Pan-cancer analysis of whole genomes reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours

Bernardo Rodriguez-Martin[1,2], Eva G. Alvarez[1,2,§], Adrian Baez-Ortega[3,§], Jonas Demeulemeester[4,5], Young Seok Ju[6], Jorge Zamora[1,2], Harald Detering[7], Yilong Li[8], Alicia L. Bruzos[1,2], Stefan C. Dentro[9,4,10], Ana Dueso-Barroso[11,12], Daniel Alderjan[13], Marta Tojo[1,2], Nicola D. Roberts[8], Miguel G. Blanco[14,15], Paul A. W. Edwards[16,17], Joachim Weischenfeldt[18,19], Martin Santamarina[1,2], Montserrat Puiggros[11], Zechen Chong[20], Ken Chen[20], Eunjung Alice Lee[21], Jeremiah A. Wala[22,23], Keiran Raine[8], Adam Butler[8], Sebastian M. Waszak[19], Fabio C. P. Navarro[24,25], Steven E. Schumacher[22,23], Jean Monlong[26], Francesco Maura[27,28,8], Niccolo Bolli[27,28], Guillaume Bourque[26], Mark Gerstein[24,25], Peter J. Park[21], Rameen Berroukhim[22,23], David Torrents[11,29], Jan Korbel[19], Inigo Martincorena[8], Peter Van Loo[4,5], Haig H. Kazazian[13], Kathleen H. Burns[30,13], Peter J. Campbell[8,31,*] & Jose M. C. Tubio[1,2,8,*]

[§]These authors contributed equally to the manuscript

* These authors contributed equally to the manuscript

[1]Mobile Genomes and Disease, The Biomedical Research Centre (CINBIO), University of Vigo, Vigo 36310, Spain
[2]Department of Biochemistry, Genetics and Immunology, Faculty of Biology, University of Vigo, Vigo 36310, Spain
[3]Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, UK
[4]The Francis Crick Institute, London, UK
[5]Department of Human Genetics, KU Leuven – University of Leuven, Leuven, Belgium
[6]Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea

[7]Evolutionary Genomics, The Biomedical Research Centre - CINBIO, University of Vigo, 36310 Vigo, Spain
[8]Cancer Ageing and Somatic Mutation Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB101SA, UK
[9]Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA. UK
[10]Big Data Institute, University of Oxford, Oxford, UK
[11]Joint BSC-CRG-IRB Research Program in Computational Biology, 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]Institute for Genetic Medicine, Johns Hopkins University School of Medicine - Baltimore, MD USA
[14]Departamento de Bioquímica e Bioloxía Molecular, CIMUS, Universidade de Santiago de Compostela, 15706 Santiago de Compostela, Spain
[15]Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Universidade de Santiago de Compostela, 15706 Santiago de Compostela, Spain
[16]Department of Pathology, University of Cambridge, Cambridge, UK
[17]Cancer Research UK Cambridge Institute, Cambridge, UK
[18]Biotech Research & Innovation Centre (BRIC); Finsen Laboratory, Rigshospitalet, Copenhagen, Denmark
[19]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
[20]Department of Bioinformatics and Computational Biology, The University of Texas Maryland Anderson Cancer Center, Houston, Texas, USA
[21]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
[22]The Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[23]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[24]Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT
[25]Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT
[26]Department of Human Genetics, McGill University, Montreal, H3A 1B1, Canada
[27]Department of Oncology and Onco-Hematology, University of Milan, Milan, Itlay
[28]Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[29]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[30]Department of Pathology, Johns Hopkins University School of Medicine - Baltimore, MD USA
[31]Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK

Address for correspondence:

Dr Jose M. C. Tubio,
Mobile Genomes and Disease,
The Biomedical Research Centre (CINBIO),
University of Vigo,
Vigo 36310,
Spain

Phone: +34 986 130 053
e-mail: jt14@sanger.ac.uk

Dr Peter Campbell,
Cancer Ageing and Somatic Mutation Programme,
Wellcome Trust Sanger Institute,
Hinxton,
Cambridgeshire CB10 1SA,
UK.

Phone: +44 1223 494951
Fax: +44 1223 494809
e-mail: pc8@sanger.ac.uk

About half of all cancers have somatic integrations of retrotransposons. To characterize their role in oncogenesis, we analyzed the patterns and mechanisms of somatic retrotransposition in 2,774 cancer genomes from 31 histological cancer subtypes. L1 insertions emerged as the third most frequent type of somatic structural variation in cancer. Occasionally, aberrant L1 integrations can remove vast, megabase-scale regions of a chromosome, sometimes involving essential regions for chromosomal instability, namely centromeres and telomeres. We find L1-mediated deletions may promote cancer-causing lesions through direct removal of tumour suppressor genes, or triggering events that result in oncogene amplification. L1 retrotransposition can also cause interchromosomal rearrangements, and tandem duplications of megabase-scale regions. These observations illuminate a relevant role of L1 retrotransposition in remodeling the cancer genome, with potential implications in the initiation and/or development of human tumours.

Long interspersed nuclear element (LINE)-1 (L1) retrotransposons are widespread repetitive elements in the human genome, representing 17% of the entire DNA content[1,2]. Using a combination of cellular enzymes and self-encoded proteins with endonuclease and reverse transcriptase activity, L1 elements copy and insert themselves at new genomic sites, a process called retrotransposition. Most of the ~500,000 L1 copies in the human gene are truncated, inactive elements not able to retrotranspose; in contrast, 100-200 L1 loci are active in the human population, of which a small number are highly active copies termed hot-L1s[3-5]. These L1 source elements are usually transcriptionally repressed in healthy genomes, but epigenetic changes occurring in tumours may promote their expression and allow them to retrotranspose[6,7]. Somatic L1 retrotransposition most often introduces a new copy of the 3' end of the L1 sequence, and

through it can also mobilize unique DNA sequences located immediately downstream of the source element, a process called L1-mediated transduction[7,8]. L1 retrotransposons can also promote the somatic trans-mobilization of processed pseudogenes, which are copies of messenger RNAs that have been reverse transcribed into DNA and inserted into the genome using the enzymes of active L1 elements[9,10].

Approximately 50% of all human tumours have somatic retrotransposition of L1 elements[7,11-13]. Previous analyses indicate that, although a fraction of somatically acquired L1 insertions in cancer may influence gene function, the majority of retrotransposon integrations in a single tumour represent passenger mutations with little or no effect on cancer development[7,11]. However, L1 insertions are capable of promoting genomic alterations apart from canonical L1 insertion events[2], and these remain largely unexplored in human cancer[14].

To further understand the roles of retrotransposons in cancer, we developed novel strategies to analyze the patterns and mechanisms of somatic retrotransposition in 2,774 cancer genomes from 31 histological cancer subtypes within the framework of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project [P. J. C. et al., manuscript in preparation], many of which have not been previously evaluated for retrotransposition. This work illuminates novel, hidden patterns and mutational mechanisms of structural variation in human cancer mediated by L1 retrotransposition. We find that aberrant integration of L1 retrotransposons has a major role in remodeling cancer genome architecture, mainly by promoting megabase-scale deletions that, occasionally, generate genomic consequences that may promote cancer development through the removal of tumour suppressor genes, such as *CDKN2A*, or triggering amplification of oncogenes,

such as *CCND1*.

RESULTS

The landscape of somatic retrotransposition in the largest cancer dataset

We identified 20,194 somatically acquired retrotransposition events. Overall, 43% of all cancer genomes account for at least one retrotransposition, being these events more frequent in lung squamous carcinoma (Lung-SCC), esophageal adenocarcinoma (Eso-AdenoCA), and colorectal adenocarcinoma (ColoRect-AdenoCA), where >90% of the samples from these tumour types bear one or more somatic events (**Fig. 1a**, **Supplementary Table 1**). Eso-AdenoCA is the tumour type with the highest retrotransposition rate, in which 29% of the cancer genomes exceeds one hundred somatic retrotranspositions, followed by head-and-neck squamous (Head-SCC, with 11%), Lung-SCC (6%), and ColoRect-AdenoCA (3%). These four tumour types alone account for 68% of all somatic events in pan-cancer, while they just represent 10% of the samples.

Retrotranspositions were classified into 7 categories (**Fig. 1b** and **Supplementary Fig. 1**). With 98% of the events, L1 integrations (i.e. solo-L1 and L1-transductions) overwhelmingly dominates the landscape of somatic retrotransposition across all tumours, making L1 insertions the third most frequent type of somatic structural variation in pan-cancer after tandem duplications (49,528 total events) and deletions (44,715), and followed by far by unbalanced translocations (13,727) [Y. L. et al., manuscript in preparation]. Trans-mobilization of processed pseudogenes and rearrangements (mainly deletions) promoted by L1 integration, with 228 and

96 events respectively, are poorly represented classes in the retrotransposition dataset that, in general, are more frequent in cancer types with high L1 activity rates. Genomic landscapes in the cohort reveal that although in the majority of cancer genomes L1 events predominate, we still observe cases where pseudogenes and L1-mediated chromosomal rearrangements makeup remarkably the retrotransposition scene (**Fig. 1c**).

The genome-wide analysis of the distribution of 19,705 somatic L1 insertions revealed a dramatic variation of L1 retrotransposition rate across the cancer genome (**Fig 2a**). At a megabase scale, we find that L1 retrotransposition density is strongly correlated with DNA replication timing (Spearman's $\rho = 0.69$, $P \sim 0$) (**Fig. 2b**), and negatively correlated with expression level (Spearman's $\rho = -0.41$, P = 6.3e-127) (**Fig. 2c**) and gene-density ($r = -0.18$; p = 7.4e-24). Poisson regression revealed that 46.2% of the total variance in L1 retrotransposition rate in cancer could be explained by combining these genetic features, with replication timing alone accounting for 44.61% of the variance. We also evaluated the association of L1 retrotransposition density with chromatin state, which revealed a rate of somatic L1 insertion six times higher in repressed chromatin than in euchromatin (P = 1.889e−243) (**Fig. 2d**). Overall, these data confirm that somatic L1 integration in cancer is heavily biased towards late replicating, lower expressed, gene-poor, heterochromatic regions of the genome [7,12].

We identified ~32% (6,317/19,906) somatic retrotranspositions inserted within gene regions including promoters, being 198 events associated with cancer genes. As a consequence of the L1 tendency to integrate in heterochromatic-like regions, we find somatic retrotranspositions in the

PCAWG dataset are enriched in lowly expressed genes compared to those that are highly expressed (**Fig. 2e**). Accordingly, although we find evidence that some L1 insertions may influence gene expression, we did not find strong support that L1 retrotransposition altered the function of any of the 35 cancer genes bearing somatic L1 retrotranspositions from 28 tumours with available transcriptome. Specifically, we identified 4 genes with L1 retrotranspositions in the proximity of promoter regions showing significant over-expression when compared to the expression in the remaining samples from the same tumour type (Student's t-test, $q < 0.10$; **Supplementary Fig. 2**). This includes one head-and-neck tumour, D015591, with a somatic L1 event of uncertain importance integrated in a promoter region of the *ABL2* oncogene. In addition, we analyzed the potential of processed pseudogenes to promote functional consequences in human tumours[10], finding evidence for aberrant fusion transcripts arising from inclusion of 14 processed pseudogenes in the target host gene, and expression of 3 processed pseudogenes landing in intergenic regions (**Supplementary Fig. 3**).

We used L1-3' transduction events mobilized somatically to trace L1 activity to specific source elements. This shows 124 germline L1 loci in the human genome are responsible for most of the genomic variation generated by retrotransposition in cancer. Fifty-two of these loci represent novel, previously unreported source elements in human cancer [S. M. W. et al., manuscript in preparation] (**Supplementary Table 2**). We analyzed the relative contribution of individual source elements to retrotransposition burden across cancer types, finding that retrotransposition is generally dominated by five hot-L1 source elements and that alone give rise to half of all somatic transductions (**Fig. 3a**). This analysis revealed different behaviors of L1 source elements, with two extreme patterns of hot-L1 activity, which we have termed Strombolian and

Plinian, marked by their similarity to the patterns of volcano eruption types (**Fig. 3b**). Strombilian source elements represent the calmest type of hot-L1 activity, characterized by the production of small amounts of retrotranspositions in individual tumour samples, but they are often active leading them to contribute significantly to overall retrotransposition in cancer. On the contrary, source elements with Plinian hot activity are rarely active in a tumour but their eruption is violent, promoting large amount of retrotranspositions in single tumour samples. At the individual tumour level, although we observe that the number of active source elements in a single cancer genome may vary from 1 to 22, typically only 1 to 3 loci are operative (**Fig. 3c**). Occasionally, somatic L1 integrations that retain a full structure may also act as source for subsequent somatic retrotransposition events[7], and may reach hot activity rates, leading them to command retrotransposition in a given tumour. For example, in a remarkable Head-SCC tumour, DO14343, we identify one somatic L1 integration at 4p16.1 that promotes 18 transductions, with the next most active element being a germline L1 locus at 22q12.1 accounting for 15 transductions (**Supplementary Fig. 4**).

Distinctive patterns of somatic L1 retrotransposition reveal hidden mutational mechanisms in cancer

In a retrotransposition analysis of cancer genomes with high somatic L1 activity rates, we observed that some L1 retrotransposition events followed a distinctive pattern consisting of one-single cluster of reads, which is associated with one breakpoint of a copy number loss, and whose mates unequivocally identify one extreme of a somatic L1 integration with, apparently, no reciprocal cluster supporting the other extreme of the L1 insertion (**Fig. 4a**). Analysis of the associated copy number changes identified the missing L1 reciprocal cluster far away, at the

second breakpoint of the copy number loss, indicating that this pattern represents a deletion occurring in conjunction with the integration of a L1 retrotransposon (**Fig. 4b**). These rearrangements, called L1-mediated deletions, are most likely the consequence of an aberrant mechanism of L1 integration, in which a molecule of L1 cDNA is paired to a distant 3´-overhang from a preexisting double strand DNA brake generated upstream of the initial integration site, and the DNA region between the break and the original target site is subsequently removed by aberrant repair[15]; although other alternative models have been proposed[15-18].

We developed specific algorithms to explore L1-mediated deletions on a large scale, across all the pan-cancer tumours, which identified 96 somatic events matching the patterns described above that promote deletions larger than 500 nucleotides (**Supplementary Table 3**). The reconstruction of the sequence at the breakpoint junctions in each case supports the presence of an L1 element – or L1-transduction – sequence and its companion polyadenylate tract, indicative of retrotransposition. No target site duplication is found, which is typically absent in L1-mediated deletions[15]. To confirm that these rearrangements are mediated by the integration of a single intervening retrotransposition event, we explored the pan-cancer dataset looking for somatic L1-mediated deletions where the L1 sequences at both breakpoints of the deletion can be unequivocally assigned to the same L1 insertion. These include small deletions and associated L1 insertions shorter than the library size, allowing sequencing read-pairs to overlay the entire structure. For example, in a lung tumour, DO27334, we identified a deletion involving a 1.1 kb region at 19q12 with hallmarks of being generated by an L1 element (**Fig. 4c**). In this rearrangement we find two different types of discordant read-pairs at the deletion breakpoints: one cluster that supports the insertion of an L1 element, and a second that spans the L1 event and

supports the deletion. Another type of L1-mediated deletion that can unequivocally be assigned to one-single L1 insertion event is represented by those deletions generated by the integration of orphan L1-transductions. These transductions represent bits of unique DNA sequence located downstream of an active L1 locus, which are mobilized without the companion L1[7]. For example, in one esophageal tumour, DO50362, we find the loss of a 2.5 kb long region at chromosome 3 in which the breakpoints of the deletion revealed one type of discordant reads only, which support the insertion of one-single DNA region transduced from an L1 loci located at chromosome 7 (**Fig. 4d**).

To further validate L1-mediated deletions, we performed whole-genome sequencing on two cancer cell-lines with high retrotransposition rates, NCI-H2009 and NCI-H2087, encompassing mate-pair libraries with long insert sizes (3kb and 10kb) that would exceed the insertion event at the deletion boundaries. In these samples, our algorithms identified 16 events with the hallmarks of L1-mediated deletions, in which the mate-pairs data confirmed one-single L1-derived (i.e., solo-L1 or L1-transduction) insertion as the cause of the copy number loss, and identified the sizes of the deletion and the associated insertion (**Supplementary Fig. 5**).

We have successfully reconstructed the L1 3'-extreme insertion breakpoint sequence for 87% (89/102) of the retrotransposition events associated with L1-mediated deletions, revealing the presence of a 3'-A/TTT-5' L1-endonuclease consensus cleavage site motif in 82% (73/89) of the events (**Fig. 4e**). This confirms that L1 machinery, through a target-site primed reverse transcription (TPRT) mechanism, is responsible for the integration of most of the L1 events causing neighboring DNA loss. Nonetheless, we observe 18% (16/89) instances where this

consensus site is not found, suggesting that a small fraction of L1-associated deletions may be the consequence of an L1-endonuclease-independent insertion mechanism [17,18]. Whatever the mechanism of L1 integration is operating here, taken together, these data indicate that the somatic integration of L1 elements is undoubtedly mediating the associated deletions.

L1 retrotransposition has a major impact on cancer genome architecture

Although L1-mediated deletions generally range from a few hundred to thousands of base pairs (**Supplementary Table 3**), occasionally, they can remove vast, megabase regions of a chromosome with potential functional consequences for a cancer genome. For example, in esophageal tumour DO50410, we find a 45.5 Mb interstitial deletion involving the p31.3-p13.3 regions from chromosome 1 (**Fig. 5a**), where both breakpoints of the rearrangement show the hallmarks of a deletion promoted by the integration of an L1 element. Here, the L1 element is 5'-truncated, which rendered a small L1 insertion, allowing a fraction of the sequencing read-pairs to span both breakpoints of the rearrangement, which unequivocally supports the same L1 event at both breakpoints of the deletion.

L1-mediated deletions can generate major rearrangements involving essential structures for the stability of a chromosome. In one lung tumour, DO27334, we found an interstitial L1-mediated deletion that promotes the loss of 51.1 Mb from chromosome X (**Fig. 5b**). Here, the deletion removes the centromere leaving two breakpoints, one at each of the chromosomal arms Xp and Xq. The analysis of the sequencing data revealed the integration of an L1 transduction from chromosome 22q12.1 within the deletion breakpoints, being the most likely cause of the

chromosomal loss. Theoretically, the rearrangement would result in a fusion of the distal regions from both chromosomal arms, generating a chromosome with no centromere.

Similarly, in another notable esophageal tumour, DO50365, we observe the integration of an L1 transduction from chromosome 14q23.1 associated with an interstitial deletion spanning 47.9 Mb at chromosome 5 (**Fig. 5c**). The deletion generates a putative shorter chromosome with no centromere that would most likely be lost during mitosis. Nonetheless, large-scale DNA loss involving a centromere is a relatively common feature in human cancers that can be also caused by both breakage-fusion-bridge[19] and chromothripsis[20], which cancer cells resolve through neocentromere formation or through acquisition of interchromosomal rearrangements that stabilize the new chromosomal configuration[20,21]. In this case, a closer analysis of the sequencing data at the 5q deletion breakpoint confirmed a fusion with 1q, where a piece of the transduction from chromosome 14q23.1 (the same region associated with the deletion) is bridging the interchromosomal rearrangement (**Fig. 5c** and **Supplementary Fig. 6**). The identification of a L1 endonuclease motif at one of the L1 integration breakpoints suggests that this complex pattern of genomic rearrangements, involving a deletion coupled with a translocation, may have been initiated by an aberrant L1 integration mechanism.

Our analyses revealed a subset of single-L1 clusters with no reciprocal cluster, and not always associated with a copy number change, suggesting that some of these rearrangements may correspond to hidden genomic translocations, linking two different chromosomes, in which L1 retrotransposition is involved. Consistent with this mechanism, we found evidence of L1 retrotransposition-associated translocations in the cancer cell-lines that were sequenced using

mate-pairs with 3kb and 10kb inserts. One of the samples, NCI-H2087, showed translocation breakpoints at 1q31.1 and 8q24.12, both with the hallmarks of L1-mediated deletions, where the mate-pair sequencing data identifies an orphan L1-transduction from chromosome 6p24 bridging both chromosomes (**Fig. 6a**). This interchromosomal rearrangement could be mediated by an aberrant operation of the mechanism of L1 integration, where a bit of the L1-transduction cDNA is wrongly paired to a second 3´-overhang from a preexisting double strand brake generated in a second chromosome[15].

We also found evidence that L1 integrations can cause tandem duplications of large genomic regions in human cancer. In the esophageal tumour DO50374, we identified two independent read clusters supporting the integration of a small L1 event, coupled with coverage drop at both breakpoints. The analysis of the copy number data revealed that the two L1 clusters demarcate the boundaries of a 22.6 Mb duplication that involves the 6q14.3-q21 region, suggesting that the L1 insertion could be the cause of such rearrangement (**Fig. 6b**). The analysis of the rearrangement data at the breakpoints identified read-pairs that traverse the length of the L1 insertion breakpoints, and the L1-endonuclease motif is the L1 3' insertion breakpoint, both confirming a single L1 event as the cause of a tandem duplication.

L1-mediated rearrangements can result in cancer-driving mutations in human cancer

Although L1-mediated deletions generally account for a low proportion of somatic retrotransposition events in a tumour, their potential to impact the function of a cancer genome is considerably higher than any other retrotransposition event. The simplest way by which L1-

mediated deletions may lead to the generation of oncogenic rearrangements is through the loss of tumour suppressor genes. In esophageal tumour DO50362, the integration of an L1-transduction from chromosome 7p12.3 into the short arm of chromosome 9 caused a 5.3 Mb deletion involving the 9p21.3-9p21.2 region. This led to loss of one copy of a key tumour suppressor gene, *CDKN2A* (**Fig. 7a**), a well-known mutational driver in many cancer types, including esophageal tumours[22-25]. Interestingly, the analysis of the sequence at the breakpoint junctions revealed that the L1 element inserted retained its original structure, meaning that it may remain active. As expected, the analysis of the sequencing data downstream revealed one somatic transduction promoted by the L1 element at the new insertion site, demonstrating that some L1 events that promote deletions are competent for retrotransposition (**Supplementary Fig. 7**).

Similarly, in a second esophageal tumour, DO50383, an L1 element integrated into chromosome 9 promotes an 8.6 Mb deletion encompassing the 9p22.1-9p21.1 region that removes one copy of the same tumour suppressor gene, *CDKN2A* (**Fig. 7b**). The analysis of other types of somatic variation in the region, revealed no inactivating structural or point mutations in the second copy of this cancer gene in any of the two samples affected and, although methylation alterations, which in this gene represent an important source of inactivation[23], could not be explored, cancer development is based on the continuous acquisition of mutations in these and other cancer genes, and it is highly likely that these L1-mediated deletions are seeding the genomic variability necessary for future clonal expansions in these tumours[26].

A second potential mechanism by which L1-mediated deletions could generate cancer-causing genomic alterations is through the promotion of telomere loss that subsequently triggers the

amplification of oncogenes. In the esophageal tumour DO50374, we identified one-single cluster of reads at the long arm of chromosome 11 with the typical hallmarks of an L1-mediated deletion that, unexpectedly, did not pair with any reciprocal cluster far away, indicating that a deletion is present but only one breakpoint could be found (**Fig. 8a**). Copy number data analysis across chromosome 11 revealed a complex pattern of copy number changes that differs upstream and downstream of the L1 integration site at 11q. Downstream of the L1 somatic event, we identified a pattern compatible with a 53.4 Mb deletion, extending from the referred breakpoint to the end of the chromosome including the telomere. Upstream of the L1 somatic event, from the relevant breakpoint towards the centromere, we observed different amplification strata of megabase regions that reach a maximum peak at a segment that contains the *CCND1* oncogene, a relevant driver gene that is commonly amplified in many human cancers[27]. Analysis of the paired-end mapping reads at the breakpoints of such amplified regions revealed that their boundaries are demarcated by fold-back inversion rearrangements (**Fig. 8a**) – a diagnostic pattern typically associated with breakage-fusion-bridge (BFB) repair [19,28]–.

Taken together, the patterns described above allowed us to reconstruct the history of the rearrangements as follows (**Fig. 8b**): First, a somatic L1 insertion event results in the deletion of a 53.4 Mb of 11q extending to the telomere. The atelomeric chromosome 11 is then subjected to BFB repair, a mechanism known to duplicate the chromosome and consequently generate an end-to-end chromosomal fusion centered around the L1 element (the "bridge") located at the breakpoint of the deletion. The resulting dicentric chromosome is broken when, in the mitosis, during cytogenesis the two centromeres are pulled to opposite poles of the dividing cell. The sequencing reads that would support this first BFB cycle are not visible, because of a limitation

of the sequencing library insert size employed, which is too short to traverse the L1 insertion at the boundaries of the bridge. Nevertheless, the newly broken chromosome is repaired again, generating a second bridge whose boundaries are demarcated by two clusters of reads from a fold-back inversion. These BFB cycles lead to rapid-fire amplification of the *CCND1* oncogene. The patterns observed here, may indicate that L1 integration is the initiating event that promotes a butterfly effect leading to the amplification of *CCND1*.

Somatic acquisition of telomere length abnormalities is one of the earliest genomic alterations occurring in the process of malignant transformation leading to cancer[19,20,28]. We looked in the pan-cancer dataset for similar patterns involving telomere loss mediated by L1 integration, and found 4 more events from 3 different cancer samples (**Supplementary Fig. 8**). Surprisingly, in one lung tumour, DO26976, we found almost identical rearrangements to the one described above (**Fig. 8c**). Here, a somatic L1-transduction promoted a 50.6 Mb deletion of the long arm of chromosome 11, including the telomere. We observe the hallmarks of L1-mediated deletion at the breakpoint, including an L1 integration that matches a dramatic coverage drop that extends downstream across 11q to the tip of the chromosome. Upstream of the L1 event, we see megabase-size amplification of chromosomal regions targeting the *CCND1* oncogene, with boundaries demarcated by a fold-back inversion indicating BFB repair. The independent occurrence of these patterns in two different tumour samples demonstrates a mutational mechanism mediated by L1 retrotransposition, which contributes to the initiation and/or development of cancer.

DISCUSSION

Here we characterize the patterns and mechanism of cancer retrotransposition on an unprecedented multidimensional scale, across thousands of cancer genomes, integrated with rearrangement, transcriptomic, and copy number data. This provides new perspective on a long-standing question: is activation of retrotransposons relevant in human oncogenesis? Our findings demonstrate that major restructuring of cancer genome can emerge out from aberrant L1 retrotransposition events in tumours with high retrotransposition rates, particularly in esophageal, lung, and head-and-neck cancers. L1-mediated deletions can promote the loss of megabase-scale regions of a chromosome that may involve centromeres and telomeres. It is likely that the majority of such genomic rearrangements would be harmful for a cancer clone. However, occasionally, L1-mediated deletions may promote cancer-driving rearrangements that involve loss of tumour suppressor genes and/or amplification of oncogenes, representing another mechanism by which cancer clones acquire new mutations that help them to survive and grow.

Relatively few germline L1 loci in a given tumour, typically 1-3 copies, are responsible for such dramatic structural remodeling. These include a subset of highly active, 'hot' L1 that are heritable structural variants in human populations and, overall, we identified 124 L1 source elements in human populations with the capacity to drive somatic retrotransposition in cancer. Given the role these L1 copies may play in some cancer types, we believe this work underscores the importance of characterizing cancer genomes in light of L1 retrotransposition.

FIGURE LEGENDS

Figure 1. Rates of somatic retrotransposition across human cancers. (A) For each cancer type included in the PCAWG project, proportions of analyzed tumour samples with more than 100 somatic retrotranspositions (red), between 10 and 100 (orange), between 1 and 10 (yellow), and zero (grey). (B) Frequency of retrotransposition events across cancers. Somatic retrotranspositions in PCAWG were classified into 7 categories. Here, only the four cancer types with higher retrotransposition rates are shown, together with a pan-cancer overview. The remaining cancer types are displayed in **Supplementary Fig. 1**. (C) Circos plots showing the genomic landscape of somatic retrotransposition in four representative samples. Chromosome ideograms are shown around the outer ring with individual rearrangements shown as arcs, each coloured according to the type of rearrangement. Note the spideweb-genome patterns that characterize samples with high L1 retrotransposition rates (DO50383, and DO14343). In DO27747 and DO27334, L1-mediated deletions and processed pseudogenes dominate. Same colors as above: total retrotranspositions (black), Solo-L1 integration (purple), L1-transduction (green), Alu (orange), SVA (yellow), processed pseudogene (blue), L1-mediated deletion (red).

Figure 2. Distribution of L1 retrotransposition density in the cancer genome and association with genome organization. (A) The variation of L1 retrotransposition density (grey bars) in the cancer genome is represented relative to the variation in other genomic features, including replication timing (blue lines), expression level (red line), proportion of heterochromatin (green bars) and euchromatin (yellow bars). The information is displayed in windows of 10 Mb. Note that L1 retrotransposition rate is elevated in windows enriched in heterochromatic domains, characterized by late replication and low expression, while L1 rate is repressed in more euchromatic regions. (B) Rate of somatic L1 insertions strongly correlates

with replication timing. (C) Rate of somatic L1 insertions strongly anti-correlates with expression. Here, gene expression profiles were an average of 91 cell lines from Cancer Cell Line Encyclopedia [29]. (D) L1 retrotranspositions acquired somatically are overrepresented in transcriptionally repressed (typically heterochromatic) regions of the cancer genome. Note that the relative abundance of L1 insertions in repressed chromatin is 6 times higher than in transcriptionally active chromatin. Error bars reflect Poisson confident intervals. Chromatin states were derived from ENCODE[30]. (E) Somatic retrotransposition is enriched in lowly expressed genes (<3 FPKM) relative to highly expressed genes. Here, expression data were pan-cancer transcriptomes.

Figure 3. The dynamics of L1 source elements activity in human cancer. (A) We analyzed the contribution of 124 germline L1 source loci to somatic retrotransposition burden in different human cancers. The total number of transductions identified for each cancer type is shown in a blue coloured scale. 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. (B) Two extreme patterns of hot-L1 activity, Strombolian and Plinian were identified. Dots show the number of transductions promoted by each source element in a given tumour sample. Arrows indicate violent eruptions in particular samples (Plinian source elements). (C) Distribution of numbers of active source elements per sample across tumour types with source element activity.

Figure 4. The hallmarks of somatic L1-mediated deletions revealed by copy number and paired-end mapping analysis. (A) In the retrotransposition analysis of DO50320, an Eso-AdenoCA sample with high L1 somatic activity rates, we found one-single cluster of reads at

chromosome X, which is associated with one breakpoint of a copy number loss, and whose mates unequivocally identify one extreme of a somatic L1 integration with, apparently, no reciprocal cluster supporting the other extreme of the L1 insertion. (B) The analysis of the associated copy number change at chromosome X identifies the missing L1 reciprocal cluster far away, at the second breakpoint of the copy number loss, and reveals a 3.9 Kb long deletion occurring in conjunction with the integration of a 2.1 Kb L1 somatic insertion. The sequencing data associated to this L1-mediated deletion show two clusters of discordant read pairs and clipped reads supporting both extremes of a L1 retrotransposon, including the poly-A at the 3' extreme of the element. (C) In a Lung squamous carcinoma, DO27334, a 34 bp truncated L1 insertion promotes a 1.1.Kb deletion at chromosome 19. Because the L1 insertion is too short, apart from the two clusters of discordant read pairs that typically support a L1 event, we also identify a pair of discordant read-pairs clusters that span the L1 event and support the deletion. (D) In an esophageal adenocarcinoma, DO50362, the integration at chromosome 3 of a 413 bp orphan L1-transduction from chromosome 7 causes a 2.5 Kb deletion, which is supported by two clusters of discordant read pairs whose mates map onto the same region at chromosome 7. (E) Reconstruction of the breakpoint sequence at the L1 3'-extreme from 39 retrotransposition events linked to deletions, revealed the presence of the 5'-TTT/A-3' L1-endonuclease consensus cleavage site motif in 32 of the events. Motifs found in 102 deletions are shown in **Supplementary Table 3**.

Figure 5. Somatic integration of L1 causes loss of megabase interstitial chromosomal regions in cancer. (A) Left: In an esophageal tumour, DO50410, we find a 45.5 Mb interstitial deletion at chromosome 1 that is generated after the integration of a short L1 event. We observe

a pair of clusters of discordant read pairs whose mates support both extremes of the L1 insertion. Because the L1 element event is smaller than the library insert size, we also identify a pair of reciprocal read clusters that span the L1 event and support the deletion. Right: Model for megabase L1-mediated interstitial deletions[15]. The integration of a L1 mRNA typically starts with an L1-endonuclease cleavage promoting a 3'-overhang necessary for reverse transcription. Then, the cDNA (-) strand invades a second 3'-overhang from a pre-existing double-strand break upstream of the initial integration site. L1-endonuclease A-TTT motif identifies TPRT L1-integration mechanism. (B) In an esophageal tumour, D024334, a transduction from chromosome 22 and its companion L1 element is integrated on chromosome X, promoting a 51.1 Mb deletion that removes the centromere. One reads cluster in positive orientation supports an inverted L1 element. One negative cluster supports a small region transduced from chromosome 22 that bears a poly-A tract. L1-endonuclease A-TTT motif identifies TPRT L1-integration mechanism. (C) Likewise, in a second esophageal adenocarcinoma, DO50365, a transduction from chromosome 14 and its companion L1 element integrates at chromosome 5, promoting a 47.9 Mb deletion involving centromere loss. A positive reads cluster supports the integration in inverted orientation of the L1 element (L1-endonuclease TTT-A motif identifies TPRT L1-integration mechanism), while two negative clusters at the 3' extreme of the insertion support both the presence of a region transduced from chromosome 14 associated with the deletion, and an interchromosomal rearrangement between chromosomes 1 and 5 bridged by a small piece of the chromosome 14 transduction (see **Supplementary Fig. 6** for details).

Figure 6. Somatic L1 integration promotes translocations and tandem duplications of megabase regions in cancer. (A) Left: In a cancer cell-line, NCI-H2087, we find an

interchromosomal translocation, between chromosomes 8 and 1, mediated by a region transduced from chromosome 6, which acts as a bridge and joins both chromosomes. Although this event was originally identified using long-insert mate-pairs sequencing data (main text), the size of the insertion event is short enough to be revealed by standard paired-end sequencing, which is shown here. We observe two read clusters, positive and negative, demarcating the boundaries of the rearrangement, whose mates support the transduction event. In addition, two reciprocal clusters span the insertion breakpoints, supporting the translocation between chromosomes 8 and 1. L1-endonuclease A-TTT motif identifies TPRT L1-integration mechanism. Right: A model for megabase L1-mediated interchromosomal rearrangements. L1-endonuclease cleavage promotes a 3'-overhang in the negative strand, retrotranscription starts, and the cDNA (-) strand invades a second 3'-overhang from a pre-existing double-strand break occurring in a different chromosome, leading to translocation. (B) Left: In an esophageal tumour, DO50374, we find a 22.6 Mb tandem duplication at the long arm of chromosome 6. Note that the copy number states vary between 2 and 3 copies because this sample was predicted to have undergone a whole-genome duplication event [S. C. D. et al., manuscript in preparation]. The analysis of the sequencing data at the boundaries of the rearrangement breakpoints reveals two clusters of discordant read pairs whose mates support the involvement of a L1 event. Because the L1 element is sorter than the library size, we also find two reciprocal clusters that align 22.6 Mb apart on the genome and in opposite orientation, spanning the insertion breakpoints and confirming the tandem duplication. L1-endonuclease TTT-A degenerate motif identifies TPRT L1-integration mechanism. Right: Model that explains the megabase tandem rearrangement shown in left. Large direct tandem duplication can be generated if the cDNA (-) strand invades a

second 3'-overhang from a pre-existing double-strand break occurring in a sister chromatid, and downstream of the initial integration site locus.

Figure 7. L1-mediated deletions promote loss of tumour suppressor genes. (A) In esophageal tumour DO50362, the somatic integration at chromosome 9 of a transduction from chromosome 7 and its companion L1 element, promotes a 5.3 Mb deletion involving loss of one copy of the tumour suppressor gene *CDKN2A*. The sequencing data shows a positive cluster of reads whose mates map onto the 5' extreme of a L1, and a negative reads cluster that contain split reads matching a poly-A and whose mates map onto a region transduced from chromosome 7. L1-endonuclease A-TTT motif identifies TPRT L1-integration mechanism. (B) Similarly, in a second esophageal adenocarcinoma, DO50383, the integration of a L1 retrotransposon generates an 8.6 Mb deletion involving the same tumour suppressor gene, *CDKN2A*. The sequencing data reveals two clusters, positive and negative, whose mates support the integration of the L1 event, together with clipped reads that precisely mark the insertion breakpoint to base pair resolution. L1-endonuclease A-TT degenerate motif identifies TPRT L1-integration mechanism.

Fig. 8. L1-mediated deletions promote amplification of oncogenes through activation of breakage-fusion-bridge repair. (A) In an esophageal adenocarcinoma, DO50362, on the long arm of chromosome 11 we identify one-single cluster of reads in positive orientation whose mates support the integration of a L1 retrotransposon, and the analysis of the sequence at the breakpoints of the rearrangement revealed the L1-endonuclease cleavage site motif. At the L1 insertion breakpoint, we also observe the origin of a copy number change supporting the loss of 53 Mb from the long arm of the chromosome that includes the telomeric region. Upstream to the

L1 event, we observe different levels of amplification, and the presence of two reciprocal clusters of discordant read pairs revealing a fold-back inversion, a diagnostic pattern typically associated with breakage-fusion-bridge repair. The temporal order of the two major rearrangement events are marked with (1) and (2). The sequencing reads that would support the first BFB cycle are not visible, because of a limitation of the sequencing library insert size employed, which is too short to traverse the L1 insertion at the boundaries of the bridge (B) Model for the loss of the long arm of chromosome 11 through integration of a L1 retrotransposon, and subsequent repair through breakage-fusion-bridge that triggers amplification of oncogene *CCND1*. (C) We found almost identical rearrangement patterns in a lung cancer, DO25976, where the integration of a L1 retrotransposon is associated with loss of 50 Mb of the long arm of chromosome 11 that includes the telomere, and activates breakage-fusion-bridge repair leading to amplification of *CCND1*.

SUPPLEMENTARY TABLES LEGENDS

Supplementary Table 1. Counts of retrotranspositions by sample and cancer type

Supplementary Table 2. List of germline L1 source elements with counts per sample

Supplementary Table 3. Features of L1-mediated deletions (>500 bp) analyzed in this study

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Frequency of retrotransposition events across all cancer types in PCAWG. Points represent the number of retrotransposition events per sample for each retrotransposition category.

Supplementary Figure 2. Gene expression effects associated with L1 retrotranspositions. (A) A volcano plot representation of the impact of L1 insertion in cancer genes showing the gene expression change (x axis) and inverted significance (y axis). Red dots indicate the significant associations under q value < 0.1. This analysis revealed 2 L1 retrotranspositions where the target cancer gene (*ABL2* and *RB1*) is significantly over-expressed compared to the remaining tumours from the same cancer type (Student's t-test, q < 0.10). Nonetheless, these two events are of uncertain importance: the first (*ABL2*) is a L1 inserted in an alternative promoter of the oncogene, but the structural analysis of the integrated L1 revealed a truncated element that has lost the promoter region; the second (*RB1*) is a tumour suppressor gene. (B) Up-regulation of the *ABL2* oncogene in tumour DO15591, a Head-SCC. The expression of the same oncogene in other Head-SCC samples from the PCAWG dataset are also shown. (C) The analysis of RNA-seq data in genes with L1-retrotransposition in promoter regions showed significant upregulation in additional three genes. Volcano plot represents gene expression change of the gene (x axis) and inverted significance (y axis). Red dots indicate the significant associations under q value < 0.1.

Supplementary Figure 3. Expression of processed pseudogene somatic insertions. We found evidence for expression of 17 processed pseudogenes mobilized somatically, including aberrant fusion transcripts arising from inclusion of 14 processed pseudogenes in the target host gene, which are represented here Arcs with arrows within the circos indicate the processed pseudogene retrotransposition event, connecting the source processed pseudogene (underlined and bold) with the corresponding integration region. Target site is denoted as intergenic, when integration

occurs out of gene boundaries, or with the host gene name in italics when integration is within a gene. In the outermost layer of the figure, we represent the predicted processed pseudogene-host gene transcripts. Green and blue boxes represent the regions in the fusion transcript that correspond to the host gene and processed pseudogene, respectively; thinner green blocks represent 3' and 5' UTRs in transcripts of the host gene; and internal arrows indicate the coding direction. Thin black lines connecting green and blue boxes represent introns, with (continuous) or without (dashed) direct RNA-seq reads support. Split and discordant read pairs supporting a fusion transcript are shown above the representation of the corresponding predicted transcript. For each host gene mRNA, we have inferred the coding potential of each fusion transcript, which is shown underneath the fusion transcript representation. Start codon is denoted as ATG, termination codon as STOP, and uncertain termination is represented using dots.

Supplementary Figure 4. Somatic source elements may dominate retrotransposition in a tumour. In an esophageal adenocarcinoma, DO50383: (Left) distribution of numbers of transductions promoted by single L1 source elements (orange: somatic source elements; blue: germline source elements). A somatic source element at 4p16.1 commands somatic retrotransposition in this sample with 18 transductions, followed by a germline source element at 22q12.1 with 15 transduction events. (Right) circos plot showing somatic transductions promoted by the somatic source element at 4p16.1 in tumour DO50343.

Supplementary Figure 5. Validation of L1-mediated deletions by mate-pair sequencing data analysis. In order to further validate L1-deletions, we performed mate-pair sequencing of long-inserts libraries (4 Kb and 10 Kb) on two cancer cell-lines with high-retrotransposition

rates. Here, it is shown validation of a deletion 10.4 Kb long promoted by a 768 bp L1 insertion in the cancer cell-line NCI-H2009. The L1 element inserted within the deletion breakpoints is too long to be characterized using standard paired-end sequencing libraries, but the mate-pairs successfully span the breakpoints of the deletion and confirm a single L1 insertion associated with the rearrangement.

Supplementary Figure 6. L1 insertion promotes complex rearrangements. In esophageal adenocarcinoma DO50365, complex rearrangements associated with L1 transduction inserted on chromosome 5. Original transcript consists of a partnered L1 transduction from chromosome 14q23.1. Companion L1 element is inserted on the short arm of chromosome 5 (TTT-A motif), while a small piece (~100 bp long) from the 14q23.1 transduction is jointly integrated on 5q and 1q, by an abnormal L1 integration mechanism that generates both a 47.9 deletions that removes the centromeric region, and an interchromosomal fusion between chromosomes 5q and 1q.

Supplementary Figure 7. Some L1s mediating deletions are transduction-competent. (A) Circos plot summarizing the three concatenated retrotransposition events shown in B. First event, an L1-transduction mobilized from chromosome 7 is integrated into chromosome 9. Second event, this insertion concomitantly causes a 5.3 Mb 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 Mb deletion generated by the integration of a transduction from chromosome 7, and reveals a L1-event with full-length structure. Five kilobases downstream, a positive reads cluster supports a transduction from this L1-retotransposition event into chromosome X.

Supplementary Figure 8. L1 integration may cause telomere loss. (A) In a Head-SCC, D14250, deletion of 1.9 Mb at the short arm of chromosome 10, which involves the telomeric region, is associated with the somatic integration of a L1 retrotransposon. (B) In another Head-SCC, DO14343, two independent L1 events promote deletion of both ends of chromosome 5. (C) In Lung-SCC DO26976, the aberrant integration of a L1 event bearing 5' and 3' transductions causes a complex rearrangement with loss of 50.5 Mb from the long arm of chromosome 11 that includes the telomere.

ONLINE METHODS

Sequencing data

We analysed whole genome sequencing data from 2,774 tumours and their matched normal samples obtained within the framework of the Pan-Cancer Analysis of Whole Genomes project (PCAWG), and integrated with RNA-sequencing data from 1,222 donors with genome data (P. J. C. et al., manuscript in preparation).

Identification of mobile element insertions and L1 source element discovery

Non-reference mobile element insertions (MEIs), including L1, L1-mediated transductions, Alu, SVA and ERVK insertions; were identified with TraFiC-mem v1.1.0 (https://gitlab.com/mobilegenomes/TraFiC), an improved version of the TraFiC (Transposon Finder in Cancer) algorithm[7]. TraFiC-mem is based on discordant read-pair analysis as TraFiC, but it uses Bwa-mem instead of RepeatMasker as search engine for the identification of

retrotransposon-like sequences in the sequencing reads and it incorporates an additional module for reconstructing the insertion breakpoints through local *de novo* assembly. TraFiC-mem was used to jointly call germline and somatic MEIs in each tumour/normal pair. Insertions length, orientation, target site duplication and structure are parameters that were inferred through assembly of the involved discordant read-pairs and subsequent alignment of the assembled contigs to consensus retrotransposon sequences. Filtering of somatic MEI candidates was performed following the same criteria defined previously[7], but with an additional step consisting of the removal of somatic candidates if they match a germline retrotransposition of the same family called in the 1,000 Genomes Project Phase 3 dataset[31]. Finally, annotation of MEIs was performed using the software ANNOVAR[32], gencode v19 annotation[33], and the Cancer Gene Census database[34].

To identify novel (previously unreported) germline L1 source elements, we used the same method described previously[7], relying on the detection of unique (non-repetitive) DNA regions retrotransposed somatically elsewhere in the cancer genome from a single locus matching the 10 Kb downstream region of a reference full-length L1 element, or a putative non-reference polymorphic L1 element detected by TraFiC in the pan-cancer dataset. When transduced regions were derived from the downstream region of a putative L1 event present in the tumour genome but not in the matched-normal genome, we catalogued these elements as somatic L1 source loci.

Identification of processed pseudogene insertions

TraFiC-mem was the principal algorithm employed in the identification of somatic insertions of processed pseudogenes. The method relies on the same principle as for the identification of somatic MEI events, through the detection of two reciprocal clusters of discordant read-pairs,

namely positive and negative, that supports an insertion in the reference genome, but differs from standard MEI calling in where the read-mates map, as here it is required that mates must map onto exons belonging to a same source gene. To avoid misclassification with inter and intrachromosomal translocations that involve coding regions, TraFiC-mem reconstructs the insertion breakpoint junctions looking for hallmarks of retrotransposition, including polyadenylate tract and target site duplication.

Evaluation of processed pseudogenes expression

We analysed the Pan-cancer RNA sequencing data to identify and characterize the transcriptional consequences of somatic processed pseudogene integrations. We interrogated RNAseq data (split reads and discordant read pairs) looking for chimeric retrotrocopies involving processed pseudogenes and target genomic region. For each processed pseudogene insertion somatic call, we extracted all the RNA sequencing reads (when available) mapping the source gene and the insertion target region, together with the RNA-seq unmapped reads for the corresponding sample. Then, we used these reads as query of BLASTn[35] searches against a database containing all isoforms of the source gene described in RefSeq[36], together with the genomic sequence in a [-5 Kb, +5 Kb] range around the processed pseudogene integration site. Finally, we looked for RNA-seq read-pairs and/or RNA-seq split-reads that support the joint expression of processed pseudogene and target site. All expression signals were confirmed by visual inspection.

Identification of L1-mediated deletions

Each independent read cluster identified by TraFiC-mem and supporting the integration of a L1 retrotransposition event (i.e., those clusters of discordant read pairs with apparently no reciprocal

cluster within the proximal 500 bp, and whose mates support a L1 retrotransposition somatic event) was interrogated for the presence of an associated copy number change in its proximity (see 'copy number analysis' section below). Briefly, we looked for copy number change calls from working group 11 of the Pan-Cancer project (PCAWG-11) where the upstream breakpoint matches an independent L1 cluster in positive orientation, the downstream breakpoint from the same copy number change matches an independent L1 cluster in negative orientation, and the reconstruction of the structure of the putative insertion causing the deletion is compatible with one-single retrotransposition event. In addition, because we detected that some small L1-mediated deletions – usually below 10 Kb – are missing when using the copy number data described above, we followed an alternative strategy for the identification of deletions below 10 Kb. Briefly, first, we looked for a coverage drop in the proximity of each independent cluster, identified by obtaining a series of read depth ratios between the downstream and upstream flanking regions of each independent cluster, using different window sizes; second, we selected those independent reciprocal clusters, located less than 10 Kb apart, that were associated with a copy number change that extends from the positive cluster towards the negative, and vice versa, and where the coverage drop size matched the length of the distance that separates both reciprocal clusters; and, third, the reconstruction of the structure of the putative insertion causing the deletion is compatible with one-single retrotransposition event. The resulting L1-mediated deletion candidates were subsequently confirmed via visual inspection using integrative genomics viewer (igv)[37].

Validation of L1-mediated rearrangements in cancer cell-lines

Due to the unavailability of pan-cancer DNA samples, we performed validation of 20 somatic L1-mediated rearrangements, mostly deletions, identified in two cancer cell-lines with high retrotransposition rates, namely NCI-H2009 and NCI-H2087. For this purpose, we performed 10x mate-pair whole genome sequencing using libraries with two different insert sizes, 4 Kb and 10 Kb, which can span the integrated L1 element that cause the deletion, allowing validation of the involvement of L1 in the generation of such rearrangements. Mate-pair reads (100 nucleotides long) were aligned to the human reference build hg19 by using BWA-mem[38] on default settings, with the exception of the mean insert size. Then, for each candidate L1-mediated rearrangement we looked for discordant mate-pair clusters that span the breakpoints and support the L1-mediated event.

Copy number analysis

Copy number profiles were derived by working group 11 of the Pan-Cancer Analysis of Whole Genomes project (PCAWG-11) using a consensus approach combining six different state-of-the-art copy number calling methods (S. C. D. et al., manuscript in preparation). GC content-corrected LogR values were extracted from Battenberg results, smoothed using a running median, and transformed into copy number space according to $n = (2(1-\rho) + \psi\rho)2^{LogR}/\rho$ where ρ and ψ are the PCAWG-11 consensus tumor purity and ploidy, respectively.

Identification of genomic rearrangements

Genomic rearrangements were derived by working group 6 of the Pan-Cancer Analysis of Whole Genomes project (PCAWG-6) by combining the structural variant calls from four independent calling pipelines. Structural variants were grouped into structural variants clusters, which were

classified into one of several somatic rearrangement events (Y. L. et al., manuscript in preparation; J. W. et al., manuscript in preparation).

Evaluation of the impact of retrotransposition insertions in gene expression

To study the transcriptional impact of a somatic L1 insertion within a gene, we used RNA-seq data to compare gene expression levels at genes with and without somatic L1 insertion. We used FPKM values calculated through Cufflinks software[39]. For each somatic L1 insertion within a gene, we compared the gene FPKM between sample having the insertion (study sample) against the remaining samples in same tumour type (control samples). Using the distribution of gene expression levels in control samples, we calculated the normalized gene expression differences.

Correlation between L1 insertion density and genomic features

Gene density was calculated as the fraction of nucleotides covered by Gencode v19 protein coding genes (including introns) per 1-Mb window. Average gene expression per Mb was calculated using 91 cell lines from the Cancer Cell Line Encyclopedia (CCLE) [29]. DNA replication timing was expressed on a scale from 100 (early) to 1,500 (late)[40,41]. Chromatin state was derived from ENCODE segmentation[30], and euchromatin and heterochromatin regions were defined as those regions where the six main ENCODE cell lines shared the same annotation. The correlation between L1 insertion rate per Mb and each genomic feature was evaluated using Spearman's rank. To study the association with multiple predictor variables we used Poisson regression (glm function in R).

ACKNOWLEDGEMENTS

J.M.C.T. is supported by European Research Council (ERC) starting grant (Grant Agreement Number: 716290 SCUBA CANCERS ERC_Stg_2016). This work was supported by the Wellcome Trust grant 09805. B.R.M is supported by a predoctoral fellowship from Xunta de Galicia (Spain). A.L.B. is supported by a predoctoral fellowship from the Spanish Ministry of Economy, Industry and Competitiveness (MINECO). R.B. received funding through the National Institutes of Health (U24CA210978 and R01CA188228). M.G.B received funding through MINECO, AEI, Xunta de Galicia and FEDER (BFU2013-41554-P, BFU2016-78121-P, ED431F 2016/019). N.B. is supported by a My First AIRC grant from the Associazione Italiana Ricerca sul Cancro (n. 17658). K.H.B. is supported by P50GM107632 and R01CA163705. J.D. is a postdoctoral fellow of the Research Foundation – Flanders (FWO) and the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie Grant Agreement No. 703594-DECODE). P.AW.E. is supported by Cancer Research UK. E.A.L. is supported by K01AG051791. I.M. is supported by Cancer Research UK (C57387/A21777). F.M. is supported by A.I.L. (Associazione Italiana Contro le Leucemie-Linfomi e Mieloma ONLUS) and by S.I.E.S. (Società Italiana di Ematologia Sperimentale). Y.S.J. is supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI14C1277). J.O.K. is supported by an ERC Starting Grant. S.M.W. received funding through a SNSF Early Postdoc Mobility fellowship (P2ELP3_155365) and an EMBO Long-Term Fellowship (ALTF 755-2014). J.W. received funding from the Danish Medical Research Council (DFF-4183-00233).

REFERENCES

1. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. Kazazian, H.H., Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626-32 (2004).
3. Sassaman, D.M. *et al.* Many human L1 elements are capable of retrotransposition. *Nat Genet* **16**, 37-43 (1997).
4. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100**, 5280-5 (2003).
5. Beck, C.R. *et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159-70 (2010).
6. Menendez, L., Benigno, B.B. & McDonald, J.F. L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas. *Mol Cancer* **3**, 12 (2004).
7. Tubio, J.M. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
8. Moran, J.V., DeBerardinis, R.J. & Kazazian, H.H., Jr. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530-4 (1999).
9. Kazazian, H.H., Jr. Processed pseudogene insertions in somatic cells. *Mob DNA* **5**, 20 (2014).
10. Cooke, S.L. *et al.* Processed pseudogenes acquired somatically during cancer development. *Nat Commun* **5**, 3644 (2014).
11. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967-71 (2012).
12. Helman, E. *et al.* Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res* **24**, 1053-63 (2014).
13. Solyom, S. *et al.* Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res* **22**, 2328-38 (2012).
14. Burns, K.H. Transposable elements in cancer. *Nat Rev Cancer* **17**, 415-424 (2017).
15. Gilbert, N., Lutz-Prigge, S. & Moran, J.V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**, 315-25 (2002).
16. Han, K. *et al.* Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res* **33**, 4040-52 (2005).
17. Sen, S.K., Huang, C.T., Han, K. & Batzer, M.A. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res* **35**, 3741-51 (2007).
18. Farkash, E.A., Kao, G.D., Horman, S.R. & Prak, E.T. Gamma radiation increases endonuclease-dependent L1 retrotransposition in a cultured cell assay. *Nucleic Acids Res* **34**, 1196-204 (2006).
19. Li, Y. *et al.* Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature* **508**, 98-102 (2014).
20. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
21. Garsed, D.W. *et al.* The architecture and evolution of cancer neochromosomes. *Cancer Cell* **26**, 653-67 (2014).
22. Zhou, C., Li, J. & Li, Q. CDKN2A methylation in esophageal cancer: a meta-analysis. *Oncotarget* (2017).
23. Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-25 (2012).
24. Cancer Genome Atlas, N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576-82 (2015).
25. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-50 (2014).
26. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719-24 (2009).
27. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905 (2010).
28. Campbell, P.J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109-13 (2010).
29. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-7 (2012).

30. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
31. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
32. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
33. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-74 (2012).
34. Futreal, P.A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83 (2004).
35. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
36. 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).
37. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-92 (2013).
38. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
39. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-78 (2012).
40. Haradhvala, N.J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538-49 (2016).
41. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).