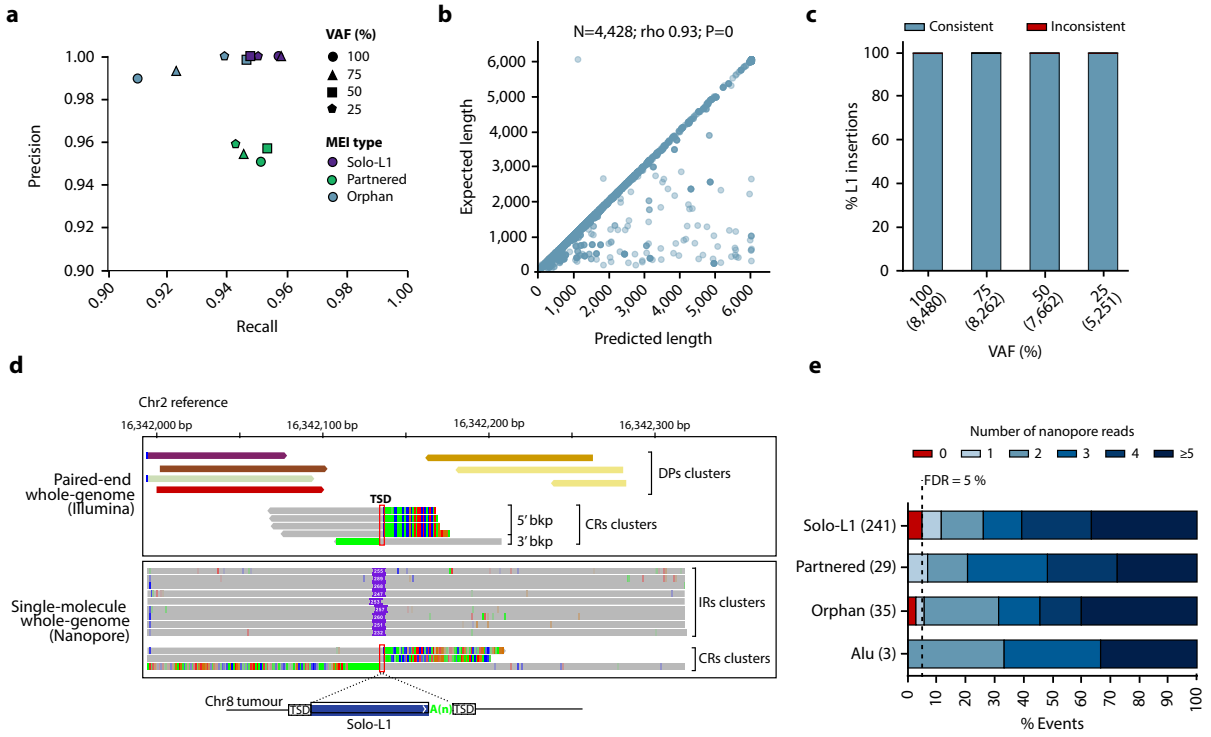


Figure 13. Validation and evaluation of TraFiC-mem. (a) Precision and recall of TraFiC-mem based on 10,000 simulated L1 insertions at multiple levels of clonality. **(b)** Correlation (Spearman) between the observed and expected lengths for 8,025 Solo-L1 insertions simulated in-silico. **(c)** Fraction of true positive Solo-L1 events with a predicted orientation consistent (green), and inconsistent (red), with the expected. Orientation consistency was assessed for four clonality levels ranging from 25 to 100. **(d)** Retrotransposition breakpoint validation approach using long-reads with ONT. Illustrative example of a Solo-L1 insertion in cancer cell line NCI-H2087 detected with short and long-reads. Top, discordant read-pairs (DPs) and clipped-reads (CRs) supporting a Solo-L1 insertion from Illumina paired-end data. Bottom, spanning-reads (SRs) and clipped-reads confirmation using ONT. **(e)** Proportion of events supported by different numbers of long-reads (from zero to more than 5 reads). Events supported by at least one long-read and absent in the matched-normal sample were considered true positives (i.e., somatic), while those not supported by ONT and/or present in the matched-normal sample were considered false positives. The total number of events assessed for each retrotransposition category is shown in parenthesis.

For each somatic retrotransposition event identified with TraFic-mem, we leveraged the ONT data to seek for orthogonal long-read validation. Two types of read signatures were considered as a source of support (Figure 13d), namely (i) “spanning-reads”, consisting on reads spanning both insertion breakpoints, so the MEI is identified as an insertion in the reference; and (ii) “clipped-reads”, which span a single insertion end, so they get clipped during their alignment in the reference. While spanning-reads are particularly useful to validate short insertions, the number of spanning-reads decreases with the insertion length, with long insertions being typically supported by clipped-reads alone. MEIs supported by at least one ONT read in the tumour and no support in the matched-normal sample were considered TP, while FP, otherwise. False discovery rate (FDR) for every retrotransposition category, including L1, Alu, partnered and orphan L1-transductions was computed as follows: FDR = FP / (TP + FP). TraFiC-mem displayed FDR under 5% for all retrotransposon classes (Figure 13e). All partnered L1 transductions and Alu insertions were validated, while only 12 solo-L1 and one orphan L1

insertions had no long-read support. Given the low ONT coverage available (i.e., 9.17X), we cannot exclude the possibility that these are genuine somatic events that were not sequenced due to insufficient coverage.

L1-mediated rearrangements validation in cancer cell lines

Two lung cancer cell lines previously reported to have high L1 retrotransposition rates[44] (NCI-H2009 and NCI-H2087) were used to validate L1-mediated rearrangements. Publicly available short-read data for both cell lines[44] was processed with TraFiC-mem to search for L1-mediated rearrangements, which led to the identification of 16 L1-mediated deletions, one L1-mediated translocation and 3 independent L1 breakends associated with a CN change. Two complementary validation approaches were used to confirm and resolve the structure of the events: (i) PCR amplification of the rearrangement breakpoints coupled with the sequencing of PCR amplicons through ONT and (ii) ONT whole genome sequencing.

For each candidate L1-mediated rearrangement, PCR-primers were designed to amplify both the insertion breakpoints and the insertion target site sequence (see section MT.4). Three amplicons with the expected molecular size corresponding to L, R and T were obtained for all the 16 L1-mediated deletions and the translocation, validating their presence in the tumour (Figure 14). Occasionally, two bands were obtained for the target site (T), corresponding to the reference and alternative alleles. L and R amplicons were absent in the normal, while primers targeting the target site (T) amplified the reference allele in those instances at which the deletion is short. As expected, for the 3 breakends a single amplicon was obtained in the tumour but not in the normal DNA. PCR amplicons were isolated, pooled and sequenced at high coverage (>1.000x) with ONT. Long-read data was aligned on the reference genome and clipped-reads were used to validate the predicted insertion breakpoints for each L1-mediated rearrangement.

In order to completely resolve the structure of L1-mediated rearrangements, both lung-cancer cell lines (NCI-H2009 and NCI-H2087) were further subjected to low depth ONT WGS (see section MT.4). Long-reads were able to completely span and resolve the structure for four L1-mediated rearrangements (Figure S9), including three deletions and a translocation. The first event corresponded to a 642 bp deletion associated with a 1.1 Kbp L1 insertion. The L1 was truncated at its 5' end, while a poly(A) tail and the L1-EN target motif was present at the 3' breakpoint, confirming a TPRT origin. The second was a 2.6 Kbp deletion bridged by a partnered transduction derived from a source L1 element at chromosome 3. The observed deletion originated via TPRT as both the poly(A) tail and the L1-EN motif were present. The third deletion was a 1.5 Kbp deletion which differed from previous events as it was associated with a 1.3 Kbp L1 insert that was truncated at both ends. No L1-EN motif was found at the insertion breakpoints, indicating that L1-mediated rearrangements can occasionally originate through an endonuclease independent insertion mechanism[135]. The last rearrangement corresponded to a translocation between chromosomes 1 and 8, which was associated with a 186bp orphan transduction bridge derived from a source element at chromosome 6.

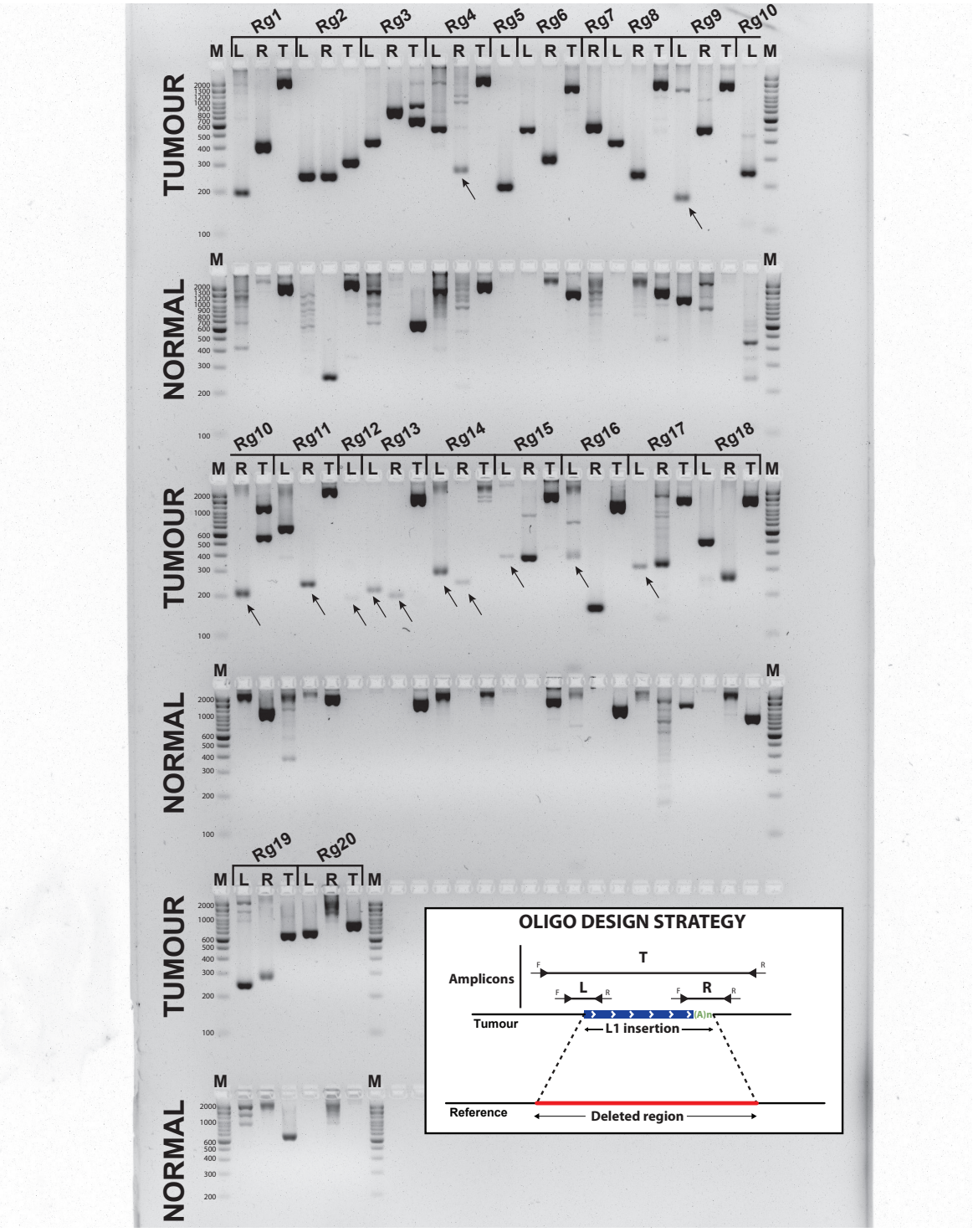


Figure 14. PCR validation of somatic L1-mediated rearrangement calls. Gel showing PCR results on cancer cell lines (NCI-H2009 and NCI-H2087) and their matched-normal cell lines (NCI-BL2009 and NCI-BL2087). We performed validation of 20 L1-mediated rearrangements (for details, see Supplementary Table 7): 16 L1-mediated deletions (Rg1, Rg2,Rg3, Rg4, Rg6, Rg8, Rg9, Rg10, Rg11, Rg13, Rg14, Rg15, Rg16, Rg17, Rg18, Rg19), 1 L1-mediated translocation (Rg20) and 3 independent L1 breakpoints associated with a CN change from an unknown rearrangement type (Rg5, Rg7, Rg12). For each rearrangement, except those where only one breakpoint is known, at least three regions were amplified in the tumours: left breakpoint (L), right breakpoint (R), and the target site (T). Arrows are used to highlight the position of some somatic amplicons. Note that the target site could also amplify in the matched-normal sample if the deletion is not too long. “M” denotes the size marker. For illustrative purposes, the oligo design strategy is shown in a panel at the bottom of the figure: amplicons (L, R and T) and oligos – forward (F) and reverse (R) – are represented.

The non-repetitive transduced sequence ended in a poly(A) tail and contained the L1-EN motif at its 3' insertion breakpoint, both TPRT hallmarks. Hence, the analysis of the complete rearrangement structures described above confirmed that L1 can mediate different rearrangement classes via the canonical TPRT pathway and occasionally through an endonuclease independent mechanism.

C4.5 Contributors

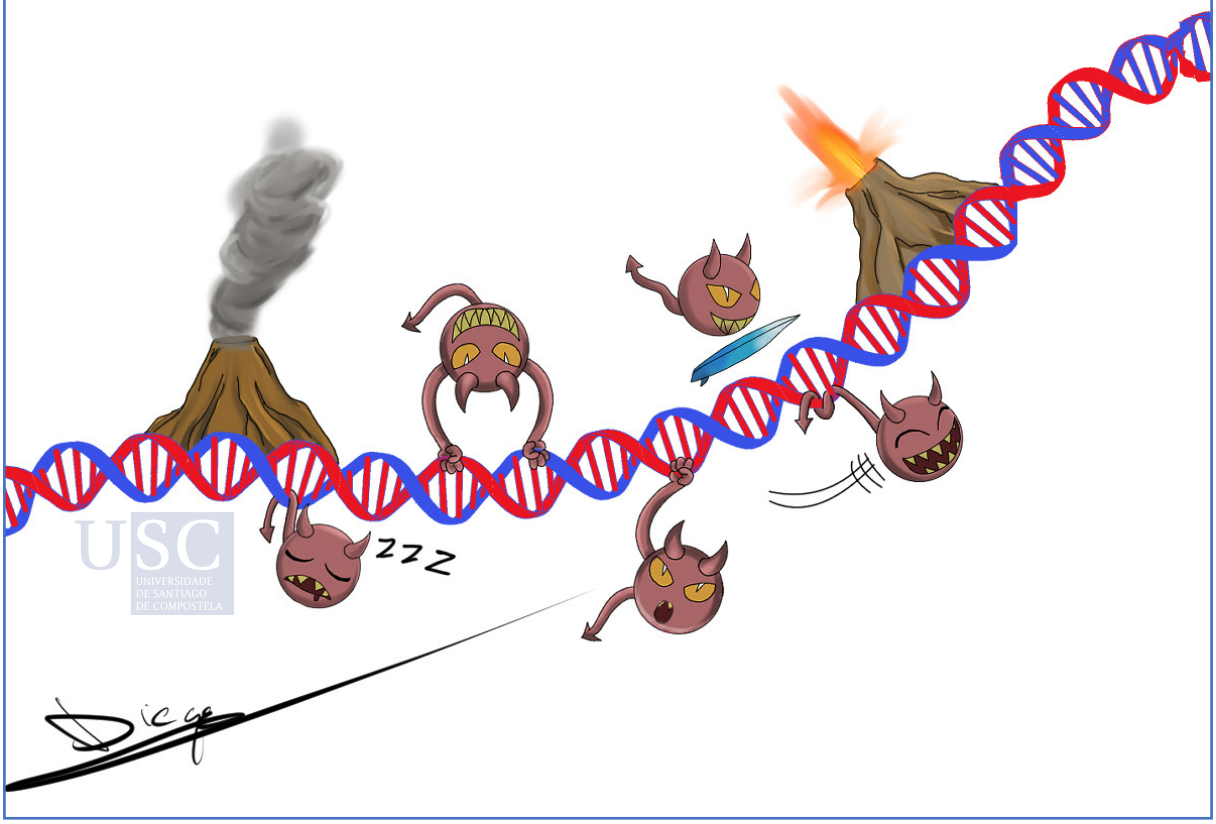
José M.C. Tubio implemented the discordant read-pair analysis module of TraFiC-mem and the bioinformatic pipelines to detect PSD insertions and L1-mediated deletions. Harald Detering contributed to the development of TraFiC-mem breakpoint analysis module. Adrian Baez-Ortega implemented components of the computational approach for L1-mediated rearrangements detection. Yilong Li and Jorze Zamora performed the simulations for the evaluation of TraFiC-mem. Javier Temes, Daniel Garcia-Souto and Jorge Rodriguez-Castro generated and managed the ONT data used for orthogonal validation of TraFiC-mem calls and L1-mediated rearrangements. Martin Santamarina and Jorge Rodriguez-Castro designed and performed PCR experiments for the validation of solo-L1 and L1-mediated rearrangements.

C4.6 Publications

This chapter's content is part of the online methods and supplementary information of a published manuscript [1]. The text was considerably rewritten, extended and adjusted to fit the flow of this thesis and provide further details. All figures included derive from supplementary figures of the aforementioned publication. The complete list of authors and their affiliations is provided in the appendix.

[1] Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat Genet 52, 306–319 (2020). DOI: https://doi.org/10.1038/s41588-019-0562-0, ISSN: 1061-4036

DISCUSSION



From my own perspective, the ultimate objective of a PhD dissertation must be to substantially expand the knowledge in a particular research topic, and to provide new hypotheses that will serve to guide future research studies. In this section, I will discuss (1) how this thesis has provided new insights on the activity patterns and impact of retrotransposons in the cancer genome; and (2) what are some of the potential research avenues to be addressed by future studies.

D.1 Oncogenic potential of somatic L1 activity

This PhD dissertation provides new insights on a long-standing question: Is the activation of endogenous retrotransposons relevant in human oncogenesis? Previous evidence was sparse, being restricted to two somatic L1 insertions disrupting the coding sequence of the tumour suppressor gene *APC* in colorectal cancer[34,51]. Nonetheless, now we know that the aberrant integration of L1 sequences can lead to diverse forms of chromosomal rearrangements in human cancers, including deletions, duplications and translocations. L1 retrotransposition can also initiate BFB-cycles, resulting in rounds of genomic instability during subsequent cell divisions[157,158].

We only detected 98 L1-mediated rearrangements, including 90 deletions, six BFB, one translocation and one duplication. While this may suggest that the generation of genomic rearrangements due to the aberrant integration of L1 sequences is an infrequent mutational process in cancer, L1-mediated rearrangements are likely to occur more frequently than we could unambiguously characterize here. The sizes of DNA fragments at illumina sequencing libraries are often too short to transverse the complete L1 insert, making a potentially large fraction of L1-mediated rearrangements undetectable.

Emerging long-read sequencing technologies, namely ONT and Pacific Biosciences, are able to produce reads averaging around 10 Kbp in length[177]. As a consequence, they should provide a more comprehensive picture about the frequency of L1-mediated rearrangements in cancer. We successfully applied low-coverage ONT sequencing to validate and sequence-resolve a subset of L1-mediated rearrangements, including three deletions and a translocation, which may serve as a proof of principle for future studies. Although long-read sequencing platforms are still subjected to limitations that prevent their application beyond the sequencing of a handful of cancer genomes, recent technological advances have increased yield and reduced sequencing error rates[177]. Further improvements in the quality of long-read sequencing data and drop costs may enable long-read based surveys of structural variation, including mobile elements, across hundreds of cancer genomes in the near future. This will require the development of specialized computational methods for the identification of retrotransposon insertions and L1-mediated rearrangements from long-read data.

D.2 Somatic retrotransposition during the life history of cancer

This PhD dissertation includes the largest genomic study of somatic retrotransposition in cancer to date. Through the analysis of a large cohort of cancer whole genomes, we provided a comprehensive portrait of L1 activity rates across cancer types, finding particularly high numbers of L1 mobilizations in esophageal, lung, head-and-neck and colon cancers. Meanwhile, L1 retrotransposons remain predominantly silent in blood, bone and brain cancers.

Sequencing of the cancer genome provides a snapshot of the mutational profile of a tumour at the time of biopsy. However, cancer development is an evolutionary process, with somatic mutations occurring during the evolution of the tumour[6]. As most cancer genome surveys have focused on primary tumours, it is not clear what is the timing for somatic retrotransposition during the course of cancer. Although L1 mobilization can be seen as a by-product of an increased genomic instability acquired during tumour development, there is evidence supporting early retrotransposition activity during oncogenesis. In 1992, an exonic somatic L1 insertion was reported to disrupt the tumour suppressor *APC* in a colon cancer patient[34], which is classically the first cancer driver hit in colon cancer progression. In addition, whole genome analysis of Barret´s oesophagus revealed somatic L1 activity in this type of precancerous lesion[45].

Genomic analysis for multiple biopsies of prostate and lung cancer patients suggested increased levels of somatic L1 retrotransposition in the later stages of cancer, with metastatic and invasive tumours having higher number of somatic insertions than primary cancers. In addition, in a recent publication I co-authored, we reported increased L1 activity in prostate cancer cell lines after the treatment with carboplatin and enzalutamide[178], two widely used chemotherapeutic agents. These observations are of potential clinical relevance, as high levels of retrotransposition at the later stages of tumorigenesis or upon chemotherapy may fuel tumour genome evolution, leading to potential phenotypic changes, including the acquisition of drug resistance or the ability to metastasize at distal tissues. The future study of L1 retrotransposition in longitudinal datasets, which include tumour biopsies collected at multiple time points and body locations, will be necessary to investigate these questions.

The research community is currently moving towards the systematic identification of somatic mutations in healthy tissue samples and pre-cancer states. Most of these studies have applied bulk sequencing for single-cell derived colonies or laser microdissection of healthy tissues, uncovering substantial amounts of somatic variation in tissues from an ample variety of body locations, including skin[179], colonic crypts[180], lung[181], bladder[182] or placenta[183]. Alternative approaches involve the application of single-cell sequencing methods, which are conceptually very powerful, but still subjected to major technical limitations[184]. Although the studies mentioned above have primarily focused on the detection of small forms of genetic variation, the datasets produced constitute an excellent opportunity to investigate the frequency of somatic retrotransposition across diverse normal tissues and relate those with the frequencies observed across cancer types.

D.3 Source L1 elements and hot activity patterns

This PhD thesis has expanded the catalogue of known active L1s in the human genome and has provided new perspectives regarding the activity patterns for hot-L1s. Through the systematic detection of L1 transductions in the PCAWG cohort, we uncovered 114 source L1s active in cancer, including 54 copies that were not previously known to be active.

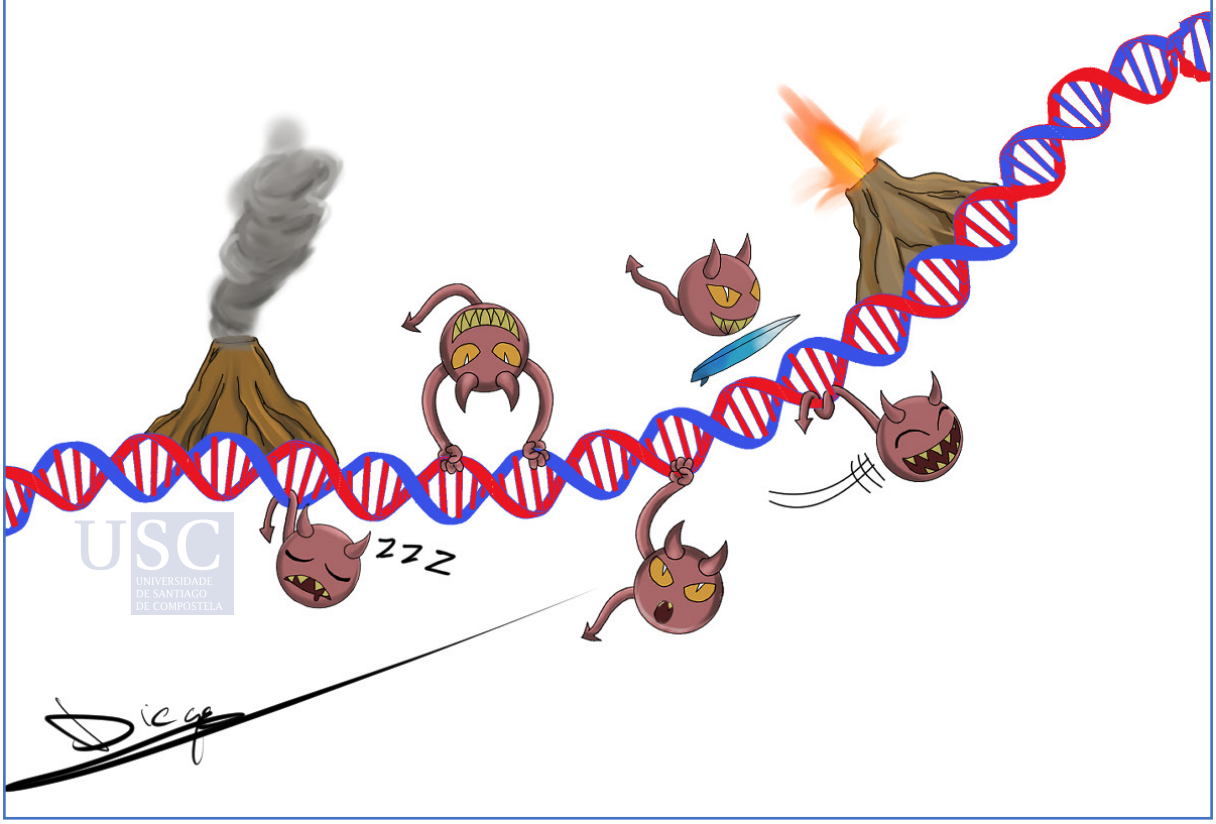
Nonetheless, the use of transductions for the discovery and characterization of active L1 copies is subjected to some limitations. First, the identification of L1 transductions relies on the search for discordant read-pairs connecting the integration point with the region located downstream of a source L1 element[44]. As a consequence, transductions spanning repeated sequences will be difficult to detect, since short-reads will map to multiple genomic positions in the human reference instead of downstream of the source L1. Second, given the abundance of repetitive sequences in the human genome[31], a substantial amount of active L1s is likely to be flanked by other repeats, making them undetectable through transduction tracing. Third, some active L1s may harbour strong polyadenylation signals, preventing them from generating transductions via transcriptional read-through.

Long-read sequencing may serve to overcome some of these limitations, providing a more comprehensive view of the collection of L1s active in the cancer genome. Kilobase-long reads are able to completely resolve the sequence for retrotransposition insertions, including transductions, potentially enabling the detection of short transduction events and transductions spanning repetitive sequences. In addition, although L1 copies typically display high levels of sequence homology between each other, they can be frequently differentiated based on internal nucleotidic changes that are specific for subsets of L1 sequences[170]. Diagnostic nucleotide positions have been traditionally used for taxonomic purposes, being the basis for the current classification of the human specific L1 lineage into 4 subfamilies[170]. However, diagnostic single nucleotide variants occurring within the sequence of active L1s can also potentially be used to infer the progenitor copy of somatic retrotransposition insertions[51].

Based on this principle, Scott and colleagues were able to map an oncogenic somatically acquired L1 insertion disrupting *APC*[51] to a hot source loci located at chromosome 17. However, their approach to resolve the sequence for L1 insertions was laborious, as involved PCR, cloning and Sanger sequencing of L1 inserts, preventing its application beyond a handful of cancer genomes. Meanwhile, due to limitations in read-length and insert size, the internal sequence for L1 insertions has remained refractory for short-read sequencing. Although long-reads are able to span the complete sequence for L1 inserts, they are subjected to high error rates, preventing reliable inferences. However, the recent development of circular consensus sequencing, a new protocol for Pacbio sequencing which provides highly accurate (99.8%) long reads[185], is likely to enable major advances in this line of research.

The bulk of somatic transductions identified in the PCAWG dataset derived from a limited collection of hot-L1s, which is consistent with a previous study based on retrotransposition assays[59]. Hot source loci displayed a dichotomous pattern of activity and allele frequencies, which resemble volcano eruption types. Similarly, as described for pathogens, such as viruses and bacteria, the observed patterns may be the consequence of the process of coevolution between hot-L1s and the human genome. Plinian L1s may represent recently acquired hot-L1s, which have not yet reached an equilibrium with our species, displaying low allele frequencies but occasionally leading to bursts of somatic retrotransposition in cancer. In contrast, Strombolian elements are likely to be older copies which may have attenuated their activity to coexist without being detrimental to the host (i.e., the human genome), resulting in high allele frequencies and prevalent but moderate levels of somatic retrotransposition. In a recent publication I co-authored[178], we leveraged PacBio assemblies for 64 human haplotypes to resolve the complete sequence for eight Strombolian and two Plinian L1s, in addition to 319 FL-L1s without hot activity in the PCAWG dataset. We used phylogenetic methods to estimate their age in million years (Myr), confirming that Plinian elements (mean = 0.32 Myr) have been more recently acquired than Strombolian copies (mean = 0.97 Myr). Indeed, two out of the three youngest active L1s, namely 2q24.1 (0.20 Myr), and 6p22.1-2 (0.45), display Plinian activity. In contrast, 1p12 is a Strombolian copy that, despite integrating into the human genome 1.8 Myr ago, remains frequently active in cancer. Overall, this data further supports the hypothesis discussed above and indicates that highly active L1s have been recently acquired in the human genome, representing a potential source for cancer risk owing to their mutagenic capabilities. Current initiatives to generate large-scale resources, including whole genome sequences and clinical information for thousands of patients, such as the UK Biobank[186], will enable the scientific community to investigate the potential relationship between hot-L1s and cancer risk through genome-wide association analysis.

CONCLUSIONS



Chapter 1

Extensive heterogeneity in the rates of somatic retrotransposition across tumour types

There are large differences in the prevalence of somatic retrotransposition insertions across cancer types. Esophageal, lung, head-and-neck and colon cancers display particularly high levels of retrotransposition activity, with mobile element insertions representing a predominant SV class in these tumour types. Other adenocarcinomas, such as those arising from the stomach, pancreas, breast, uterus, ovary, cervix and prostate, have moderate levels of retrotransposition. Meanwhile, skin, bone, brain and blood cancers have low levels of somatic retrotransposition. *TP53* mutation is associated with increased levels of somatic L1 insertions, which may contribute to the observed differences in the number of L1 insertions between tumours.

Somatic retrotransposition occasionally alter gene expression and splicing

The somatic insertion of retrotransposons within gene boundaries can lead to occasional gene expression or splicing alterations. Splicing aberrations are driven by the exonization of intronic insertions, which includes PSD. The exonization of an L1 insertion within the second intron of the tumour suppressor *RB1* leads to the expression of multiple *L1-RB1* fusion transcripts and is associated with increased levels of *RB1* expression.

Multiple genomic features shape the distribution for somatic L1 insertions

The genome-wide distribution of somatic L1 insertions through the cancer genome is highly heterogeneous. The density of L1-EN motifs and replication timing are major factors shaping the observed distribution, with multiple genomic features (i.e., DNA accessibility, promoters, enhancers and gene expression) having a more moderate effect. These patterns are consistently observed across different tumour types.

Chapter 2

Hot source L1 elements account for the bulk of somatic retrotransposition in cancer

A total number of 114 germline source L1s, including 54 novel copies, are active in the PCAWG dataset. Somatic retrotransposition is predominantly driven by a small set of 16 hot elements. Hot loci display a dichotomous pattern of activity and allele frequencies, which resembles volcano eruption classes.

High retrotransposition rates are driven by the cumulative activation of multiple L1s

The number of active L1 per cancer genome is heterogeneous and correlates with the number of detected L1 retrotranspositions, indicating that the cumulative contribution of multiple somatically active L1 source elements drives the high retrotransposition rates observed in some tumours.

Chapter 3

Aberrant L1 retrotransposition can mediate genomic rearrangements with occasional oncogenic consequences

The aberrant integration of L1 sequences can lead to diverse forms of genomic rearrangements, which include deletions, duplications, translocations and the initiation of BFB-cycles. The majority of L1-mediated rearrangements identified are deletions, which occasionally can be large, spanning up to 50 Mbp of the genomic sequence. L1-mediated rearrangements are predominantly driven by target primed reverse transcription, while a minority may occur through endonuclease independent retrotransposition. Occasionally, L1-mediated rearrangements can be cancer driver events, as illustrated by the recurrent deletion of the tumour suppressor gene *CDKN2A* and the amplification of the oncogene *CCND1*.

Chapter 4

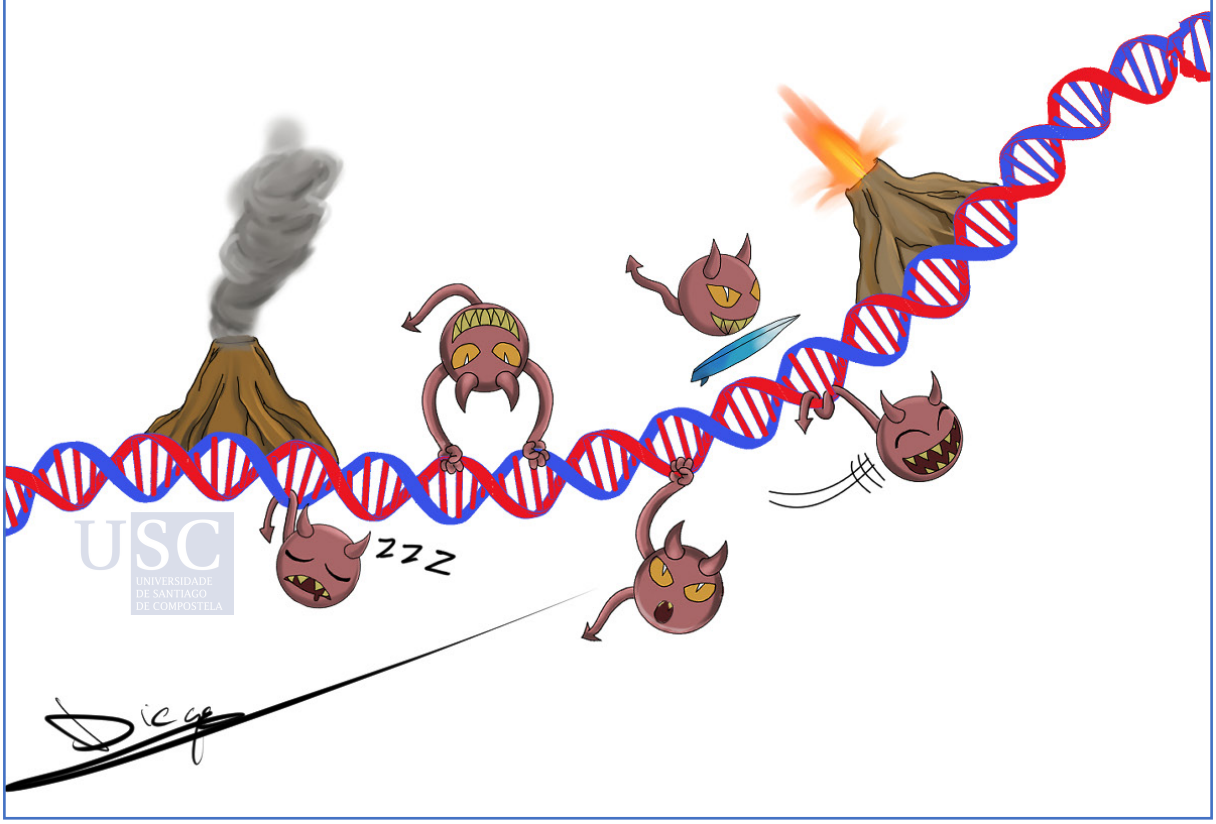
TraFiC-mem, a new algorithm for somatic retrotransposition detection

We developed TraFiC-mem, a novel computational method for the detection of somatic mobile element insertions, including L1-mediated transductions, from short-read cancer genome data. TraFiC-mem repository is publicly available in order to facilitate future studies for somatic retrotransposition in cancer (https://gitlab.com/mobilegenomes/TraFiC). In addition, two independent modules for the detection of processed pseudogene insertions and L1-mediated rearrangements were implemented.

Validation of TraFiC-mem and L1-mediated rearrangements

TraFiC-mem displays high sensitivity (>90%) and low false discovery rates (<5%) when evaluated using a simulated cancer genome containing synthetic L1 inserts, including L1-mediated transductions. Further evaluation using long-read sequencing data (i.e., ONT) confirms low false discovery rates (<5%). We validated all the L1-mediated rearrangements (i.e., 16 L1-mediated deletions and one translocation) detected at two lung cancer cell lines (NCI-H2009 and NCI-H2087) via long-reads and PCR. We also used the long-read data to resolve the complete rearrangement structure for five events, confirming for all of them the existence of an L1 or transduction bridge at the rearrangement breakpoints.

BIBLIOGRAPHY



1. Bianconi, E. *et al.* An estimation of the number of cells in the human body. Ann. Hum. Biol. 40, 463–471 (2013).

2. Mattiuzzi, C. & Lippi, G. Current Cancer Epidemiology. J. Epidemiol. Glob. Health 9, 217–222 (2019).

3. Chaffer, C. L. & Weinberg, R. A. A perspective on cancer cell metastasis. Science 331, 1559–1564 (2011).

4. Hansemann, D. Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. Archiv für pathologische Anatomie und Physiologie und für klinische Medicin 119, 299–326 (1890).

5. Calkins, G. N. Zur Frage der Entstehung maligner Tumoren. By Th. Boveri. Jena, Gustav Fischer. 1914. 64 pages. Science 40, 857–859 (1914).

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

7. Pulciani, S. *et al.* Oncogenes in human tumor cell lines: molecular cloning of a transforming gene from human bladder carcinoma cells. Proc. Natl. Acad. Sci. U. S. A. 79, 2845–2849 (1982).

8. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. Nature 378, 789–792 (1995).

9. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 47, D941–D947 (2019).

10. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. Nat. Rev. Cancer 20, 555–572 (2020).

11. Zaman, A., Wu, W. & Bivona, T. G. Targeting Oncogenic BRAF: Past, Present, and Future. Cancers 11, (2019).

12. Aubrey, B. J., Strasser, A. & Kelly, G. L. Tumor-Suppressor Functions of the TP53 Pathway. Cold Spring Harb. Perspect. Med. 6, (2016).

13. Leroy, B., Anderson, M. & Soussi, T. TP53 mutations in human cancer: database reassessment and prospects for the next decade. Hum. Mutat. 35, 672–688 (2014).

14. Nielsen, F. C., van Overeem Hansen, T. & Sørensen, C. S. Hereditary breast and ovarian cancer: new genes in confined pathways. Nat. Rev. Cancer 16, 599–612 (2016).

15. Balmus, G. *et al.* ATM orchestrates the DNA-damage response to counter toxic non-homologous end-joining at broken replication forks. Nat. Commun. 10, 87 (2019).

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

17. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human

cancer genome. Nature 463, 191–196 (2010).

18. International Cancer Genome Consortium *et al.* International network of cancer genome projects. Nature 464, 993–998 (2010).

19. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120 (2013).

20. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).

21. Feeney, J., Birdsall, B., Ostler, G., Carr, M. D. & Kairi, M. A novel method of preparing totally alpha-deuterated amino acids for selective incorporation into proteins. Application to assignment of 1H resonances of valine residues in dihydrofolate reductase. FEBS Lett. 272, 197–199 (1990).

22. Ray, M., Salgia, R. & Vokes, E. E. The role of EGFR inhibition in the treatment of non-small cell lung cancer. Oncologist 14, 1116–1130 (2009).

23. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337 (2012).

24. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of endometrial carcinoma. Nature 497, 67–73 (2013).

25. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016).

26. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. Nature 578, 112–121 (2020).

27. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. Cell 174, 758–769.e9 (2018).

28. Meric-Bernstam, F. *et al.* Advances in HER2-Targeted Therapy: Novel Agents and Opportunities Beyond Breast and Gastric Cancer. Clin. Cancer Res. 25, 2033–2041 (2019).

29. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020).

30. 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).

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

32. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10, 691–703 (2009).

33. 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).

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

35. Chalitchagorn, K. *et al.* Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. Oncogene 23, 8841–8846 (2004).

36. Phokaew, C., Kowudtitham, S., Subbalekha, K., Shuangshoti, S. & Mutirangura, A. LINE-1 methylation patterns of different loci in normal and cancerous cells. Nucleic Acids Res. 36, 5704–5712 (2008).

37. Skowronski, J. & Singer, M. F. Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proc. Natl. Acad. Sci. U. S. A. 82, 6050–6054 (1985).

38. Ardeljan, D., Taylor, M. S., Ting, D. T. & Burns, K. H. The Human Long Interspersed Element-1 Retrotransposon: An Emerging Biomarker of Neoplasia. Clin. Chem. 63, 816–822 (2017).

39. Iskow, R. C. *et al.* Natural mutagenesis of human genomes by endogenous retrotransposons. Cell 141, 1253–1261 (2010).

40. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. Science 337, 967–971 (2012).

41. Solyom, S. *et al.* Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res. 22, 2328–2338 (2012).

42. Shukla, R. *et al.* Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell 153, 101–111 (2013).

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

44. Tubio, J. M. C. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science 345, 1251343 (2014).

45. Doucet-O'Hare, T. T. *et al.* LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. Proc. Natl. Acad. Sci. U. S. A. 112, E4894–900 (2015).

46. 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).

47. Ewing, A. D. *et al.* Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res. 25, 1536–1545 (2015).

48. Rodić, N. *et al.* Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. Nat. Med. 21, 1060–1064 (2015).

49. Achanta, P. *et al.* Somatic retrotransposition is infrequent in glioblastomas. Mob. DNA

7, 22 (2016).

50. Carreira, P. E. *et al.* Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. Mob. DNA 7, 21 (2016).

51. Scott, E. C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res. 26, 745–755 (2016).

52. Scott, A. F. *et al.* Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics 1, 113–125 (1987).

53. Martin, S. L., Voliva, C. F., Burton, F. H., Edgell, M. H. & Hutchison, C. A., 3rd. A large interspersed repeat found in mouse DNA contains a long open reading frame that evolves as if it encodes a protein. Proc. Natl. Acad. Sci. U. S. A. 81, 2308–2312 (1984).

54. Kolosha, V. O. & Martin, S. L. *In vitro* properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proc. Natl. Acad. Sci. U. S. A. 94, 10155–10160 (1997).

55. Martin, S. L. Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. RNA Biol. 7, 706–711 (2010).

56. Feng, Q., Moran, J. V., Kazazian, H. H., Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87, 905–916 (1996).

57. Mathias, S. L., Scott, A. F., Kazazian, H. H., Jr, Boeke, J. D. & Gabriel, A. Reverse transcriptase encoded by a human transposable element. Science 254, 1808–1810 (1991).

58. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription *in vitro*. EMBO J. 21, 5899–5910 (2002).

59. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U. S. A. 100, 5280–5285 (2003).

60. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81 (2015).

61. Hancks, D. C. & Kazazian, H. H., Jr. Roles for retrotransposon insertions in human disease. Mob. DNA 7, 9 (2016).

62. Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F. & Kazazian, H. H., Jr. Isolation of an active human transposable element. Science 254, 1805–1808 (1991).

63. Holmes, S. E., Dombroski, B. A., Krebs, C. M., Boehm, C. D. & Kazazian, H. H., Jr. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. Nat. Genet. 7, 143–148 (1994).

64. Brouha, B. *et al.* Evidence consistent with human L1 retrotransposition in maternal meiosis I. Am. J. Hum. Genet. 71, 327–336 (2002).

65. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H., Jr. Exon shuffling by L1

retrotransposition. Science 283, 1530–1534 (1999).

66. Goodier, J. L., Ostertag, E. M. & Kazazian, H. H., Jr. Transduction of 3'-flanking sequences is common in L1 retrotransposition. Hum. Mol. Genet. 9, 653–657 (2000).

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

68. Sassaman, D. M. *et al.* Many human L1 elements are capable of retrotransposition. Nat. Genet. 16, 37–43 (1997).

69. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. Nat. Genet. 37, 727–732 (2005).

70. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. Nat. Genet. 38, 75–81 (2006).

71. Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 16, 1182–1190 (2006).

72. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56–64 (2008).

73. Beck, C. R. *et al.* LINE-1 retrotransposition activity in human genomes. Cell 141, 1159–1170 (2010).

74. Ostertag, E. M. & Kazazian, H. H., Jr. Biology of mammalian L1 retrotransposons. Annu. Rev. Genet. 35, 501–538 (2001).

75. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 13, 2541–2558 (2003).

76. Damert, A. *et al.* 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res. 19, 1992–2008 (2009).

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

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

79. 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).

80. Cost, G. J. & Boeke, J. D. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry 37, 18081–18093 (1998).

81. Kazazian, H. H., Jr & Moran, J. V. The impact of L1 retrotransposons on the human genome. Nat. Genet. 19, 19–24 (1998).

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

83. Lee, J., Ha, J., Son, S.-Y. & Han, K. Human Genomic Deletions Generated by SVA-Associated Events. Comp. Funct. Genomics 2012, 807270 (2012).

84. Vogt, J. *et al.* SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. Genome Biol. 15, R80 (2014).

85. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am. J. Hum. Genet. 97, 199–215 (2015).

86. Weir, B., Zhao, X. & Meyerson, M. Somatic alterations in the human cancer genome. Cancer Cell 6, 433–438 (2004).

87. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467 (1977).

88. Britten, R. J. & Kohne, D. E. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161, 529–540 (1968).

89. Schmid, C. W. & Deininger, P. L. Sequence organization of the human genome. Cell 6, 345–358 (1975).

90. Adams, J. W., Kaufman, R. E., Kretschmer, P. J., Harrison, M. & Nienhuis, A. W. A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. Nucleic Acids Res. 8, 6113–6128 (1980).

91. Sun, L., Paulson, K. E., Schmid, C. W., Kadyk, L. & Leinwand, L. Non-Alu family interspersed repeats in human DNA and their transcriptional activity. Nucleic Acids Res. 12, 2669–2690 (1984).

92. Fanning, T. & Singer, M. The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. Nucleic Acids Res. 15, 2251–2260 (1987).

93. Levy, J. Sequencing the yeast genome: an international achievement. Yeast 10, 1689–1706 (1994).

94. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. Science 287, 2185–2195 (2000).

95. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815 (2000).

96. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87 (2005).

97. Wang, J. *et al.* Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms. Gene 365, 11–20 (2006).

98. Konkel, M. K., Wang, J., Liang, P. & Batzer, M. A. Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies. Gene 390, 28–38 (2007).

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

100. Faulkner, G. J. Retrotransposons: mobile and mutagenic from conception to death. FEBS Lett. 585, 1589–1594 (2011).

101. 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).

102. Xing, J., Witherspoon, D. J. & Jorde, L. B. Mobile element biology: new possibilities with high-throughput sequencing. Trends Genet. 29, 280–289 (2013).

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

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

105. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420–426 (2007).

106. Ewing, A. D. Transposable element detection from whole genome sequence data. Mob. DNA 6, 24 (2015).

107. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. 40, 722–729 (2008).

108. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. Nature 526, 68–74 (2015).

109. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods 6, 677–681 (2009).

110. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333–i339 (2012).

111. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15, R84 (2014).

112. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 28, 581–591 (2018).

113. Thung, D. T. *et al.* Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. 15, 488 (2014).

114. Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 27, 1916–1929 (2017).

115. Supek, F. & Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell 170, 534–547.e23 (2017).

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

117. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).

118. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013).

119. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).

120. PCAWG Transcriptome Core Group *et al.* Genomic basis for RNA alterations in cancer. Nature 578, 129–136 (2020).

121. Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell 184, 2239–2254.e39 (2021).

122. Danecek, P. *et al.* The variant call format and VCFtools. Bioinformatics 27, 2156–2158 (2011).

123. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature 578, 102–111 (2020).

124. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer 18, 696–705 (2018).

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

126. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

127. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–45 (2016).

128. Robinson, J. T. *et al.* Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).

129. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat. Genet. 49, 1779–1784 (2017).

130. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).

131. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. Nucleic Acids Res. 40, e115 (2012).

132. Wang, H. *et al.* SVA elements: a hominid-specific retroposon family. J. Mol. Biol. 354, 994–1007 (2005).

133. Piskareva, O. & Schmatchenko, V. DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template *in vitro*. FEBS Lett. 580, 661–668 (2006).

134. Ostertag, E. M. & Kazazian, H. H., Jr. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. 11, 2059–2065 (2001).

135. Morrish, T. A. *et al.* DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat. Genet. 31, 159–165 (2002).

136. Cooke, S. L. *et al.* Processed pseudogenes acquired somatically during cancer development. Nat. Commun. 5, 3644 (2014).

137. Faulkner, G. J. & Billon, V. L1 retrotransposition in the soma: a field jumping ahead. Mob. DNA 9, 22 (2018).

138. Maura, F. *et al.* Genomic landscape and chronological reconstruction of driver events in multiple myeloma. Nat. Commun. 10, 3835 (2019).

139. Maura, F. *et al.* CDKN2A deletion is a frequent event associated with poor outcome in patients with peripheral T-cell lymphoma not otherwise specified (PTCL-NOS). Haematologica Online ahead of print, (2020).

140. Wylie, A. *et al.* p53 genes function to restrain mobile elements. Genes Dev. 30, 64–77 (2016).

141. Jung, H., Choi, J. K. & Lee, E. A. Immune signatures correlate with L1 retrotransposition in gastrointestinal cancers. Genome Res. 28, 1136–1146 (2018).

142. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell 176, 1282–1294.e20 (2019).

143. Conticello, S. G. The AID/APOBEC family of nucleic acid mutators. Genome Biol. 9, 229 (2008).

144. Futreal, P. A. *et al.* A census of human cancer genes. Nat. Rev. Cancer 4, 177–183 (2004).

145. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–218 (2013).

146. Gao, G. & Smith, D. I. Very large common fragile site genes and their potential role in cancer development. Cell. Mol. Life Sci. 71, 4601–4615 (2014).

147. Kazazian, H. H., Jr & Moran, J. V. Mobile DNA in Health and Disease. N. Engl. J. Med. 377, 361–370 (2017).

148. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 171, 1029–1041.e21 (2017).

149. Mita, P. *et al.* LINE-1 protein localization and functional dynamics during the cell cycle. Elife 7, (2018).

150. Flasch, D. A. *et al.* Genome-wide *de novo* L1 Retrotransposition Connects Endonuclease Activity with Replication. Cell 177, 837–851.e28 (2019).

151. 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).

152. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. Cell 129, 823–837 (2007).

153. Rodic, N. LINE-1 activity and regulation in cancer. Front. Biosci. 23, 1680–1686 (2018).

154. Chang, H. H. Y., Pannunzio, N. R., Adachi, N. & Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat. Rev. Mol. Cell Biol. 18, 495–506 (2017).

155. Xu, W. & Ji, J.-Y. Dysregulation of CDK8 and Cyclin C in tumorigenesis. J. Genet. Genomics 38, 439–452 (2011).

156. Bunting, S. F. & Nussenzweig, A. End-joining, translocations and cancer. Nat. Rev. Cancer 13, 443–454 (2013).

157. Bignell, G. R. *et al.* Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. Genome Res. 17, 1296–1303 (2007).

158. Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat. Genet. 52, 891–897 (2020).

159. McClintock, B. The Behavior in Successive Nuclear Divisions of a Chromosome Broken at Meiosis. Proc. Natl. Acad. Sci. U. S. A. 25, 405–416 (1939).

160. McClintock, B. The Stability of Broken Ends of Chromosomes in Zea Mays. Genetics 26, 234–282 (1941).

161. Dewhurst, S. M. *et al.* Structural variant evolution after telomere crisis. Nat. Commun. 12, 2093 (2021).

162. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 467, 1109–1113 (2010).

163. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. Nature 463, 899–905 (2010).

164. Li, Y. *et al.* Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. Nature 508, 98–102 (2014).

165. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 34, 3600 (2018).

166. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).

167. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277 (2000).

168. Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res. 12, 656–664 (2002).

169. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993 (2011).

170. Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol. Biol. Evol. 17, 915–928 (2000).

171. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18, 821–829 (2008).

172. Wang, J. *et al.* dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum. Mutat. 27, 323–329 (2006).

173. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010).

174. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774 (2012).

175. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

176. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. Bioinformatics 28, 593–594 (2012).

177. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. 19, 329–346 (2018).

178. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372, (2021).

179. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348, 880–886 (2015).

180. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. Nature 574, 532–537 (2019).

181. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial

epithelium. Nature 578, 266–272 (2020).

182. Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. Science 370, 75–82 (2020).

183. Coorens, T. H. H. *et al.* Inherent mosaicism and extensive mutation of human placentas. Nature 592, 80–85 (2021).

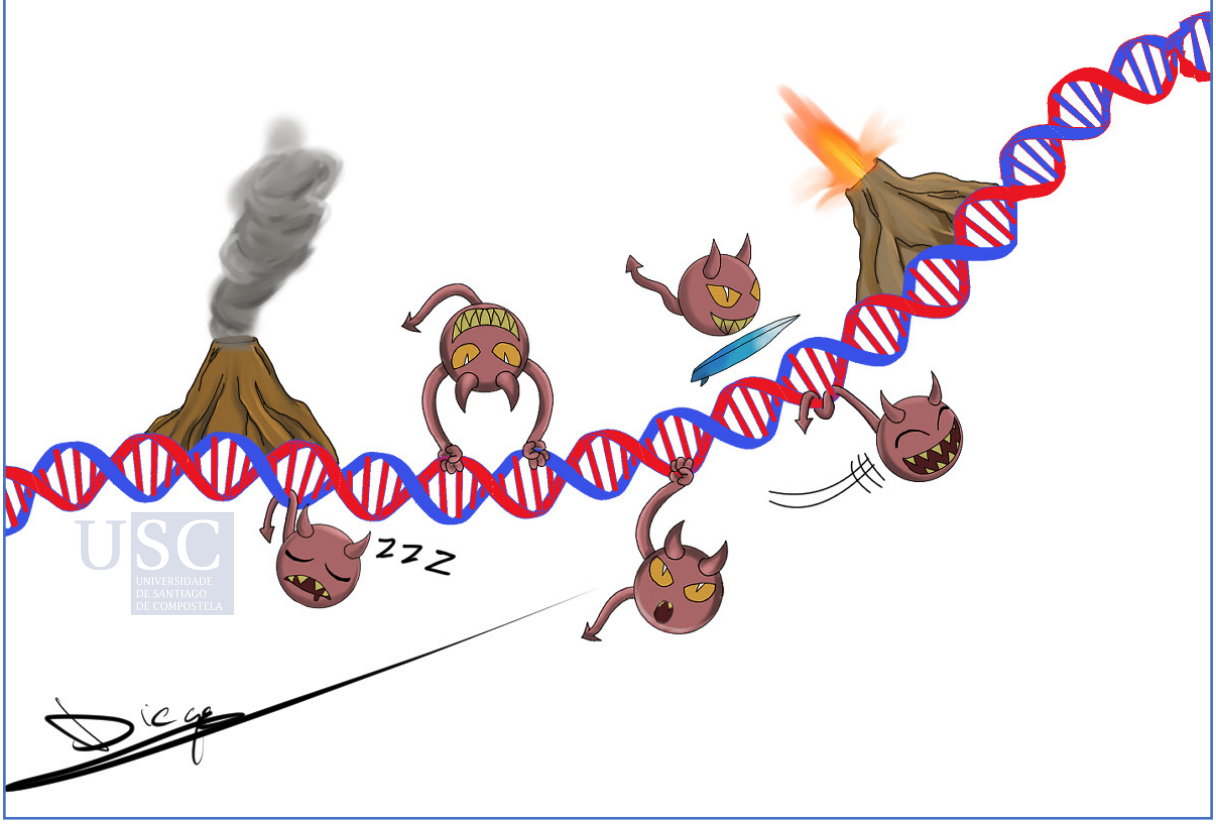
184. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. Nat. Rev. Genet. 17, 175–188 (2016).

185. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 37, 1155–1162 (2019).

186. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018).

187. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics Chapter 4, Unit 4.10 (2004).

APPENDIX



SUPPLEMENTARY FIGURES

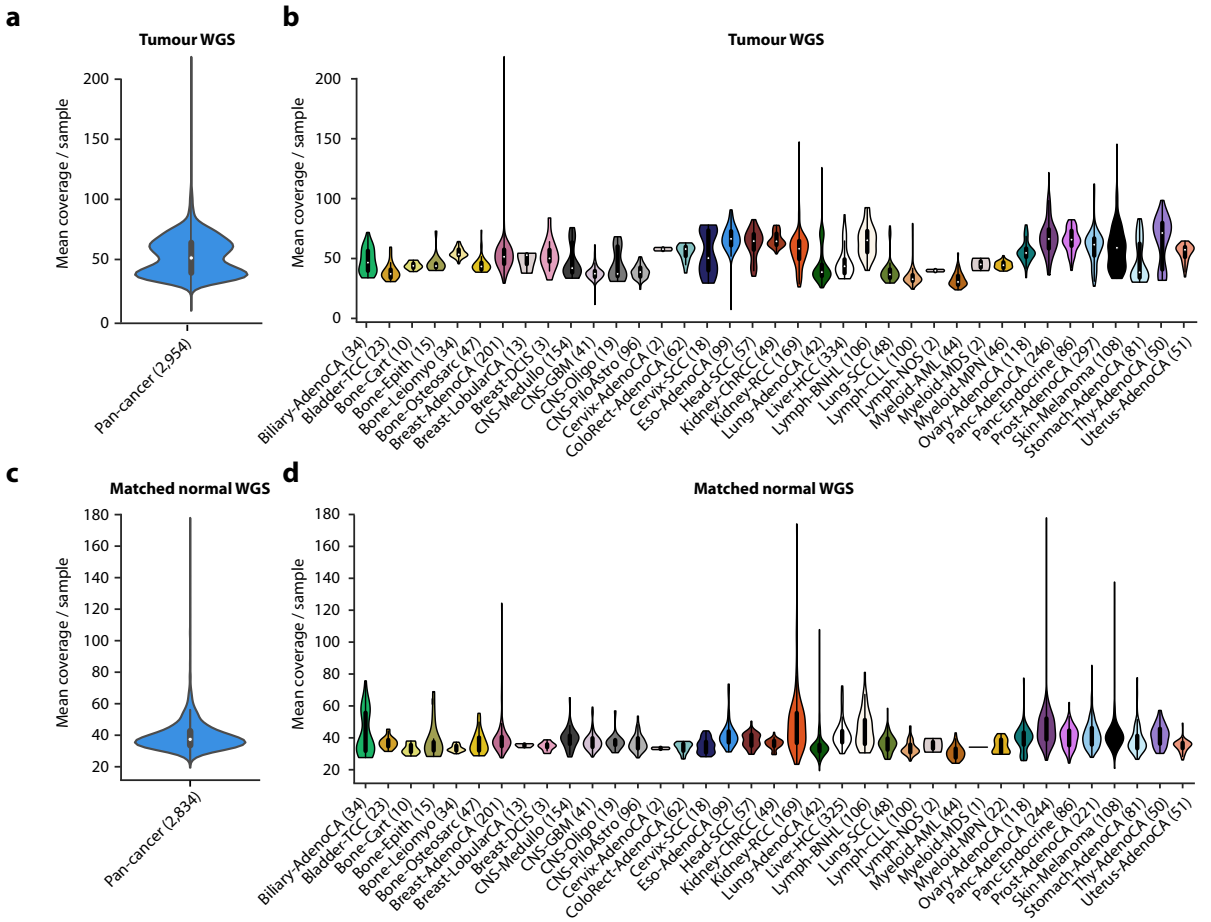


Figure S1. Coverage of whole genome sequencing data for tumours and matched-normal samples included in the PCAWG cohort. (a) Violin plot for the distribution of the mean coverage from all PCAWG tumours analyzed in this study shows a bimodal distribution with maxima at 38 and 60 reads per bp. **(b)** Distribution of the mean coverage from PCAWG tumours by cancer type. **(c)** Violin plot for the distribution of the mean coverage from all PCAWG matched-normal samples analyzed in this study shows a mean coverage of 30 reads per genome. **(d)** Distribution of the mean coverage from PCAWG matched-normal samples by cancer type.

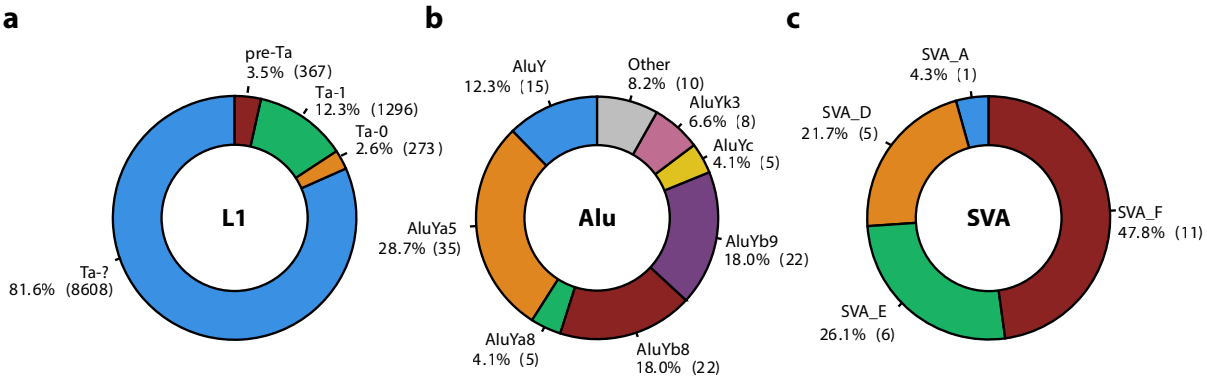


Figure S2. Distribution of somatic retrotranspositions according to subfamily. (a) L1 subfamilies based on diagnostic nucleotides[170]. The category "Ta-?" contains Ta sequences for which it was not possible to detect the Ta-0 or Ta-1 diagnostic nucleotides. **(b)** Alu subfamilies based on RepeatMasker annotation[187]. **(c)** SVA subfamilies based on RepeatMasker annotation[187].

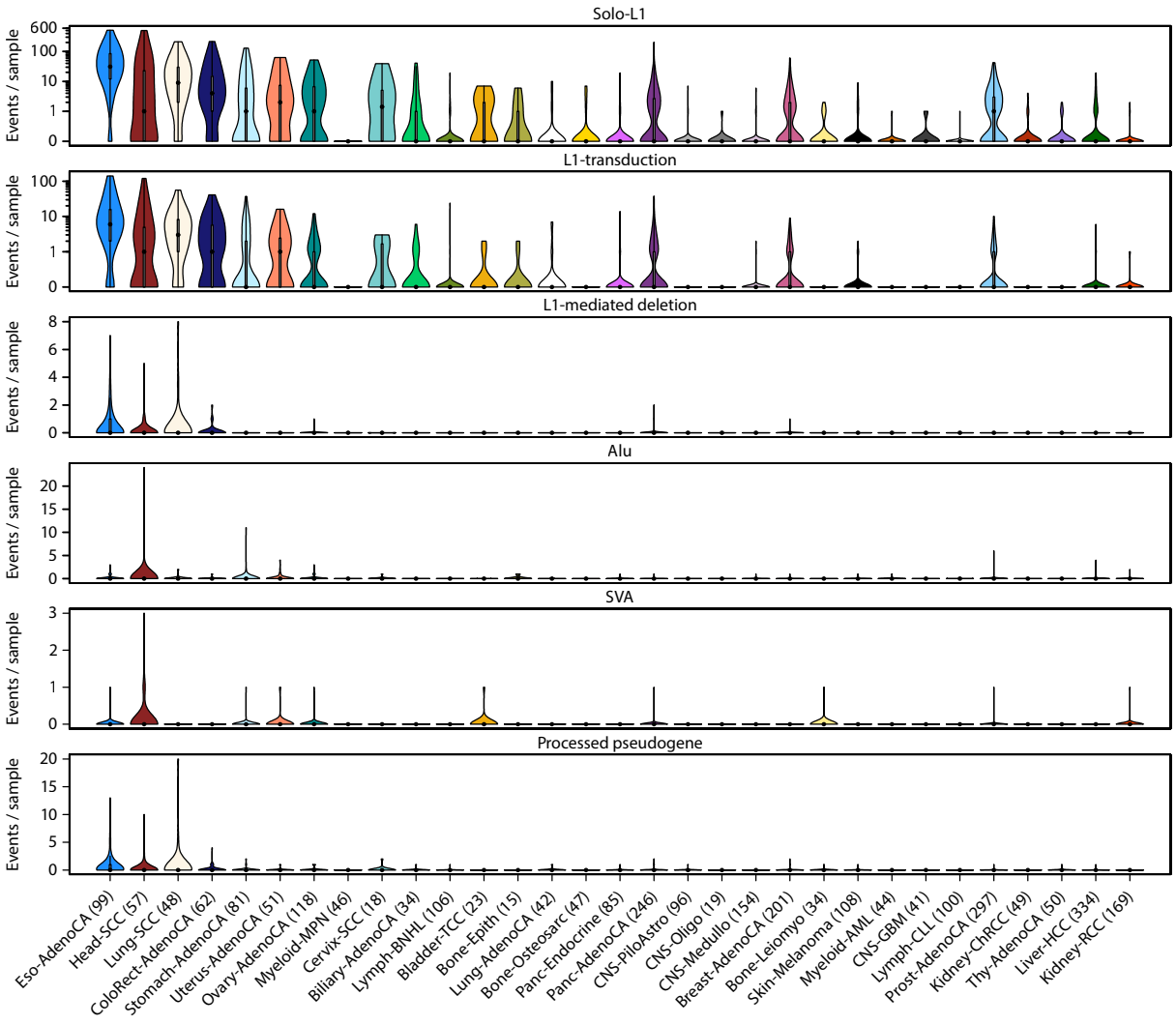


Figure S3. Rates of somatic retrotransposition across PCAWG tumour types. Violin plots showing the distributie number of retrotranspositions per sample across cancer types, for the six different categories of retrotranspositions that were analyzed (Solo-L1, L1-transductions, L1-mediated deletions, Alu, SVA and Processed pseudogenes). Y-axis is represented in a logarithmic scale.

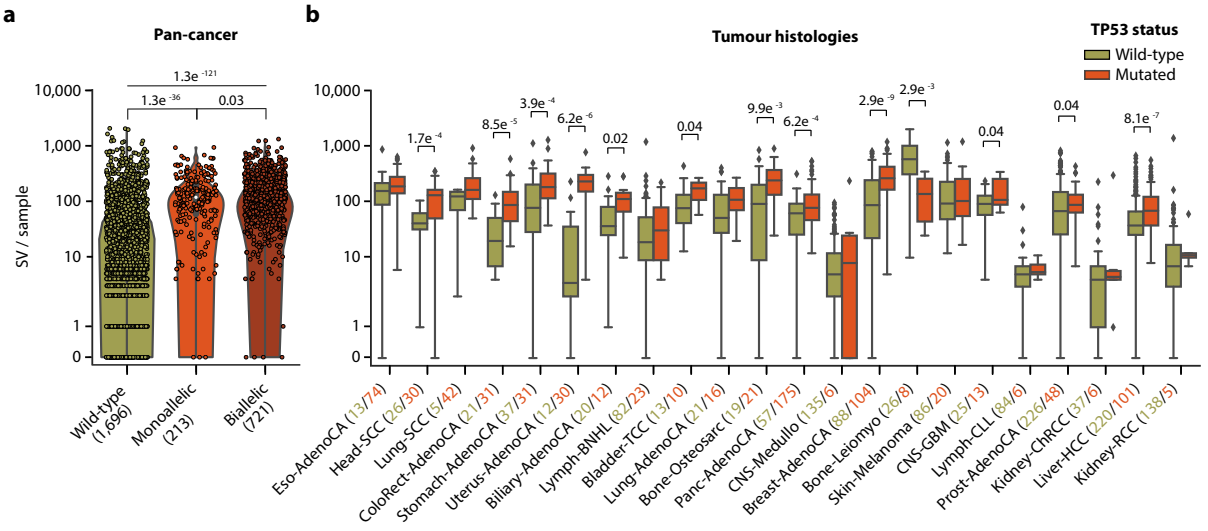


Figure S4. TP53 mutation is associated with high rates of structural variation. (a) Distribution of SV counts for three groups of samples according to their TP53 mutational status: wild-type, monoallelic and biallelic driver mutation. Each data point corresponds to one tumour sample. Groups are compared though Mann–Whitney U. **(b)** Box-and-whisker plots showing the distribution of SV counts across tumour types with samples grouped in two categories: TP53 wild-type and TP53-mutated (monoallelic or biallelic). Within a given tumour type, the two groups (wild-type and mutated) are compared using Mann-Whitney U. P-values are shown only when differences between groups are significant.

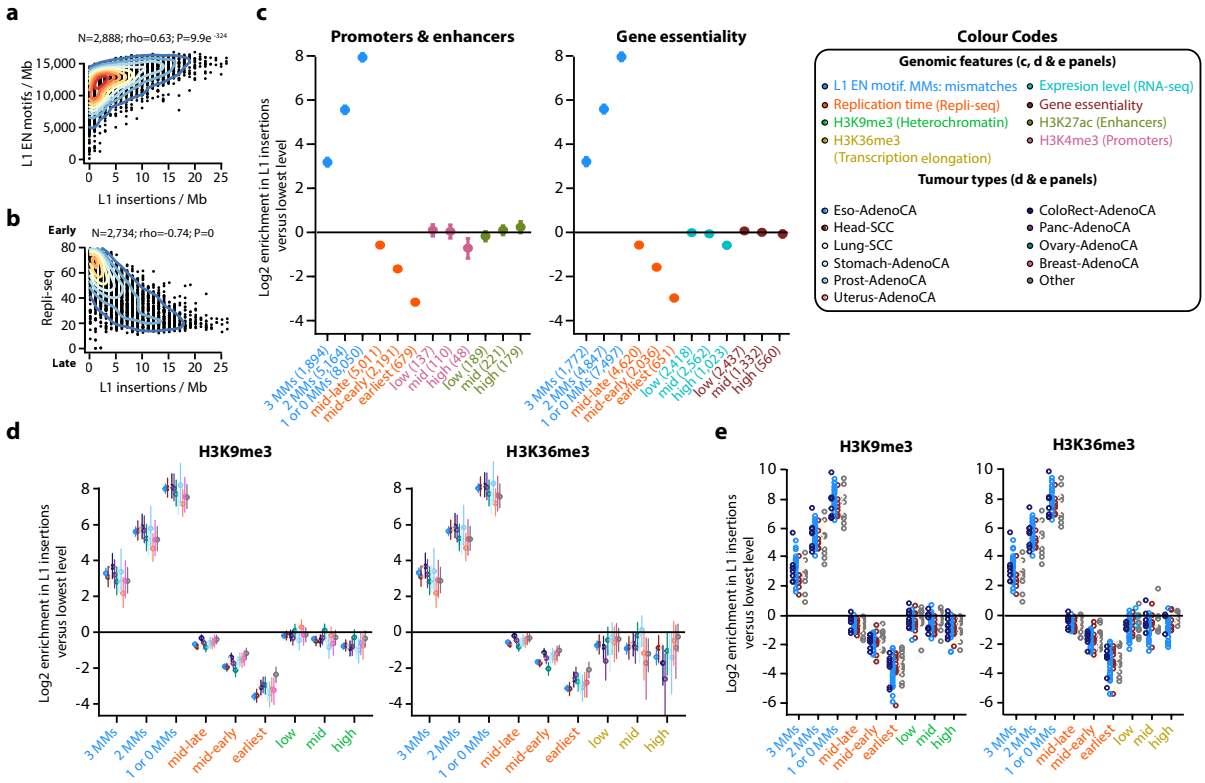


Figure S5. L1 integration and genomic features. (a) Correlation between the number of somatic L1 insertions detected in 1 Mbp bins and L1-EN motif density. 2D Kernel density estimate (KDE) is displayed over the data points in a blue to red gradient. **(b)** Correlation between the L1 insertion rate and replication timing, which is measured through Repli-seq wavelet-smoothed signal and averaged per Mb. **(c-e)** Enrichment scores resulting from comparing the L1 insertion rate in bins 1-3 for a particular genomic feature (see genomic features and colour codes in the legend panel above) versus bin 0 of the same feature, which therefore always has log enrichment=0 by definition and is not shown. Enrichment scores have been adjusted for multiple covariates. For replication time, bin 0 is the latest-replicating quarter of the genome. For gene essentiality, bin 0 is the non-essential genes. For the L1 motif, bin 0 denotes a non-match (4 or more mismatches). MMs stands for the number of mismatches relative to the consensus L1-EN motif. **(d)** Enrichments for each tumour type with at least 100 L1 insertions. Each distribution is coloured according to tumour type. **(e)** Enrichments for each sample with at least 100 L1 insertions. Each data point is coloured according to tumour type.

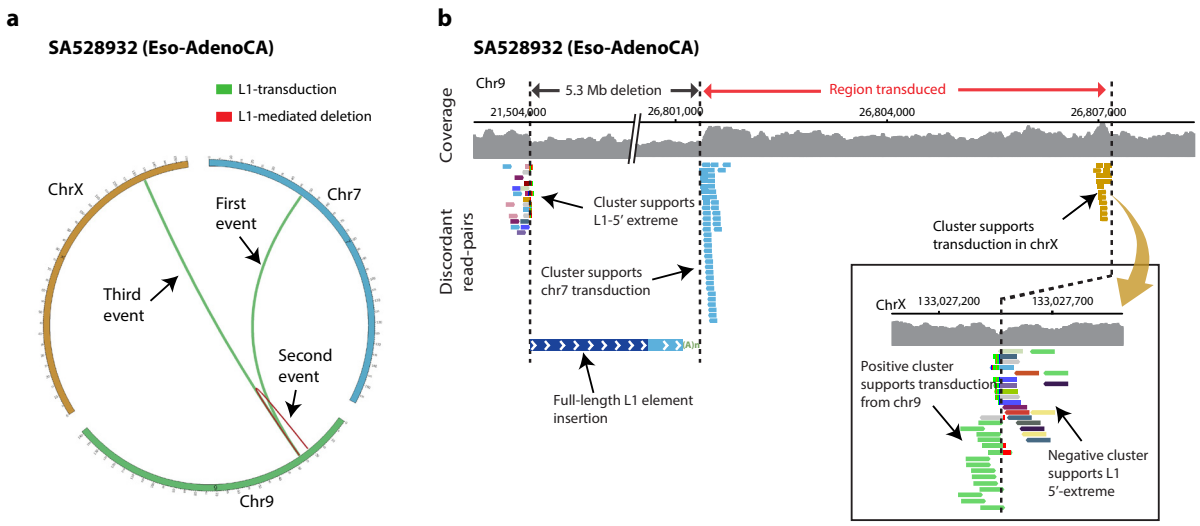


Figure S6. L1-mediated deletions can be transduction-competent. (a) Circos plot summarizing the three consecutive retrotransposition events shown in the panel b. First event, an L1 transduction mobilized from chromosome 7 is integrated into chromosome 9. Second event, this insertion concomitantly causes a 5.3 Mbp deletion in the acceptor chromosome 9. Third event, the L1 element causing the deletion is subsequently able to promote a transduction that integrates into chromosome X. **(b)** Discordant read-pairs in chromosome 9 supports a 5.3 Mbp deletion generated by the integration of a transduction from chromosome 7, and reveals an L1-event with full-length structure. Five kilobases downstream, a positive cluster of reads supports a transduction from this L1-retrotransposition event into chromosome X.

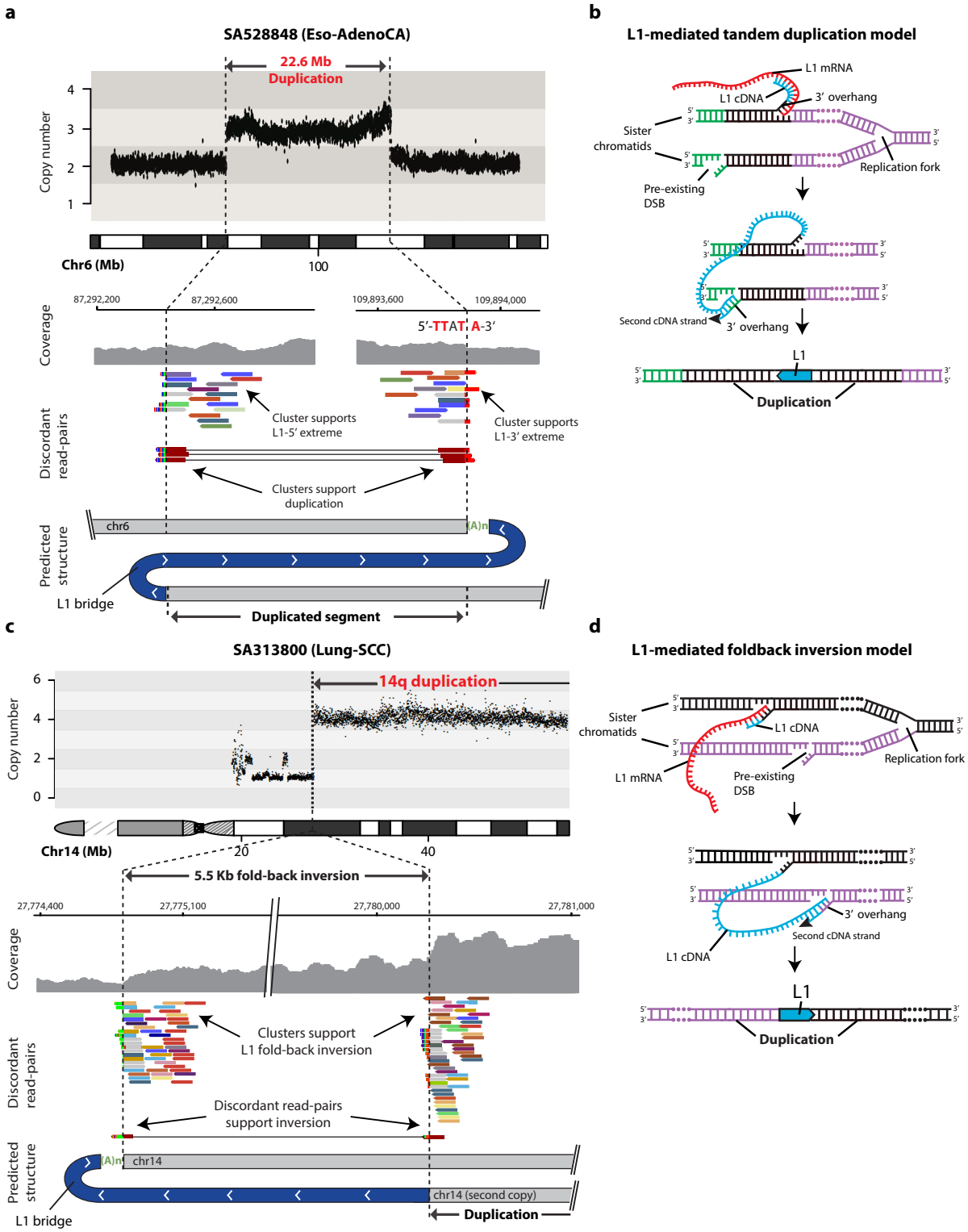


Figure S7. L1-mediated tandem duplication and fold-back inversion. (a) In an esophageal adenocarcinoma sample (SA528848), we found a 22.6 Mbp tandem duplication in the long arm of chromosome 6. The analysis of the sequencing data at the boundaries of the rearrangement breakpoints reveals two clusters of multi-coloured discordant read-pairs supporting an L1 insertion event. As the L1 element was shorter than the library size, we also found two additional clusters, aligning 22.6 Mbp apart in opposite orientation, which span the L1 insert and confirm the tandem duplication. An L1-endonuclease 5′-TTT/A-3′ degenerate motif was found. **(b)** Large tandem duplications can be generated if the cDNA (-) strand invades a second 3′ overhang from a pre-existing double-strand break that occurred on a sister chromatid, and downstream to the initial L1-EN cleavage site. **(c)** In a lung

tumour sample (SA313800), a small L1 insertion generates a dicentric chromosome through a fold-back inversion rearrangement, which is supported by two clusters of discordant read-pairs with the same orientation and located 5.5 Kbp apart. Two additional multi-coloured clusters supporting the integration of an L1 colocalize with the inversion breakpoints, indicating a L1-mediated mechanism. **(d)** L1-mediated fold-back inversion model.

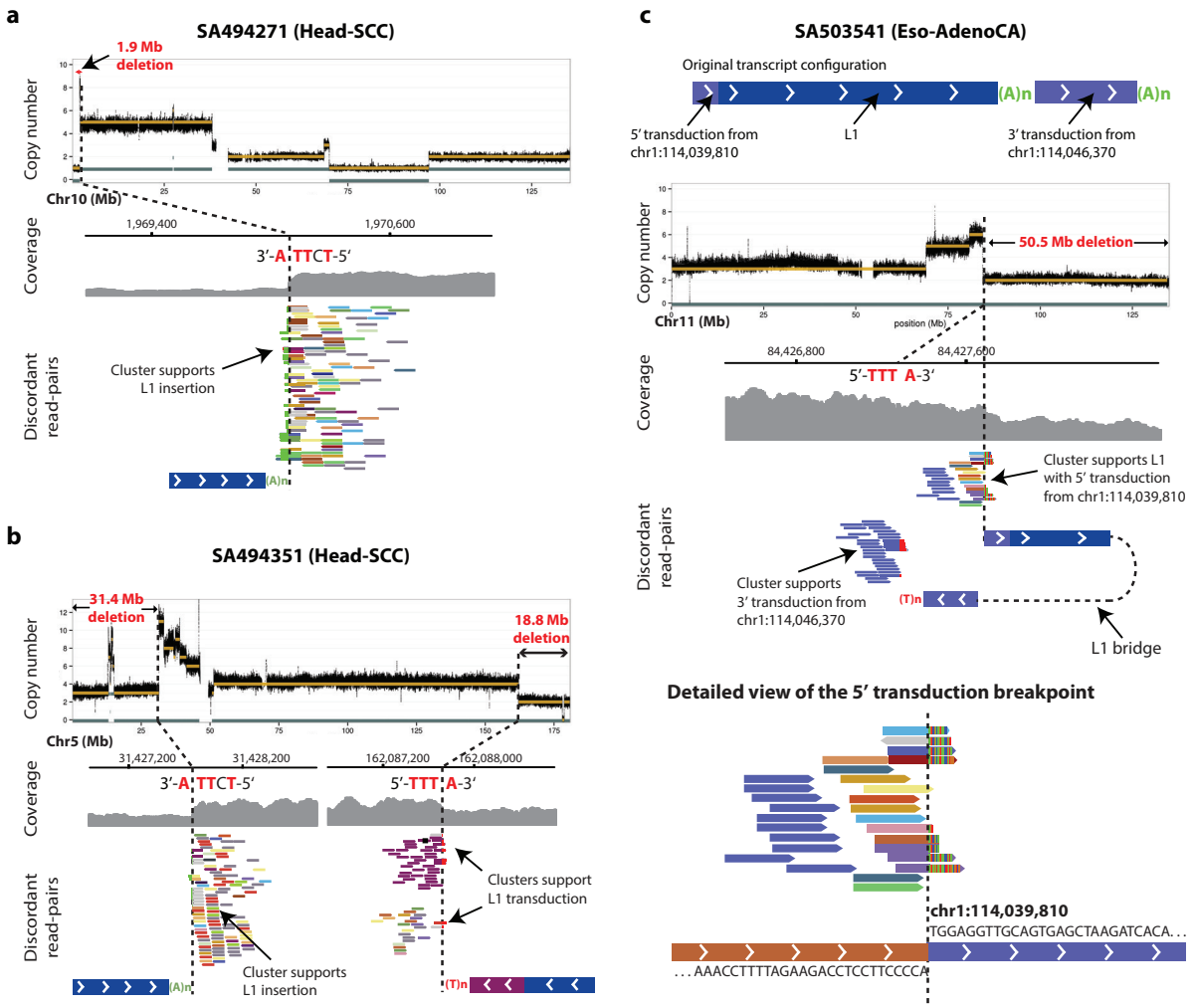
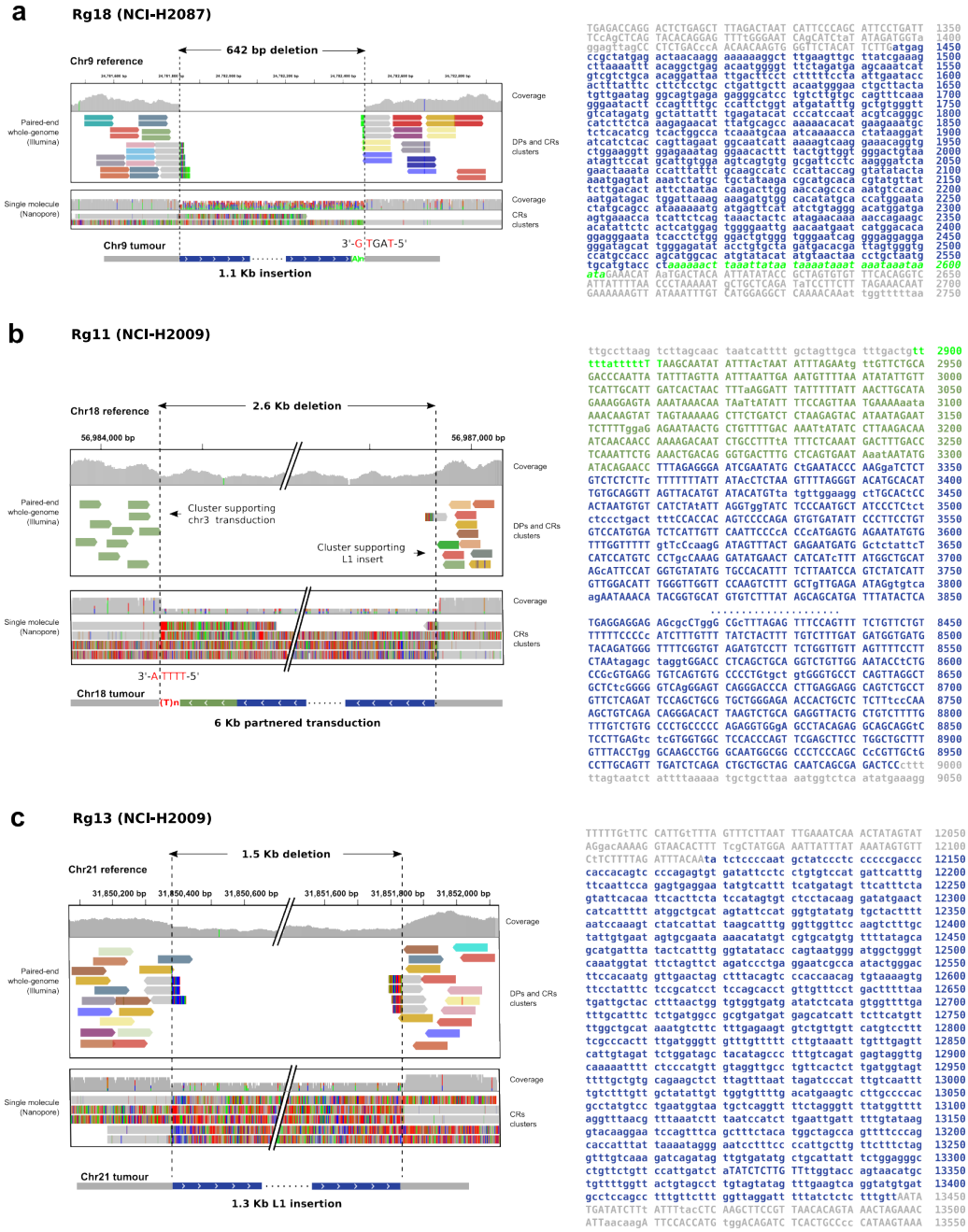


Figure S8. Somatic integration of L1 and telomere loss. The total CN and the minor allele CN are plotted as gold and gray bands, respectively. (a) In a head-and-neck tumour, SA494271, the aberrant integration of L1s produces telomeric deletion of 1.9 Mbp in size at the short arm of chromosome 10. Multi-coloured clusters at the deletion breakpoints, poly(A) and L1-EN confirm a L1-mediated origin. **(b)** In another head-and-neck tumour, SA494351, two independent L1 retrotransposition events promote the large deletions at both ends of chromosome 5. **(c)** In a lung squamous carcinoma, SA503541, the aberrant integration of an L1 event bearing 5' and 3' transductions causes a complex rearrangement with loss of 50.5 Mbp from the long arm of chromosome 11 that includes the telomere. Discordant read-pair clusters supporting the presence of an insertion containing both 5' and 3' transduced sequences are found at the telomeric deletion boundary.



Supplementary Figure 11

Figure S9. Single-molecule sequencing validation of somatic L1-mediated rearrangement calls. We deeply sequenced (>1,000X) the PCR amplicons shown in Figure 14 using ONT. We also performed shallow whole genome single-molecule sequencing (<10X) for the two tumour cell lines subjected to PCR validation (NCI-H2009 and NCI-H2087). For illustrative purposes, this figure only shows the validation of four representative rearrangements (Rg18, Rg11, Rg13, Rg20). On the left side of each panel, paired-end and ONT reads supporting a given rearrangement are displayed over a reconstruction of the rearrangement conformation. On the right side of each panel, an arbitrary ONT long-read validating each of the rearrangement structures is shown. Nucleotide colours match those in the reconstruction of the rearrangement (blue for L1, bright-green for poly-A, grey for target region, light-green for transduction). **(a)** Solo-L1 insertion mediating a 642 bp deletion. **(b)** Partnered transduction promoting a 2.6 Kbp long deletion. **(c)** A 1.5 Kbp deletion generated through an endonuclease independent L1 integration. Long reads confirm the truncation of the L1 element at its 5′ and 3′ ends. **(d)** Translocation between 1q31.1 and 8q24.12 mediated by an orphan transduction (same rearrangement as in Figure 10b). ONT reads validate the orphan transduction bridge between both chromosomes

EXTENDED ABSTRACT IN GALICIAN

Os retrotransposóns son a clase máis abundante de repeticións de ADN no xenoma humano, representando aproximadamente un terzo da secuencia xenómica. A súa notable prevalencia é unha consecuencia da súa capacidade de propagación mediante un mecanismo de copia e pega denominado retrotransposición. Aínda que os retrotransposóns foron considerados tradicionalmente como ADN "lixo", a súa mobilización nas células xerminais impactou profundamente na evolución do xenoma humano, contribuíndo á xeración de novos xenes e secuencias reguladoras. As crecentes liñas de evidencia indican que os retrotransposóns tamén pódense mobilizar máis aló da liña xerminal, con investigacións recentes sobre o xenoma do cancro que informan dunha ampla retrotransposición somática de LINE-1 (L1) en diversos tipos de tumores. Non obstante, a retrotransposición somática no cancro investigouse ata o de agora nun número limitado de xenomas, quedando por explorar múltiples histoloxías tumorais Os estudos do xenoma do cancro tamén centráronse normalmente na identificación de eventos de inserción de elementos móbiles canónicos, mentres que os retrotransposóns son capaces de mediar formas máis complexas de alteracións xenómicas, que permanecen sen identificar no contexto do cancro.

Esta tese de doutoramento ten como obxectivo investigar os patróns de actividade e as consecuencias da retrotransposición somática no cancro mediante a análise dunha gran cohorte de xenomas do cancro recopilados polo Consorcio Pan-Cancer (PCAWG). Dado o volume de datos a analizar, que incluía 2.954 xenomas completos de 35 tipos de tumores diferentes, desenvolvemos TraFiC-mem, un método computacional para a detección de insercións de elementos móbiles adquiridas somáticamente. Avaliamos a precisión e o recall de TraFiC-mem a través da reanálise de 4 xenomas sintéticos de cancro que conteñen eventos de retrotransposición somática en frecuencias alélicas que van do 25% ao 100%. A precisión e o recall de TraFiC-mem foron superiores ao 95% e ao 90% para todas as clases de VAF de inserción e L1 avaliadas, respectivamente. Aínda que non se observaron diferenzas entre as clases de inserción en canto ao recall, a precisión media de TraFiC-mem foi do 95% para as transducións asociadas, aumentando ata o 99% e o 100% para as transducións orfas e as insercións en solitario, respectivamente. Ademais, avaliamos a precisión das anotacións de TraFiC-mem mediante a comparación das lonxitudes e orientacións de inserción previstas coas expectativas baseadas nas simulacións. As lonxitudes inferidas e esperadas estaban fortemente correlacionadas (Spearman rho = 0,93; valor P = 0,0), mentres que a orientación de inserción foi consistente no 99% dos casos. Debido á non dispoñibilidade de ADN para mostras de PCAWG, utilizamos NCI-H2087, unha liña celular de cancro de pulmón coñecida por ter un alto número de insercións L1, para avaliar TraFiC-mem en datos reais. TraFiC-mem presenta un FDR inferior ao 5% para todas as clases de retrotransposóns. Todas as transducións de L1 asociadas e as insercións de Alu foron validadas, mentres que só 12 insercións solitarias de L1 e unha L1 orfa non tiñan soporte de lecturas longas. Dada a baixa

cobertura ONT dispoñible (9.17X), non podemos excluír a posibilidade de que trátense de eventos somáticos xenuínos que non se secuenciaron debido a unha cobertura insuficiente.

Identificáronse un total de 19.166 insercións somáticas de retrotransposóns en todos os xenomas do cancro. En consonancia cos informes anteriores, as insercións L1 dominan de forma abrumadora a paisaxe de retrotransposición na cohorte PCAWG, mentres que só se atoparon 130 insercións somáticas de Alu e 23 SVA. Ademais, detectamos 274 insercións de pseudoxenes procesados en 105 mostras de cancro, un número que supera con moito a estimación anterior de PSD nos xenomas do cancro. Observamos niveis particularmente altos de retrotransposición de L1 en cancros de esófago, cabeza e pescozo, pulmón e colorrectal, onde as insercións de elementos móbiles representaron unha clase predominante de variación estrutural. Outros adenocarcinomas, como os que se orixinan no estómago, páncreas, mama, útero, ovario, cérvix e próstata, presentan niveis moderados de retrotransposición. Mentres tanto, os cancros de pel, óso, cerebro e sangue teñen baixos niveis de retrotransposición somática. A mutación de *TP53* está asociada a un aumento dos niveis de insercións somáticas de L1, o que pode contribuír ás diferenzas observadas no número de insercións de L1 entre os tumores.

As integracións L1 ocorren con frecuencia dentro dos límites dos xenes, incluíndo promotores e intróns, con 66 eventos dirixidos a xenes catalogados como condutores do cancro. Polo tanto, utilizamos os datos de RNA-seq dispoñibles para o 35% dos tumores de PCAWG para investigar o impacto funcional das insercións somáticas de retrotransposóns detectadas no conxunto de datos PCAWG. En primeiro lugar, analizamos os xenes expresados de forma diferencial despois dunha inserción L1 no seu promotor. Entre as 83 insercións L1 dirixidas a promotores, catro xenes foron sobreexpresados (test t de Student, $q < 0,1$). Isto inclúe un aumento de 6 veces na expresión xénica do oncoxene *ABL2* nunha mostra de carcinoma escamoso de cabeza e pescozo, SA494343, en relación aos restantes tumoresde cabeza e pescozo sen a inserción L1. Ampliamos aínda máis a análise transcriptómica ás insercións que afectan a calquera compoñente dos xenes do cancro (é dicir, intróns ou exóns), o que revelou a sobreexpresión do xene supresor de tumores *RB1* nun tumor de vexiga (test t de Student, $q < 0,10$). Unha análise adicional indicou que o cambio observado na expresión dos xenes probablemente sexa causado por unha integración L1 no segundo intrón de *RB1*. Máis precisamente, identificamos pares de lectura discordantes que admiten o empalme aberrante entre o inserto L1 e 7 exóns *RB1* diferentes, que indican a existencia de múltiples isoformas *L1-RB1* xeradas mediante a exonización L1. Detectamos 23 casos adicionais de exonización, incluíndo 5 insercións L1 e Alu dentro dos exóns, unha inserción L1 adicional no intrón do xene *NCOR2* e 17 pseudoxenes procesados.

As 18.739 insercións L1 adquiridas somáticamente detectadas no conxunto de datos PCAWG proporcionaron unha excelente oportunidade para investigar os patróns de distribución da inserción L1 no xenoma do cancro. A distribución das retrotransposicións somáticas L1 foi marcadamente heteroxénea ao longo do xenoma do cancro. Como a integración de L1

depende dunha endonuclease autocodificada que se dirixe a unha secuencia diana consenso dexenerada (5-TTTT/A-3), primeiro investigamos se a distribución de L1 somáticas a través do xenoma do cancro podía ser determinada pola aparición de motivos diana L1-EN. Usando un enfoque estatístico baseado na regresión binomial negativa para deconvolucionar a influencia de múltiples características xenómicas, incluíndo o tempo de replicación, diversas marcas epixenéticas, o estado da cromatina e a expresión xénica, observamos un enriquecemento de 244 veces das insercións L1 en secuencias moi semellantes aos motivos L1-EN. Como sábese que o tempo de replicación ten un gran impacto nas taxas de mutación locais nos xenomas do cancro, investigamos a súa asociación coa retrotransposición somática L1. Axustando polo efecto potencialmente confuso dos motivos L1-EN, observamos unha forte asociación entre a retrotransposición somática de L1 e o tempo de replicación do ADN.

Como se informou previamente que a retrotransposición somática L1 estaba enriquecida en rexións heterocromáticas, tamén examinamos as taxas de L1 en rexións heterocromáticas pechadas analizando a histona H3 trimetilada K9 (H3K9me3). Ao axustar os efectos de confusión do contido do motivo L1-EN e o tempo de replicación, descubrimos que as insercións somáticas están esgotadas en heterocromatina e enriquecidas en cromatina aberta. Esta discrepancia coas análises anteriores é probablemente a consecuencia do efecto de confusión entre a heterocromatina e as rexións de ADN de replicación tardía, que antes non se tiña en conta. Tamén atopamos unha asociación negativa entre a taxa de insercións de L1 e as características xenómicas relacionadas coa transcrición activa da cromatina, caracterizada por menos eventos L1 nos promotores activos, unha redución lixeira pero significativa das taxas de L1 en xenes altamente expresados e un esgotamento en H3K36me3, unha marca de rexións activamente transcritas depositadas no corpo e no extremo 3′ dos xenes activos.

Usando o noso algoritmo TraFiC-mem, detectamos 3.696 transducións somáticas no conxunto de datos de tumores PCAWG. As transducións orfas, nas que se retrotranspón unha secuencia downstream dunha L1 activa sen a devandita L1, representaron o 64% de todos os eventos, sendo o resto transducións asociadas (é dicir, cunha L1 acompañante). Estimamos que o tamaño medio das secuencias transducidas é de 333 pb, aínda que ocasionalmente se detectaron transducións longas que alcanzaban tamaños de ata 1,5 Kb. En consonancia cos patróns descritos anteriormente para as insercións de solo-L1, as transducións foron particularmente abundantes nos cancros de esófago, cabeza e pescozo, pulmón e colon; só estes catro tipos de tumores abarcan o 70% de todas as transducións. Dado que as transducións defínense pola retrotransposición dunha secuencia xenómica única, pódense usar para identificar sen ambigüidades os loci L1 de onde derivan. Descubrimos que 114 L1 da liña xerminal foron responsables de todas as transducións identificadas na cohorte PCAWG. Aínda que anteriormente se informou de que 60 elementos fonte L1 estaban activos, 54 elementos son de feito novas copias activas, o que amplía o catálogo de L1 activos en humanos.

Observamos unha variabilidade considerable na actividade entre as fontes L1, cun conxunto reducido de 16 loci L1 altamente activos (é dicir, “quentes”) que foron responsables do 67% de todas as transducións detectadas. Isto é consistente con estudos anteriores en tumores humanos de orixe natural e ensaios *in vitro*, apoiando ademais a idea de que a maioría das mobilizacións de retrotransposóns na liña xerminal e no soma orixínanse dun número reducido de copias L1 con actividade “quente”. Os Hot-L1 mostran dous patróns de actividade diferenciados, que denominamos “Estromboliano” / “Stromboliano”” e “Pliniano” debido á súa semellanza cos patróns de erupción dos volcáns. Os volcáns plinianos, como o Vesubio en Pompeia, caracterízanse por unha actividade volcánica esporádica pero particularmente intensa. Mentres tanto, o monte Stromboli, en Italia, estivo en actividade case continua durante 2.000 anos, producindo erupcións levemente explosivas. Por analoxía, os L1 plinianos raramente son activos a través dos tumores, pero producen intensos estalidos de retrotransposición somática, mentres que os L1 estrombolianos son frecuentemente activos no cancro, pero só median erupcións de actividade L1 somática de pequenas a moderadas. Mentres que os elementos estrombolianos adoitan ser relativamente comúns e ás veces mesmo alelos fixados na poboación humana, todos os elementos plinianos son polimorfismos raros. Este notable patrón dicotómico de actividade e frecuencia de alelos pode ser consecuencia de diferenzas na súa idade e na presión selectiva que actúa sobre estes loci L1, sendo os elementos plinianos os que probablemente representen L1 “quentes” adquiridos recentemente, que aínda non alcanzaron un equilibrio coa nosa especie.

Observamos que o número de copias de elementos fonte activos por xenoma individual do cancro é moi heteroxéneo, oscilando entre cero e 22 L1 activos por xenoma, e que se correlaciona fortemente co número de retrotransposicións somáticas. Do mesmo xeito, as mostras de tipos de cancro con altas taxas de retrotransposición (é dicir, de colon, cabeza e pescozo, pulmón e esófago) teñen de media 2-4 elementos fonte activos, o que é 4-8 veces superior á media de todo o conxunto de datos do proxecto Pan-Cancer. Estes datos indican que a contribución acumulada de múltiples copias activas determina en gran medida a carga de retrotransposición nun tumour determinado, e que os tumores que presentan altas taxas de retrotransposición caracterízanse por un elevado número de elementos fonte L1 activos.

Durante a análise da retrotransposición somática, observamos un patrón de integración L1 moi intrigante nalgúns xenomas do cancro con altos niveis de actividade L1. Consistía nun único grupo de lecturas que admitía un dos extremos dunha integración L1, que estaba asociada cunha perda de número de copia. Unha análise máis detallada do cambio do número de copia revelou o clúster recíproco que faltaba, que admitía o segundo extremo para a integración L1, no extremo máis afastado da perda do número de copia, o que suxire que a eliminación ocorreu xunto coa integración dun L1. Un tramo poli(A) estaba presente nun dos puntos de ruptura da perda do número de copias xunto co motivo diana da L1 EN. Estes distintivos aseméllanse aos que se describiron anteriormente para os reordenamentos mediados por L1, o que suxire que a integración aberrante de L1 xerou a perda de ADN descrita.

Desenvolvemos un método computacional para buscar sistemáticamente delecións mediadas por L1 en todos os xenomas do cancro de PCAWG. Detectamos 90 eventos somáticos que coincidían cos patróns descritos anteriormente, abarcando eliminacións de diferentes tamaños, que van desde 0,5 Kbp ata 53,4 Mb. Reconstruímos con éxito as unións do punto de ruptura para as 90 eliminacións mencionadas anteriormente, atopando unha secuencia derivada de L1 en todas elas. Ademais, o 82% das delecións caracterizadas contiña unha secuencia que se asemella a sitios de clivaxe de consenso L1-EN nos seus puntos de ruptura 3' (motivo dexenerado 5'-TTTT/A-3'). Todas as delecións asociadas aos motivos L1-EN tamén contiñan un tracto de poliadenilato nos seus puntos de ruptura 3', indicativo do paso a través dun intermedio de ARN. En xeral, estas características suxiren que a maquinaria L1, mediante a transcrición reversa mediada por diana, é responsable da integración da maioría das secuencias L1 que causan a perda de ADN veciño.

Aínda que a maioría das delecións mediadas por L1 caracterizadas varían entre uns centos e miles de pares de bases, en ocasións a integración aberrante dos elementos L1 pode provocar a perda de rexións cromosómicas do tamaño da megabase, afectando aos xenes supresores de tumores. Por exemplo, nunha mostra de tumor de esófago (SA528932), unha transdución de L1 xerada a partir dun loci L1 localizado en 7p12.3 causou unha perda de 5,3 Mbp que afectaba ao brazo curto do cromosoma 9 que eliminou *CDKN2A*, un xene supresor de tumores relevante que se muta con frecuencia en moitos tipos de cancro, incluídos os tumores de esófago. Nunha segunda mostra de tumor de esófago (SA528899), unha integración L1 no cromosoma 9 promoveu unha deleción de 8,6 Mbp de tamaño que, de novo, elimina *CDKN2A*. A análise das frecuencias alélicas das variantes revelou que ambas delecións son clonais, o que suxire que puideron ocorrer cedo durante a evolución destes tumores. Estes achados destacan o potencial da integración aberrante de L1 para promover as perdas de ADN con funcións oncoxénicas.

Mentres buscamos delecións mediadas por L1, observamos que a retrotransposición aberrante de L1 pode estar implicada na xeración doutras formas de variación estrutural. Nunha mostra de tumor de esófago (SA528896), observáronse dúas translocacións separadas mediadas por L1 no contexto dun complexo grupo de reordenamentos. En primeiro lugar, unha transdución L1 a partir dun elemento fonte en 14q23.1 asociouse cunha translocación desequilibrada que implica 1p e 5q. En segundo lugar, atopouse outra secuencia L1 na unión entre 5p e un locus xenómico descoñecido, completando unha gran perda intersticial de número de copias no cromosoma 5 que afectou ao centrómero. Estas observacións suxiren que os retroelementos L1 poden salvar roturas de dobre cadea localizadas en diferentes cromosomas. Para investigar máis esta pregunta, extraemos datos da liña celular de cancro de pulmón para buscar translocacións que contiveran sinaturas de pares de lectura discordantes para a integración de L1 nos seus puntos de interrupción (é dicir, poli(A), motivo EN e pares de lectura compatibles). Esta procuradescubriu unha translocación que conecta 1q31.1 con 8q24.12 na liña celular tumoral NCI-H2087. Curiosamente, ambos os puntos de ruptura de translocación

están flanqueados por grupos de pares de lectura discordantes que soportan unha transdución L1 orfa derivada dun elemento fonte L1 situado no cromosoma 6p24. É probable que este reordenamento intercromosómico se orixinase a través da operación aberrante da reacción canónica de TPRT, que leva ao apareamento da secuencia de ADNc transducida cun saliente 3′ derivado dunha rotura de dobre cadea preexistente nun segundo cromosoma.

Tamén atopamos evidencia de que as integracións L1 poden causar duplicacións de grandes rexións xenómicas no cancro. Por exemplo, noutra mostra de tumor esofáxico relevante (SA528848), identificamos dous grupos de lectura discordantes que apoian a integración dun elemento L1 truncado, xunto cun aumento da cobertura delimitada por ambos os puntos de ruptura de inserción de L1. A análise do número de copias indica que os dous grupos de lecturas L1 demarcan os límites dunha duplicación de 22,6 Mb, o que suxire que a inserción de L1 podería ser a causa de tal reordenación ao unir cromátidas irmás durante ou despois da replicación do ADN. A análise detallada dos datos de secuenciación revela a presenza de grupos de lectura discordantes adicionais que admiten tanto unha duplicación en tándem como un motivo L1-EN no punto de ruptura de inserción 3', confirmando un único evento L1 como a causa desa duplicación en tándem. Notablemente, este reordenamento aumenta o número de copias do xene da ciclina C, *CCNC*, que está desregulado nalgúns tumores.

Os ciclos de rotura-fusión-ponte (BFB) son un mecanismo de inestabilidade xenómica que se inicia coa fusión de extremo a extremo de cromátidas rotas, do mesmo ou de dous cromosomas diferentes, xerando un cromosoma dicéntrico. Durante a mitose, os dous centrómeros dun cromosoma dicéntrico son arrastrados a polos opostos da célula en división, creando unha ponte anfásica, que se resolve mediante a rotura do ADN de dobre cadea nunha posición arbitraria entre ambos os centrómeros. Como os produtos cromosómicos resultantes adoitan presentar máis deficiencias de telómeros, é probable que o cromosoma experimente varias roldas de BFB ata que finalmente se estabiliza mediante a reparación dos seus extremos. Os BFB son especialmente relevantes no contexto do cancro, xa que as sucesivas roldas de BFB adoitan levar á amplificación de oncoxenes, aportando as alteracións xenómicas necesarias para a transformación maligna.

Aínda que McClintock as descubriu inicialmente a finais da década de 1930, cando observou frecuentes roturas de cromosomas e fusións nas células de millo mitóticas despois da exposición aos raios X, actualmente sábese que as fusións de extremo a extremo orixínanse a través de múltiples mecanismos, incluíndo a desgaste dos telómeros e a cromotripsis. Durante a análise da retrotransposición somática, descubrimos que a retrotransposición aberrante de L1 é un mecanismo alternativo para a fusión de extremo a extremo das cromátidas irmás e a formación de cromosomas dicéntricos. Nunha mostra de tumor de pulmón (SA313800), identificamos dous grupos de lectura discordantes coa mesma orientación e situados a 5,5 Kbp de distancia que apoiaban a presenza dunha inserción L1 ao longo do brazo longo do cromosoma 14. Ambos os grupos colocalízanse con dous grupos de pares de lectura discordantes en orientación "cabeza-a-cabeza", que representan a sinatura de lectura clásica

dunha inversión de repregamento, o que suxire que o evento de inserción L1 está implicado na xeración da reordenación invertida. Ademais, os puntos de ruptura de inversión do foldback están exactamente no límite entre unha deleción a gran escala que afecta aos primeiros 27 Mbp do cromosoma 14 e unha duplicación de 79,6 Mbp do brazo 14q. A única estrutura xenómica que se axusta a este patrón é unha inversión de repregamento na que as dúas cromátidas irmás están unidas mediante unha inserción L1, xerando un isocromosoma (14q). De novo, isto pódese explicar mediante unha variante da reacción TPRT, onde o ADNc de L1 usa unha rotura de dobre cadea preexistente para invadir a cromátida irmá durante a replicación do ADN, dando como resultado a conformación xenómica descrita.

No exemplo descrito anteriormente, non se produciron máis roturas e o isocromosoma mantívose estable. Non obstante, atopamos exemplos nos que a fusión de dúas cromátidas por unha ponte L1 inducía máis ciclos de reparación de BFB. Nunha mostra de tumor de esófago, SA528848, identificamos un grupo de lecturas no brazo longo do cromosoma 11 que tiña as características típicas dun reordenamento mediado por L1. A análise dos datos de número de copia mostrou que os puntos de ruptura da inserción L1 delimitaban unha deleción de 53 Mb, que implicaba a perda da rexión telomérica e unha amplificación masiva no cromosoma 11. A rexión amplificada no cromosoma 11 contén o oncoxene *CCND1*, que se amplifica no cromosoma 11 en moitos cancros humanos. O outro extremo desta amplificación estaba unido por un reordenamento de inversión dobrada convencional, o que é indicativo da reparación do BFB.

Estes patróns suxiren a seguinte secuencia de eventos. Durante ou pouco despois da fase S, unha retrotransposición somática L1 atravesa as cromátidas irmás en orientación invertida, rompendo os extremos teloméricos de 11q, que se perden durante a división celular posterior (modelo de inversión dobrada). As cromátidas unidas pola inserción L1 producen agora un cromosoma dicéntrico. Durante a mitose, os dous centrómeros lévanse a polos opostos da célula en división, creando unha ponte anafase, que se resolve mediante unha nova rotura do DNA. Isto induce un segundo ciclo de reparación de BFB, aínda que non mediado pola retrotransposición de L1. Estes ciclos conducen a unha amplificación rápida do oncoxene *CCND1*. Alternativamente, un reordenamento intercromosómico mediado pola retrotransposición L1 (modelo de reordenamento intercromosómico) seguido de dous ciclos de reparación de BFB podería xerar patróns de número de copias similares con perda de telómeros e amplificación de *CCND1*.

Identificamos catro casos adicionais de ciclos BFB iniciados pola retrotransposición L1 no conxunto de datos PCAWG. Sorprendentemente, nun adenocarcinoma de pulmón, SA503541, atopamos un reordenamento mediado por L1 que se asemella claramente aos patróns descritos anteriormente, incluíndo a perda de telómeros, a amplificación de *CCND1* e os grupos de lecturas de inversión dobrada dentro do amplicón. Neste caso, os datos son consistentes con dúas roldas, o que leva á adquisición de dúas copias adicionais de *CCND1*. A aparición independente de reordenamentos semellantes, que implican a amplificación do

mesmo oncoxene, en dúas mostras tumourais diferentes (SA528848 e SA503541) mostra a relevancia deste novo proceso mutacional mediado por L1. En xeral, estes datos revelan que a retrotransposición de L1 é un mecanismo alternativo para a formación de cromosomas dicéntricos e o inicio dos ciclos BFB. Se isto ocorre preto dun oncoxene, como *CCND1*, a amplificación resultante pode proporcionar unha poderosa vantaxe selectiva para o clon e potencialmente levar ao desenvolvemento de cancro.

DECLARATIONS: CONFLICTS OF INTEREST AND USE OF PUBLISHED MATERIAL

I declare not to be under any conflict of interest in relation to this PhD dissertation. All the cancer genomic data included in this thesis derives from anonymous donors and therefore it is not necessary to be approved by an ethics committee (For additional information refer to the manuscripts 1-5). All the figures included in the thesis derive from published research pieces under an open source license. Further licensing details for each manuscript are provided bellow:

Manuscript 1

Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition

Nat Genet 52, 306–319 (2020). DOI: https://doi.org/10.1038/s41588-019-0562-0, ISSN: 1061-4036

Authors

Bernardo Rodriguez-Martin[1,2,3], Eva G. Alvarez[1,2,3§], Adrian Baez-Ortega[4,§], Jorge Zamora[1,2,§], Fran Supek[5,§], Jonas Demeulemeester[6,7], Martin Santamarina[1,2,3], Young Seok Ju[8], Javier Temes[1], Daniel Garcia-Souto[1], Harald Detering[9], Yilong Li[10], Jorge Rodriguez-Castro[1], Ana Dueso-Barroso[11,12], Alicia L. Bruzos[1,2,3], Stefan C. Dentro[13,6,14], Miguel G. Blanco[15,16], Gianmarco Contino[17], Daniel Ardeljan[18], Marta Tojo[9], Nicola D. Roberts[10], Sonia Zumalave[1], Paul A. W. Edwards[19,20], Joachim Weischenfeldt[21,22], Montserrat Puiggros[11], Zechen Chong[23], Ken Chen[23], Eunjung Alice Lee[24], Jeremiah A. Wala[25,26], Keiran Raine[10], Adam Butler[10], Sebastian M. Waszak[22], Fabio C. P. Navarro[27,28], Steven E. Schumacher[25,26], Jean Monlong[29], Francesco Maura[30,31,10], Niccolo Bolli[30,31], Guillaume Bourque[29], Mark Gerstein[27,28], Peter J. Park[24], David Wedge[14,10,32], Rameen Beroukhim[25,26], David Torrents[11,33], Jan O. Korbel[22], Inigo Martincorena[10], Rebecca C. Fitzgerald[17], Peter Van Loo[6,7], Haig H. Kazazian[18], Kathleen H. Burns[34,18], Peter J. Campbell[10,35,*] & Jose M. C. Tubio[1,2,10,3,*], on behalf of the PCAWG Structural Variation Working Group, and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network

§These authors contributed equally to the manuscript

*These authors contributed equally to the manuscript

[1]Mobile Genomes and Disease, Molecular Medicine and Chronic diseases Centre (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain

[2]Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain

[3]The Biomedical Research Centre (CINBIO), University of Vigo, Vigo 36310, Spain

[4]Transmissible Cancer Group, Department of Veterinary Medicine, University of Cambridge,

Cambridge CB3 0ES, UK

[5]Genome Data Science, Institute for Research in Biomedicine IRBB, The Barcelona Institute of Science and Technology (BIST), Barcelona 08028, Spain

[6]The Francis Crick Institute, London, UK

[7]Department of Human Genetics, University of Leuven, Leuven, Belgium

[8]Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea

[9]Evolutionary Genomics, The Biomedical Research Centre - CINBIO, University of Vigo, 36310 Vigo, Spain

[10]Cancer Ageing and Somatic Mutation Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB101SA, UK

[11]Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain

[12]Faculty of Science and Technology. University of Vic - Central University of Catalonia (UVic-UCC), Vic, Spain

[13]Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA. UK

[14]Oxford Big Data Institute, University of Oxford, Oxford, UK

[15]Molecular Medicine and Chronic diseases Centre (CIMUS) – IDIS, University of Santiago de Compostela, Santiago de Compostela 15706, Spain

[16]Departamento de Bioquímica e Bioloxía Molecular, CIMUS, Universidade de Santiago de Compostela, 15706 Santiago de Compostela, Spain

[17]Medical Research Council (MRC) Cancer Unit, University of Cambridge, Cambridge, UK

[18]Institute for Genetic Medicine, Johns Hopkins University School of Medicine - Baltimore, MD USA

[19]Department of Pathology, University of Cambridge, Cambridge, UK

[20]Cancer Research UK Cambridge Institute, Cambridge, UK

[21]Biotech Research & Innovation Centre (BRIC); Finsen Laboratory, Rigshospitalet, Copenhagen, Denmark

[22]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany

[23]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

[24]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

[25]The Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[26]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[27]Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT

[28]Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT

[29]Department of Human Genetics, McGill University, Montreal, H3A 1B1, Canada

[30]Department of Oncology and Onco-Hematology, University of Milan, Milan, Itlay

[31]Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei tumouri, Milan, Italy

[32]Oxford NIHR Biomedical Research Centre, Oxford, UK

[33]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[34]Department of Pathology, Johns Hopkins University School of Medicine - Baltimore, MD USA

[35]Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK

Specific contributions to the publication

I developed computational methods for the detection of somatically acquired mobile element insertions in cancer genome data. I performed variant calling of somatic retrotransposition insertions for the cancer whole genomes included in the pan-cancer cohort. I analyzed the rates of somatic retrotransposition across tumour types and investigated their association with genomic features, other structural variant classes and mutations in cancer genes. I studied the patterns of activity for source L1 elements across cancer histologies. I characterized the breakpoints for L1-mediated rearrangements, finding the involvement of NHEJ machinery in their mechanisms of formation. Finally, I contributed with multiple main and supplementary figures in addition to text pieces.

Quality metrics

The journal Nature Genetics has an impact factor of 38.33 (2020 Journal Citation Reports), a Scopus CiteScore of 50.5 (22th of November 2021) and is in the first quartile (Q1) in Genetics according to Scimago (SJR 2020 18.86).

Journal authorization

Manuscript 2

Pan-cancer analysis of whole genomes.

Nature 578, 82–93 (2020). DOI: https://doi.org/10.1038/s41586-020-1969-6, ISSN: 0028-0836

Authors

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium. Given the large size of the PCAWG Consortium, involving several hundreds of researchers, the list of members and their affiliations is provided in the online version of the manuscript, but not in this PhD dissertation.

Specific contributions to the publication

I characterized the patterns of somatic activity for source L1 elements, finding two types of hot-L1s, which resemble volcano eruption types. These analyses are described in the “Germline effects on somatic mutations” main section. I drafted the text and, together with Eva and Javi, made Fig. 6d. I also generated a catalogue of germline mobile element insertions based on the analysis of the matched-normal WGS. General statistics regarding the germline mobile element callset and evaluation are provided in Extended Data Fig. 12.

Quality metrics

The journal Nature has an impact factor of 49.96 (2020 Journal Citation Reports), a Scopus CiteScore of 56.9 (22th of November 2021) and is in the first quartile (Q1) in Multidisciplinary according to Scimago (SJR 2020 15.99).

Journal authorization

Manuscript 3

Genomic landscape and chronological reconstruction of driver events in multiple myeloma.

Nat Commun 10, 3835 (2019). DOI: https://doi.org/10.1038/s41467-019-11680-1, ISSN: 2041-1723

Authors

Francesco Maura[1,2,3,14], Niccoló Bolli[3,4,14], Nicos Angelopoulos[2,5], Kevin J. Dawson[2], Daniel Leongamornlert[2], Inigo Martincorena[2], Thomas J. Mitchell[2], Anthony Fullam[2], Santiago Gonzalez[6], Raphael Szalat[7], Federico Abascal[2], Bernardo Rodriguez-Martin[8], Mehmet Kemal

Samur[7], Dominik Glodzik[2,9], Marco Roncador[2], Mariateresa Fulciniti[7], Yu Tzu Tai[7], Stephane Minvielle[10], Florence Magrangeas[10], Philippe Moreau[10], Paolo Corradini[3,4], Kenneth C. Anderson[7], Jose M.C. Tubio[2,8], David C. Wedge[11], Moritz Gerstung[6], Hervé Avet-Loiseau[12], Nikhil Munshi[7,13,15] & Peter J. Campbell[2,15]

[1]Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[2]The Cancer, Ageing and Somatic Mutation Programme, Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK.

[3]Department of Medical Oncology and Hemato-Oncology, University of Milan, Milan, Italy.

[4]Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei tumouri, Milan, Italy.

[5]School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK.

[6]European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Hinxton, UK.

[7]Jerome Lipper Multiple Myeloma Center, Dana–Farber Cancer Institute, Harvard Medical School, Boston, MA, USA.

[8]CIMUS - Molecular Medicine and Chronic Diseases Research Centre, University of Santiago de Compostela, Santiago de Compostela, Spain.

[9]Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[10]CRCINA, INSERM, CNRS, Université d'Angers, Université de Nantes, Nantes, France.

[11]University of Oxford, Big Data Institute, Oxford, UK.

[12]IUC-Oncopole, and CRCT INSERM U1037, 31100 Toulouse, France.

[13]Veterans Administration Boston Healthcare System, West Roxbury, MA, USA.

[14]These authors contributed equally: Francesco Maura, Niccoló Bolli.

[15]These authors jointly supervised this work: Nikhil Munshi and Peter J. Campbell.

Specific contributions to the publication

I contributed through the analysis of somatic retrotransposition in 67 multiple myeloma samples. This includes insertion calling, manual inspection and callset refinement.

Quality metrics

The journal Nature Communications has an impact factor of 14.91 (2020 Journal Citation Reports), a Scopus CiteScore of 20.00 (22th of November 2021) and is in the first quartile (Q1) in Biochemistry, Genetics and Molecular Biology (miscellaneous) according to Scimago (SJR 2020 5.56).

Journal authorization

Manuscript 4

CDKN2A deletion is a frequent event associated with poor outcome in patients with peripheral T-cell lymphoma not otherwise specified (PTCL-NOS).

Haematologica 106, 11 (2021). DOI: https://doi.org/10.3324/haematol.2020.262659, ISSN: 0390-6078

Authors

Francesco Maura[1-4], Anna Dodero[5], Cristiana Carniti[5], Niccolò Bolli[2,5,] Martina Magni[5], Valentina Monti[6], Antonello Cabras[6], Daniel Leongamornlert[3], Federico Abascal[3], Benjamin Diamond[1], Bernardo Rodriguez-Martin[7], Jorge Zamora[7], Adam Butler[3], Inigo Martincorena[3], Jose M. C. Tubio[7], Peter J. Campbell[3], Annalisa Chiappella[8°], Giancarlo Pruneri[2,6] and Paolo Corradini[2,5]

[1]Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[2]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy.

[3]The Cancer, Aging and Somatic Mutation Program, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK.

[4]Weill Cornell Medical College, New York, NY, USA.

[5]Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei tumouri, Milan, Italy;

[6]Department of Pathology and Laboratory Medicine, Fondazione IRCCS Istituto Nazionale dei tumouri, Milan, Italy.

[7]CIMUS - Molecular Medicine and Chronic Diseases Research Center, University of Santiago de Compostela, Santiago de Compostela, Spain and 8Department of Hematology Azienda Ospedaliera Città della Salute e della Scienza, Turin, Italy.

[8°]Current address: Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei tumouri, Milan, Italy.

Specific contributions to the publication

I contributed through the analysis of somatic retrotransposition in eleven peripheral T-cell lymphoma samples. This includes insertion calling, manual inspection and callset refinement.

Quality metrics

The journal Haematologica has an impact factor of 9.94 (2020 Journal Citation Reports), a Scopus CiteScore of 8.90 (22th of November 2021) and is in the first quartile (Q1) in Hematology according to Scimago (SJR 2020 2.78).

Journal authorization

Manuscript 5

Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis.

Cell 176, 1282–1294.e20 (2019). DOI: https://doi.org/10.1016/j.cell.2019.02.012, ISSN: 0092-8674

Authors

Mia Petljak[1], Ludmil B. Alexandrov[1,2], Jonathan S. Brammeld[1], Stacey Price[1], David C. Wedge[3,4], Sebastian Grossmann[1], Kevin J. Dawson[1], Young Seok Ju[5], Francesco Iorio[1,6], Jose M.C. Tubio[1,7,8,9], Ching Chiek Koh[1], Ilias Georgakopoulos-Soares[1], Bernardo Rodriguez–Martin[7,8,9], Burçak Otlu[2], Sarah O'Meara[1], Adam P. Butler[1], Andrew Menzies[1], Shriram G. Bhosle[1], Keiran Raine[1], David R. Jones[1], Jon W. Teague[1], Kathryn Beal[1], Calli Latimer[1], Laura O'Neill[1], Jorge Zamora[7,8,9], Elizabeth Anderson[1], Nikita Patel[1], Mark Maddison[1], Bee Ling Ng[10], Jennifer Graham[10], Mathew J. Garnett[1], Ultan McDermott[1], Serena Nik-Zainal[1,11], Peter J. Campbell[1], and Michael R. Stratton[1,12,*]

[1]Cancer, Ageing and Somatic Mutation, Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK

[2]Department of Cellular and Molecular Medicine and Department of Bioengineering, Moores Cancer Center, University of California, San Diego, La Jolla, CA 92093, USA

[3]Oxford Big Data Institute, Old Road Campus, Oxford OX3 7LF, UK

[4]Oxford NIHR Biomedical Research Centre, Oxford, OX4 2PG, UK

[5]Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science

and Technology, Daejeon 305-701, Republic of Korea

[6]European Molecular Biology Laboratory - European Bioinformatics Institute, Hinxton, Cambridgeshire CB10 1SA, UK

[7]Mobile Genomes and Disease, Molecular Medicine and Chronic Diseases Centre (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain

[8]Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela 15706, Spain

[9]The Biomedical Research Centre (CINBIO), Universidade de Vigo, Vigo 36310, Spain

[10]Cytometry Core Facility, Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK

[11]Department of Medical Genetics, The Clinical School, University of Cambridge, Cambridge, CB2 0QQ, UK

[12]Lead Contact. *Correspondence: mrs@sanger.ac.uk

Specific contributions to the publication

I contributed through the analysis of somatic retrotransposition in 57 clones cultured *in vitro*. This includes insertion calling, manual inspection and callset refinement. We also provided the L1 insertion counts for PCWAG tumours, which were used to investigate the association between L1 retrotransposition and APOBEC signatures.

Quality metrics

The journal Cell has an impact factor of 41.58 (2020 Journal Citation Reports), a Scopus CiteScore of 63.40 (22th of November 2021) and is in the first quartile (Q1) in Biochemistry, Genetics and Molecular Biology (miscellaneous) according to Scimago (SJR 2020 26.30).

Journal authorization

DIEGO PIDAL SUAREZ, con DNI 76963132A, autoriza a Bernardo Rodriguez Martín a incluir sus ilustraciones en la tesis doctoral titulada: "THE IMPACT OF TRANSPOSABLE ELEMENTS ON THE STRUCTURE AND FUNCTION OF THE CANCER GENOME".

En Villaviciosa, a 29 de Marzo de 2022.

El autor de las ilustraciones,

Diego Pidal Suárez

This PhD dissertation provides further insights on the impact of retrotransposon mobilization in the cancer genome. Through the analysis of a large cohort of 2,954 cancer whole genomes compiled by the Pan-Cancer Analysis of Whole Genomes Consortium, we have investigated the patterns of somatic activity for mobile elements across 38 different tumour types, uncovered novel mechanisms of mutation mediated by LINE-1 elements and characterised more than 100 full-length LINE-1 loci active in cancer.

Mobile elements, cancer genomes and computational biology are the best cocktail for fun. Enjoy reading.