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], Gianmarco Contino[9], Stefan C. Dentro[10,4,11], Alicia L. Bruzos[1,2], Ana Dueso-Barroso[12,13], Daniel Ardeljan[14], Marta Tojo[1,2], Nicola D. Roberts[8], Miguel G. Blanco[15,16], Paul A. W. Edwards[17,18], Joachim Weischenfeldt[19,20], Martin Santamarina[1,2], Montserrat Puiggros[12], Zechen Chong[21], Ken Chen[21], Eunjung Alice Lee[22], Jeremiah A. Wala[23,24], Keiran Raine[8], Adam Butler[8], Sebastian M. Waszak[20], Fabio C. P. Navarro[25,26], Steven E. Schumacher[23,24], Jean Monlong[27], Francesco Maura[28,29,8], Niccolo Bolli[28,29], Guillaume Bourque[27], Mark Gerstein[25,26], Peter J. Park[22], Rameen Berroukhim[23,24], David Torrents[12,30], Jan O. Korbel[20], Inigo Martincorena[8], Rebecca C. Fitzgerald[9], Peter Van Loo[4,5,8,9], Haig H. Kazazian[14], Kathleen H. Burns[31,14], Peter J. Campbell[8,32,*] & Jose M. C. Tubio[1,2,8,*], 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, 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, 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]Medical Research Council (MRC) Cancer Unit, University of Cambridge, Cambridge, UK
[10]Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA. UK
[11]Big Data Institute, University of Oxford, Oxford, UK
[12]Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain
[13]Faculty of Science and Technology. University of Vic - Central University of Catalonia (UVic-UCC), Vic, Spain
[14]Institute for Genetic Medicine, Johns Hopkins University School of Medicine - Baltimore, MD USA
[15]Departamento de Bioquímica e Bioloxía Molecular, CIMUS, Universidade de Santiago de Compostela, 15706 Santiago de Compostela, Spain
[16]Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Universidade de Santiago de Compostela, 15706 Santiago de Compostela, Spain
[17]Department of Pathology, University of Cambridge, Cambridge, UK
[18]Cancer Research UK Cambridge Institute, Cambridge, UK
[19]Biotech Research & Innovation Centre (BRIC); Finsen Laboratory, Rigshospitalet, Copenhagen, Denmark
[20]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany
[21]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[22]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
[23]The Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[24]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[25]Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT
[26]Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT
[27]Department of Human Genetics, McGill University, Montreal, H3A 1B1, Canada
[28]Department of Oncology and Onco-Hematology, University of Milan, Milan, Itlay
[29]Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[30]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[31]Department of Pathology, Johns Hopkins University School of Medicine - Baltimore, MD USA
[32]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: jmctubio@uvigo.es

Dr Peter J. 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 37 histological cancer subtypes. We identified 20,230 somatically acquired retrotransposition events, affecting 43% of samples, and spanning a range of event types. L1 insertions emerged as the third most frequent type of somatic structural variation in cancer. Aberrant L1 integrations can delete megabase-scale regions of a chromosome, sometimes removing tumour suppressor genes, as well as inducing complex translocations and large-scale duplications. Somatic retrotranspositions can also initiate breakage-fusion-bridge cycles of genomic instability, leading to high-level amplification of oncogenes. 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 reference genome are truncated, inactive elements not able to retrotranspose. A small subset of them, maybe 100-150 L1 loci, remain active in the average human genome, acting as source elements, of which a small number are highly active copies termed hot-L1s[3-5]. These L1 source elements are usually transcriptionally repressed, 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 can also mobilize unique DNA sequences located immediately downstream of the source element, a process called 3’ transduction[7-9]. L1 retrotransposons can also promote the somatic trans-mobilization of Alu, SVA and processed pseudogenes, which are copies of messenger RNAs that have been reverse transcribed into DNA and inserted into the genome using the machinery of active L1 elements[10-12].

Approximately 50% of human tumours have somatic retrotransposition of L1 elements[7,13-15]. 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[7,13]. Nonetheless, L1 insertions are capable of promoting other types of genomic structural alterations in the germline and somatically, apart from canonical L1 insertion events[16-18], which remain largely unexplored in human cancer[19,20].

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 cancers mediated by L1 retrotransposition. We find that aberrant integration of L1 retrotransposons has a relevant role in remodeling cancer genome architecture in some human tumours, 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 whole-genomes dataset

We identified 20,230 somatically acquired retrotransposition events. Overall, 43% of all cancer genomes have at least one retrotransposition, most frequently in lung squamous carcinoma (Lung-SCC), esophageal adenocarcinoma (Eso-AdenoCA), and colorectal adenocarcinoma (ColoRect-AdenoCA), where >90% of samples bear one or more somatic events (**Fig. 1a** and **Supplementary Table 1**). Some patients' tumours carried more than 100 separate somatic retrotranspositions: 29% of Eso-AdenoCA patients, 11% of head-and-neck squamous (Head-SCC), 6% of Lung-SCC, and 3% of ColoRect-AdenoCA. These four tumour types alone account for 68% of all somatic events in the PCAWG dataset, while representing just 10% of the samples.

Retrotranspositions were classified into seven categories (**Fig. 1b** and **Supplementary Fig. 1**). With 97% (19,705/20,230) of the events, L1 integrations (16,069 solo-L1 and 3,636 L1-transductions) overwhelmingly dominate the landscape of somatic retrotransposition across all tumours, while Alu and SVA, with 169 and 32 elements respectively, represent minor categories, and no endogenous retrovirus (ERV) somatic events were found. This makes L1 insertions the third most frequent type of somatic structural variation in the PCAWG cohort after deletions

(54,097 events) and tandem duplications (45,539). L1 retrotranspositions are thus considerably more common than unbalanced translocations (6,381) [Y. L. et al., manuscript in preparation]. Trans-mobilization of processed pseudogenes and genomic rearrangements (mainly deletions) promoted by L1 integration, with 228 and 96 events respectively, are rare 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 L1 events predominate in the majority of cancer genomes, we do observe cases where pseudogenes and L1-mediated chromosomal rearrangements have a particularly strong representation among retrotransposition events (**Fig. 1c**).

The genome-wide analysis of the distribution of 19,705 somatic L1 insertions (i.e., solo-L1 and L1 transductions) revealed a dramatic variation of L1 retrotransposition rate across the cancer genome (**Fig 2a** and **Supplementary Table 2**). At a megabase scale, we find that L1 retrotransposition sites density is strongly correlated with late DNA replication timing (Spearman's $\rho = 0.70$, $P \sim 0$) (**Fig. 2b**), and negatively correlated with expression level (Spearman's $\rho = -0.69$, $\mathrm{P} \sim 0$) (**Fig. 2c**) and gene-density (Spearman's $\rho = -0.26$; $P \sim 0$, **Supplementary Fig. 2**). Poisson regression revealed that 48.3% of the total variance in the L1 retrotransposition rate across the genome could be explained by combining these genetic features, with replication timing alone accounting for 46.8% of the variance. We also evaluated the association of L1 retrotransposition density with chromatin state, which revealed a rate of somatic L1 insertion five times higher in repressed than in transcribed chromatin (Exact Poisson test; $P = 9.032\mathrm{e}{-237}$) (**Fig. 2d**).

We identified ~32% (6,317/19,906) somatic retrotranspositions of L1 elements or transductions inserted within gene regions including promoters, of which 198 events hit cancer genes. Reflecting the tendency of L1 elements 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**). The analysis of expression levels in samples with available transcriptome revealed four 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. 3a-c**). This includes one head-and-neck tumour, D015591, with a somatic L1 element integrated into the promoter region of the *ABL2* oncogene. The structural analysis of RNA-seq data identified instances in which portions of a somatic retrotransposition within a gene exonize, being incorporated into the transcript sequence of the affected gene, a process that sometimes involved cancer genes (**Supplementary Fig. 3d**). In addition, we analyzed the potential of processed pseudogenes to generate functional consequences in cancer[11]. We found 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. 4**).

L1 source elements' contribution to Pan-Cancer retrotransposition burden

We used somatically mobilized L1-3' transduction events to trace L1 activity to specific source elements[7]. This shows 124 germline L1 loci in the human genome are responsible for most of the genomic variation generated by retrotransposition in PCAWG (**Supplementary Table 3**). Fifty-two of these loci represent novel, previously unreported source elements in human cancer [S. M.

W. et al., manuscript in preparation]. 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 that alone give rise to half of all somatic transductions (**Fig. 3a** and **Supplementary Fig. 5**). This analysis revealed different behaviours 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**). Strombolian source elements represent the calmest type of hot-L1 activity, characterized by the production of small numbers of retrotranspositions in individual tumour samples, but they are often active leading them to contribute significantly to overall retrotransposition in PCAWG. On the contrary, source elements with Plinian hot activity are rarely active across tumours, but when they are active, their eruption is violent, promoting large numbers of retrotranspositions in one or a few 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**). There is a correlation of somatic retrotranspositions with number of active germline L1 source elements among PCAWG samples (**Supplementary Fig. 6**); this is likely one of the factors that explain why Eso-AdenoCA, Lung-SCC, and Head-SCC account for higher retrotransposition rates – in these three tumour types we also observed higher numbers of active germline L1 loci (**Fig. 3c**). Occasionally, somatic L1 integrations that retain their full length may also act as source for subsequent somatic retrotransposition events[7,21], and may reach hot activity rates, leading them to dominate retrotransposition in a given tumour. For example, in a remarkable head-and-neck tumour, DO14343, we identify one somatic L1 integration at 4p16.1 that then triggers 18 transductions from its new site, with the next most active element being a germline L1 locus at 22q12.1 accounting for 15 transductions (**Supplementary Table 3**).

Genomic deletions mediated by somatic L1 retrotransposition

In cancer genomes with high somatic L1 activity rates, we observed that some L1 retrotransposition events followed a distinctive pattern consisting of a single cluster of reads, associated with copy number loss, whose mates unequivocally identify one extreme of a somatic L1 integration with, apparently, no local, 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 at the far end 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, have been observed to occur somatically with engineered L1s in cultured human cells[16,17] and naturally in the brain[18], and are most likely the consequence of an aberrant mechanism of L1 integration. One potential mechanism is that a molecule of L1 cDNA pairs with a distant 3´-overhang from a preexisting double strand DNA break 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[17] (Fig. 4c); although alternative models have been proposed[17,22-24].

We developed specific algorithms to systematically identify L1-mediated deletions, and applied this across all PCAWG tumours. This identified 90 somatic events matching the patterns described above causing deletions of different size, ranging from ~0.5 Kb to 53.4 Mb (Fig. 4d and **Supplementary Table 4**). 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 passage through an RNA intermediate. No target site duplication is found, which is also the typical pattern for L1-mediated deletions[17].

To confirm that these rearrangements are mediated by the integration of a single intervening retrotransposition event, we explored the PCAWG dataset 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. 4e**). 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 fragments of unique DNA sequence located downstream of an active L1 locus, which are mobilized without the companion L1[7,15]. For example, in one oesophageal tumour, DO50362, we find a deletion of 2.5 Kb on chromosome 3 mediated by the orphan transduction of sequence downstream of an L1 locus on chromosome 7 (**Fig. 4f**).

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 (3 Kb and 10 Kb) that would exceed the insertion event at the deletion boundaries. In these samples, our algorithms confirmed 16 events with the hallmarks of L1-mediated deletions, in which the mate-pair data confirmed a single L1-derived (i.e., solo-L1 or L1-transduction) retrotransposition as the cause of the copy number loss, and identified the sizes of the deletion and the associated insertion (**Supplementary Fig. 7**).

We successfully reconstructed the L1 3'-extreme insertion breakpoint sequence for 91% (82/90) of the retrotransposition events associated with PCAWG L1-mediated deletions, revealing that 88% (72/82) of the L1 events causing deletions preferentially inserted into sequences that resemble L1-endonuclease consensus cleavage sites (e.g., 5'-TTTT/A-3' and related sequences[25]) (**Supplementary Table 4)**. 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[25]. Interestingly, in 17% (14/82) of the events the endonucleotidic cleavage occurred at the phosphodiester bond between a T/G instead of at the standard T/A. In addition, we observed 12% (10/82) instances where the endonuclease motif is not found, suggesting that a small fraction of L1-associated deletions may be the consequence of an L1-endonuclease-independent insertion mechanism[23-25]. Whatever mechanism of L1 integration operating here, taken together, these data indicate that the somatic integration of L1 elements induces the associated deletions.

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

Most L1-mediated deletions ranged from a few hundred to thousands of base pairs, but occasionally deleted megabase regions of a chromosome (**Fig. 4d** and **Supplementary Table 4**). For example, in oesophageal 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 mediated by 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. This unequivocally supports the model that the observed copy number change is indeed a deletion mediated by retrotransposition of an L1 element. Similarly, in a lung tumour, DO27334, we found an interstitial L1-mediated deletion with loss of 51.1 Mb from chromosome X including the centromere (**Fig. 5b**).

L1-mediated deletions were, on occasion, driver events, causing loss of tumour suppressor genes. In oesophageal tumour DO50362, the integration of an L1-transduction from chromosome 7p12.3 into the short arm of chromosome 9 caused a 5.3 Mb clonal deletion involving the 9p21.3-9p21.2 region. This led to loss of one copy of a key tumour suppressor gene, *CDKN2A* (**Fig. 5c**), deleted in many cancer types including oesophageal tumours[26-29]. In this case, the inserted L1 element retained its original structure, meaning that it could have remained active[21]. Interestingly, the sequencing data revealed a somatic transduction arising from this L1 element at its new insertion site, demonstrating that L1 events that promote deletions can be competent for retrotransposition (**Supplementary Fig. 8)**. In a second oesophageal tumour, DO50383, an L1 element integrated into chromosome 9 promotes an 8.6 Mb clonal deletion encompassing the

9p22.1-9p21.1 region that removes one copy of the same tumour suppressor gene, *CDKN2A* (**Fig. 5d**). Thus, L1-mediated deletions have clear oncogenic potential.

L1 retrotransposition generates other types of structural variation in human tumours

Somatic retrotransposition can also be involved in mediating or repairing more complex structural variants. In one oesophageal tumour, DO50365, two separate L1-mediated structural variants were present within a complex cluster of rearrangements (**Fig. 6a**). In the first, an L1 transduction from a source element on chromosome 14q23.1 bridged an unbalanced translocation from chromosome 1p to 5q. A second somatic retrotransposition event bridged from chromosome 5p to an unknown part of the genome, completing a large interstitial copy number loss on chromosome 5 that involves the centromere. This case suggests that retrotransposon transcripts and their reverse transcriptase machinery can mediate breakage and repair of complex dsDNA breaks, spanning two chromosomes.

To explore this further, we identified single-L1 clusters with no reciprocal cluster in the cancer cell-lines that were sequenced using mate-pairs with 3kb and 10kb inserts. Such events may correspond to hidden genomic translocations, linking two different chromosomes, in which L1 retrotransposition is involved. 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. 6b**). This interchromosomal rearrangement is 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 break generated in a second chromosome[25] (**Fig. 6c**).

We also found evidence that L1 integrations can cause duplications of large genomic regions in human cancer. In the oesophageal tumour DO50374 (**Fig. 7a**), 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 by bridging sister chromatids during or after DNA replication (**Fig. 7b**). 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 (**Fig. 7a**). Interestingly, this duplication increases the copy number of the cyclin C gene, *CCNC*, which is dysregulated in some tumours[30].

L1-mediated rearrangements can induce breakage-fusion-bridge cycles that trigger oncogene amplication

L1 retrotranspositions can also induce genomic instability through triggering breakage-fusion-bridge cycles. This form of genetic instability starts with end-to-end fusion of broken sister chromatids, leading to a dicentric chromosome that forms an anaphase bridge during mitosis. Classically, the end-to-end chromosome fusions are thought to arise from telomere attrition[31-33] We found, however, that somatic retrotransposition can induce that first inverted rearrangement generating end-to-end fusion of sister chromatids. In lung tumour DO27334 (**Fig. 7c**), we found a small L1 event inserted on chromosome 14q demarcating a copy number change that involves an 79.6 Mb amplification of the 14q. The analysis of the sequencing data at the breakpoint revealed two discordant read-clusters with the same orientation, which are 5.5 Kb apart and support the integration of a L1. Both discordant clusters demarcate an increment of the sequencing coverage, where density is much greater on the right cluster. The only genomic structure that can explain this pattern is a fold-back inversion in which the two sister chromatids are bridged by a L1 retrotransposition in head-to-head (inverted) orientation (**Fig. 7d**).

In the previous example (**Fig. 7c-d**), no further breaks occurred, and the L1 retrotransposition generated an isochromosome (14q). Beyond this, we found examples in which the fusion of two chromatids by a L1 bridge induced further cycles of breakage-fusion-bridge repair. In the oesophageal tumour DO50374, we identified a cluster of reads at the long arm of chromosome 11 with the typical hallmarks of an L1-mediated rearrangement (**Fig. 8a**). Copy number data analysis showed that this L1 insertion point demarcated a 53 Mb deletion, involving telomeric region loss, from a region of massive amplification on chromosome 11. The amplified region of chromosome 11 contains the *CCND1* oncogene, which is amplified in many human cancers[34]. The other end of this amplification was bound by a conventional fold-back inversion rearrangement (**Fig. 8a**), indicative of breakage-fusion-bridge repair[35,36].

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 to the clone during the subsequent cell division ('fold-back inversion model', **Fig. 8b**). The chromatids bridged by the L1 insertion now make 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 breakage-fusion-bridge repair, albeit not one mediated by L1 retrotransposition. These cycles lead to rapid-fire amplification of the *CCND1* oncogene. Alternatively, an interchromosomal rearrangement mediated by L1 retrotransposition ('interchromosomal rearrangement model', **Fig. 8b**) followed by two cycles of breakage-fusion-bridge repair could generate similar copy number patterns with telomere loss and amplification of *CCND1*.

Our data show that L1-mediated retrotransposition is an alternative mechanism of creating the first dicentric chromosome that seeds subsequent rounds of chromosomal breakage and repair. If this occurs near an oncogene, the resulting amplification can provide a powerful selective advantage to the clone. We looked in the PCAWG dataset for other rearrangements demarcating copy number amplification from telomeric deletions that were mediated by L1 integration. We found four more such events across three cancer samples (**Supplementary Fig. 9**). In a lung tumour, DO26976, we found almost identical rearrangements to the one described above (**Fig. 8c**). Here, a somatic L1 event also generated telomere loss that seeded a second cycle of breakage-fusion-bridge repair. The megabase-size amplification of chromosomal regions also targeted the *CCND1* oncogene, with boundaries demarcated by the L1 insertion breakpoint and a fold-back inversion indicating breakage-fusion-bridge repair. The independent occurrence of these patterns, involving the amplification of *CCND1*, in two different tumour samples, DO50374 and DO26976, demonstrates a mutational mechanism mediated by L1 retrotransposition, which likely contributes to the initiation and/or development of human cancer.

DISCUSSION

Here we characterize the patterns and mechanisms 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 genomes can sometimes emerge from aberrant L1 retrotransposition events in tumours with high retrotransposition rates, particularly in oesophageal, 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. We believe that structural variants induced by somatic retrotransposition in human cancer may be much more frequent than we could unambiguously characterise here, due to a limitation of the insert size of the paired-end sequencing libraries. Long-read sequencing technologies should be able to give a more complete picture of how frequent such events are.

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 for patterns of L1 retrotransposition.

FIGURE LEGENDS

Figure 1. Rates of somatic retrotransposition across human cancers. (A) For each cancer type included in the PCAWG project, it is represented the proportion of tumour samples with more than 100 somatic retrotranspositions (red), between 10 and 100 (orange), between 1 and 10 (yellow), and zero (grey). The number of samples analyzed in shown on the right side of the panel. (B) Frequency of retrotransposition events across cancers. Somatic retrotranspositions in PCAWG were classified into 7 categories (ERVs are not represented, as they accounted for zero copies). Only the four cancer types with higher retrotransposition rates are shown. 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 'spider web genome' pattern that characterizes 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. L1 retrotransposition density distribution across the cancer genome and association with genome organization features. (A) L1 retrotransposition density (grey bars) across the cancer genome is represented together with 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) The rate of somatic L1 insertions strongly correlates with replication timing. The plot represents the Kernel Density Estimate (RDE). (C) The 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[37]. (D) L1 retrotranspositions acquired somatically are overrepresented in transcriptionally repressed (typically heterochromatic) regions of the cancer genome. Error bars reflect Poisson confident intervals. Chromatin states were derived from ENCODE[38]. (E) Somatic retrotransposition is enriched in lowly expressed genes (<3 FPKM) relative to highly expressed genes. Here, expression data were PCAWG 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 PCAWG 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. For an

extended version of this figure see **Supplementary Fig. 5.** (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 esophageal tumour with high L1 somatic activity rates, we found one-single cluster of reads on 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 local reciprocal cluster supporting the other extreme of the L1 insertion. (B) Analysis of the associated copy number change on 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 deletion occurring in conjunction with the integration of a 2.1 Kb L1 somatic insertion. The sequencing data associated with this L1-mediated deletion shows two clusters of discordant read-pairs and clipped-reads supporting both extremes of a L1 retrotransposon. (C) Model of L1-mediated deletion. 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. (D) Distribution of the sizes of 90 L1-mediated deletions identified in the PCAWG dataset. (E) 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 so short, we also identify discordant

read-pairs that span the L1 event and support the deletion. (F) In an esophageal adenocarcinoma, DO50362, the integration in 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.

Figure 5. Somatic integration of L1 causes loss of megabase interstitial chromosomal regions in cancer. (A) In an esophageal tumour, DO50410, we find a 45.5 Mb interstitial deletion on 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 read-pairs that span the L1 event and support the deletion. L1-endonuclease 5'-TTTT/A-3' motif identifies TPRT L1-integration mechanism. (B) In an esophageal tumour, D027334, 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 cluster of reads 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. The L1-endonuclease 5'-TTTT/A-3' motif was identfied. (C) L1-mediated deletions promote loss of tumour suppressor genes. 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 cluster that contain split-reads matching a poly-A and whose mates map onto a region transduced from chromosome 7. (D) 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 L1 event integration, together with clipped-reads that precisely mark the insertion breakpoint to base pair resolution.

Figure 6. Somatic L1 integration promotes translocations in human cancers. (A) In esophageal adenocarcinoma DO50365, two separate L1 events mediate interchromosomal rearragements. In the first, an L1 transduction from a source element on chromosome 14q23.1 bridged an unbalanced translocation from chromosome 1p to 5q. A second somatic retrotransposition event bridged from chromosome 5p to an unknown part of the genome, completing a 47.9 Mb interstitial copy number loss on chromosome 5 that removes the centromere. (B) 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. 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. (C) 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 on a different chromosome, leading to translocation.

Figure 7. Somatic L1 integration promotes duplications of megabase regions in human cancers. (A) In an esophageal tumour, DO50374, we find a 22.6 Mb tandem duplication at the long arm of chromosome 6. 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 shorter 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 5'-TTT/A-3' degenerate motif identifies TPRT L1-integration mechanism. (B) 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 to the initial integration site locus. (C) In lung tumour DO27334, a small L1 insertion causes a 79.6 Mb duplicatiob of the 14q through the induction of a fold-back inversion rearrangement. The analysis of the sequencing data at the breakpoint revealed two clusters of discordant read-pairs with the same orientation, which are 5.5 Kb apart and support the integration of an L1. Sequencing density is much greater on one of the clusters (right cluster), which is another hallmark of fold-back inversions. (D) L1-mediated fold-back inversion model.

Fig. 8. Somatic integration of L1 can trigger breakage-fusion-bridge cycles that lead to oncogene amplification. (A) In an esophageal adenocarcinoma, DO50374, the integration of an L1 retrotransposon on the long arm of chromosome 11 promotes the loss of 53 Mb region that includes the telomere, and activates breakage-fusion-bridge repair leading to amplification of the *CCND1* oncogene. The presence of an L1-endonuclease cleavage site motif 5'-TTT/A-3' together with a single cluster of discordant reads support the integration of an L1 event. This somatic retrotransposition demarcates a 53 Mb deleted region, involving loss of the telomere, from a region of massive amplification that involves *CCND1*. The presence of two reciprocal clusters of discordant read-pairs support a fold-back inversion, a diagnostic pattern associated

with breakage-fusion-bridge repair. The patterns described suggest two independent breakage-fusion-bridge cycles, marked with (1) and (2). (B) Models for the patterns described in A. The fold-back inversion model involves two breakage-fusion-bridge cycles, one induced by L1-mediated fold-back inversion (see Fig. 7d), and a second is by standard breakage-fusion-bridge repair. The intechromosomal rearrangement model involves an interchromosomal rearrangement mediated by an L1, followed by one extra cycle of breakage-fusion-bridge repair. (C) In a lung cancer, DO26976, the integration of an 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. Distribution of L1 retrotransposition density and genome organization features in PCAWG genomes

Supplementary Table 3. List of germline L1 source elements with counts per sample

Supplementary Table 4. Features of PCAWG L1-mediated rearrangements 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. Somatic L1 retrotransposition density anticorrelates with gene-density.

Supplementary Figure 3. 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 an 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. (D) We found 6 instances where bits of somatic retrotranspositions exonize, being incorporated into the host gene transcripts. These include the cancer genes *PTPN11* and *NCOR2*.

Supplementary Figure 4. 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 5. Contribution of 124 germline L1 source loci to somatic retrotransposition burden in PCAWG. The total number of transductions identified for each histological cancer subtype is shown in a blue coloured scale. The total number of transductions identified from source element locus is shown in a green 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.

Supplementary Figure 6. Correlation of number of somatic L1 insertions with number of germline L1 source elements. Points represent tumour samples.

Supplementary Figure 7. Validation of L1-mediated deletions by mate-pair sequencing data analysis. In order to further validate L1-mediated deletions, we performed mate-pair sequencing of long-inserts libraries (3 Kb and 10 Kb) on two cancer cell-lines with high-retrotransposition rates. Here, it is shown the 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 8. 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 an L1-event with full-length structure. Five kilobases downstream, a positive cluster of reads supports a transduction from this L1-retotransposition event into chromosome X.

Supplementary Figure 9. Somatic integration of L1 causes 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 an 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 an 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. Only the two clusters supporting both extremes of a putative L1-mediated fold-back inversion are shown.

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[39]. Finally, annotation of MEIs was performed using the software ANNOVAR[40], gencode v19 annotation[41], and the Cancer Gene Census database[42].

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 across cancer types. 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 PCAWG RNA-seq data to identify and characterize the transcriptional consequences of somatic processed pseudogene integrations. We interrogated RNA-seq data (split-reads and discordant read-pairs) looking for chimeric retrocopies involving processed pseudogenes and target genomic region. For each processed pseudogene insertion somatic call, we extracted all the RNA-seq 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[43] searches against a database containing all isoforms of the source gene described in RefSeq[44], 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 an 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 an 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)[45].

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[46] with 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 in samples with and without somatic L1 insertion. We used FPKM values calculated through Cufflinks software[47]. 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) [37]. DNA replication timing was expressed on a scale from 100 (early) to 1,500 (late)[48,49]. Chromatin state was derived from ENCODE segmentation[38], 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). P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. The Francis Crick Institute which receives its core funding from Cancer Research UK (FC001202), the UK Medical Research Council (FC001202), and the Wellcome Trust (FC001202).

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. 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**, 143-8 (1994).
9. Moran, J.V., DeBerardinis, R.J. & Kazazian, H.H., Jr. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530-4 (1999).
10. Kazazian, H.H., Jr. Processed pseudogene insertions in somatic cells. *Mob DNA* **5**, 20 (2014).
11. Cooke, S.L. *et al.* Processed pseudogenes acquired somatically during cancer development. *Nat Commun* **5**, 3644 (2014).
12. Ewing, A.D. *et al.* Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol* **14**, R22 (2013).
13. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967-71 (2012).
14. Helman, E. *et al.* Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res* **24**, 1053-63 (2014).
15. Solyom, S. *et al.* Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res* **22**, 2328-38 (2012).
16. Symer, D.E. *et al.* Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110**, 327-38 (2002).
17. Gilbert, N., Lutz-Prigge, S. & Moran, J.V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**, 315-25 (2002).
18. Erwin, J.A. *et al.* L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci* **19**, 1583-1591 (2016).
19. Burns, K.H. Transposable elements in cancer. *Nat Rev Cancer* **17**, 415-424 (2017).
20. Kazazian, H.H., Jr. & Moran, J.V. Mobile DNA in Health and Disease. *N Engl J Med* **377**, 361-370 (2017).
21. Kimberland, M.L. *et al.* Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Hum Mol Genet* **8**, 1557-60 (1999).
22. 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).

23. 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).

24. 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).

25. Gilbert, N., Lutz, S., Morrish, T.A. & Moran, J.V. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25**, 7780-95 (2005).

26. Zhou, C., Li, J. & Li, Q. CDKN2A methylation in esophageal cancer: a meta-analysis. *Oncotarget* (2017).

27. Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-25 (2012).

28. Cancer Genome Atlas, N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576-82 (2015).

29. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-50 (2014).

30. Xu, W. & Ji, J.Y. Dysregulation of CDK8 and Cyclin C in tumorigenesis. *J Genet Genomics* **38**, 439-52 (2011).

31. Artandi, S.E. *et al.* Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature* **406**, 641-5 (2000).

32. O'Hagan, R.C. *et al.* Telomere dysfunction provokes regional amplification and deletion in cancer genomes. *Cancer Cell* **2**, 149-55 (2002).

33. Maciejowski, J. & de Lange, T. Telomeres in cancer: tumour suppression and genome instability. *Nat Rev Mol Cell Biol* **18**, 175-186 (2017).

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

35. Campbell, P.J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109-13 (2010).

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

37. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-7 (2012).

38. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

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

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

41. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-74 (2012).

42. Futreal, P.A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83 (2004).

43. 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).

44. 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).

45. 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).

46. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).

47. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-78 (2012).

48. Haradhvala, N.J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538-49 (2016).

49. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).

Figure 1

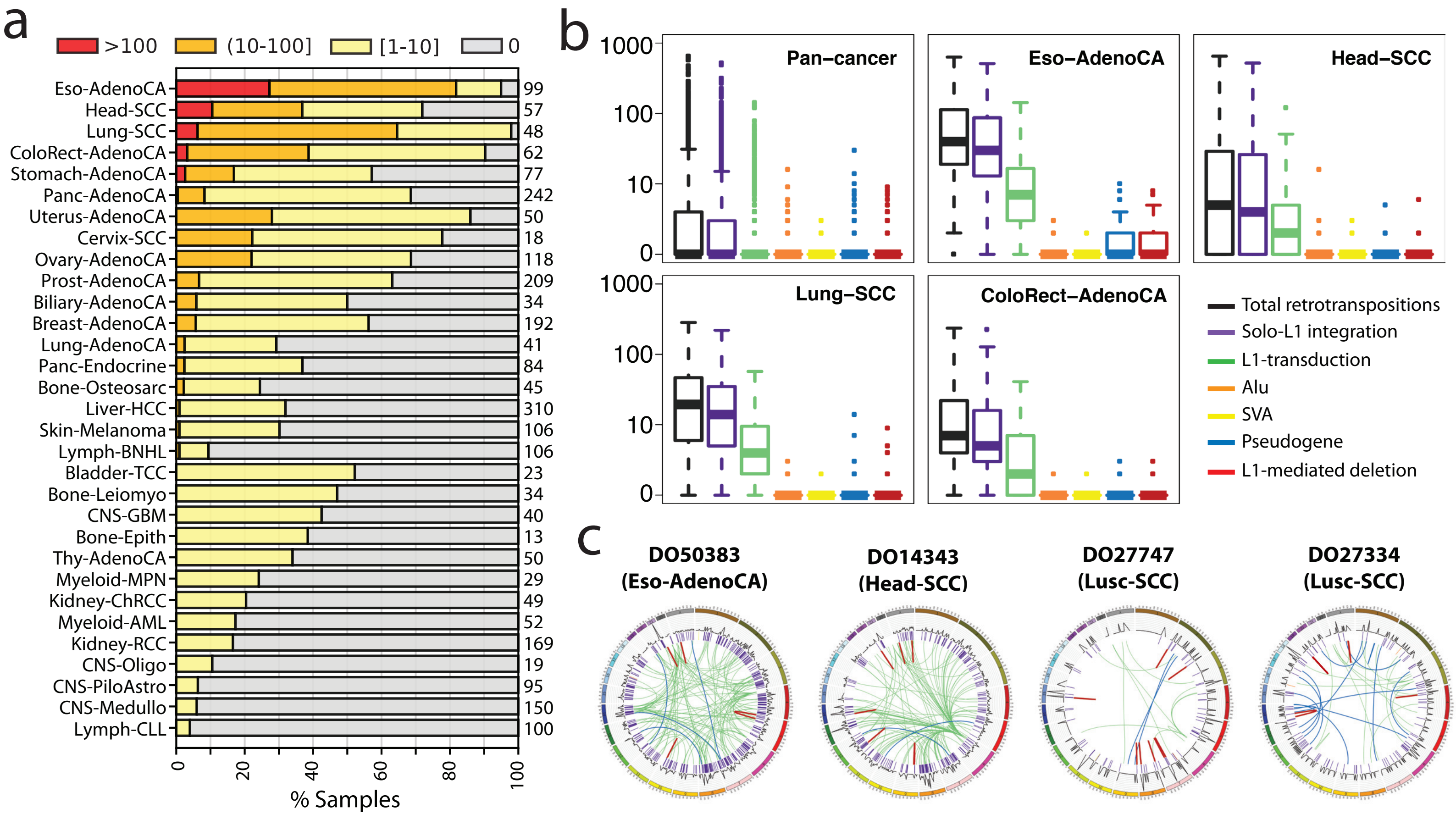

a
>100
(10-100]
[1-10]
0
Eso-AdenoCA 99
Head-SCC 57
Lung-SCC 48
ColoRect-AdenoCA 62
Stomach-AdenoCA 77
Panc-AdenoCA 242
Uterus-AdenoCA 50
Cervix-SCC 18
Ovary-AdenoCA 118
Prost-AdenoCA 209
Biliary-AdenoCA 34
Breast-AdenoCA 192
Lung-AdenoCA 41
Panc-Endocrine 84
Bone-Osteosarc 45
Liver-HCC 310
Skin-Melanoma 106
Lymph-BNHL 106
Bladder-TCC 23
Bone-Leiomyo 34
CNS-GBM 40
Bone-Epith 13
Thy-AdenoCA 50
Myeloid-MPN 29
Kidney-ChRCC 49
Myeloid-AML 52
Kidney-RCC 169
CNS-Oligo 19
CNS-PiloAstro 95
CNS-Medullo 150
Lymph-CLL 100
0
20
40
60
80
100
% Samples
b
1000
100
10
0
Pan–cancer
Eso–AdenoCA
Head–SCC
Lung–SCC
ColoRect–AdenoCA
Total retrotranspositions
Solo-L1 integration
L1-transduction
Alu
SVA
Pseudogene
L1-mediated deletion
c
DO50383 (Eso-AdenoCA)
DO14343 (Head-SCC)
DO27747 (Lusc-SCC)
DO27334 (Lusc-SCC)

Figure 2

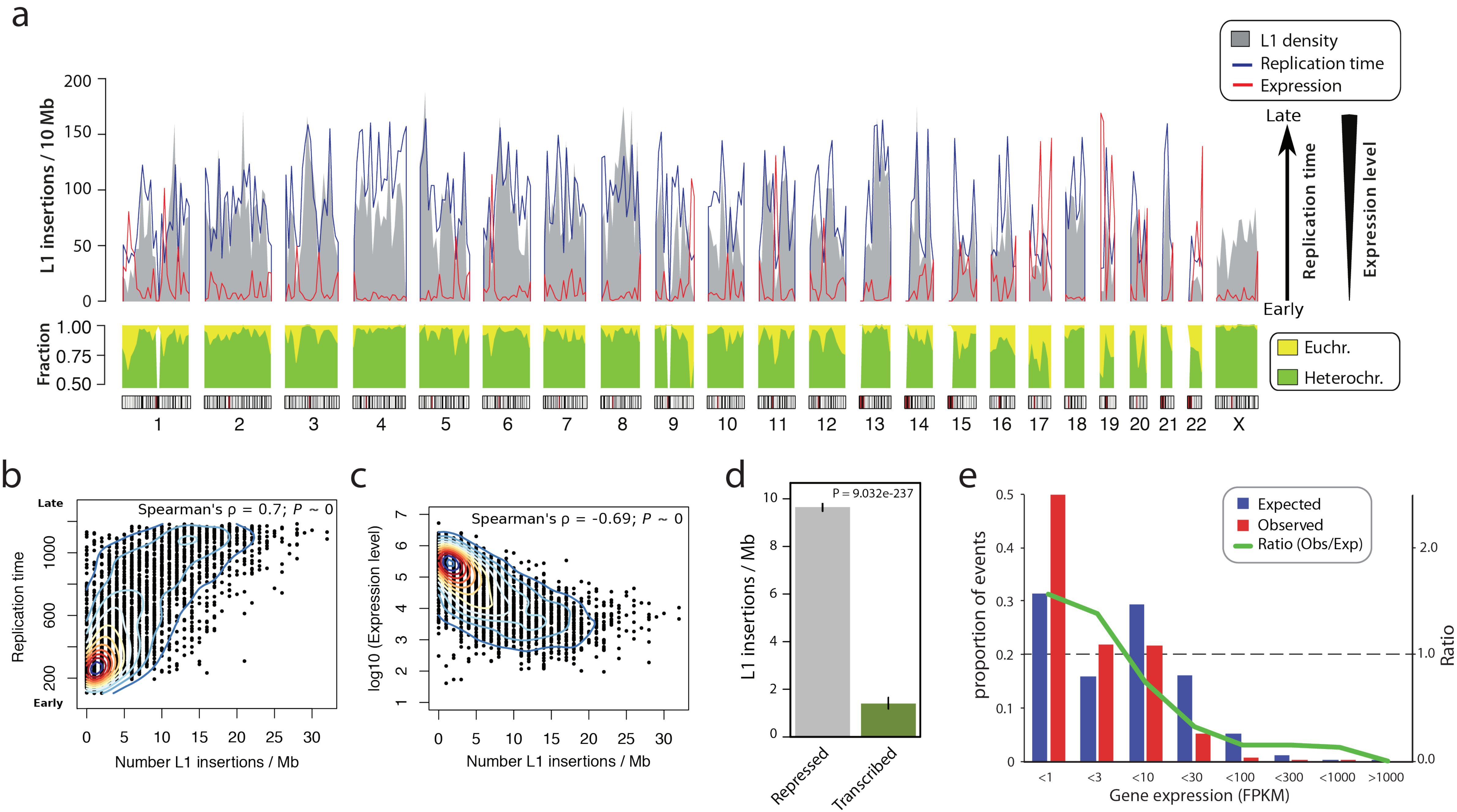
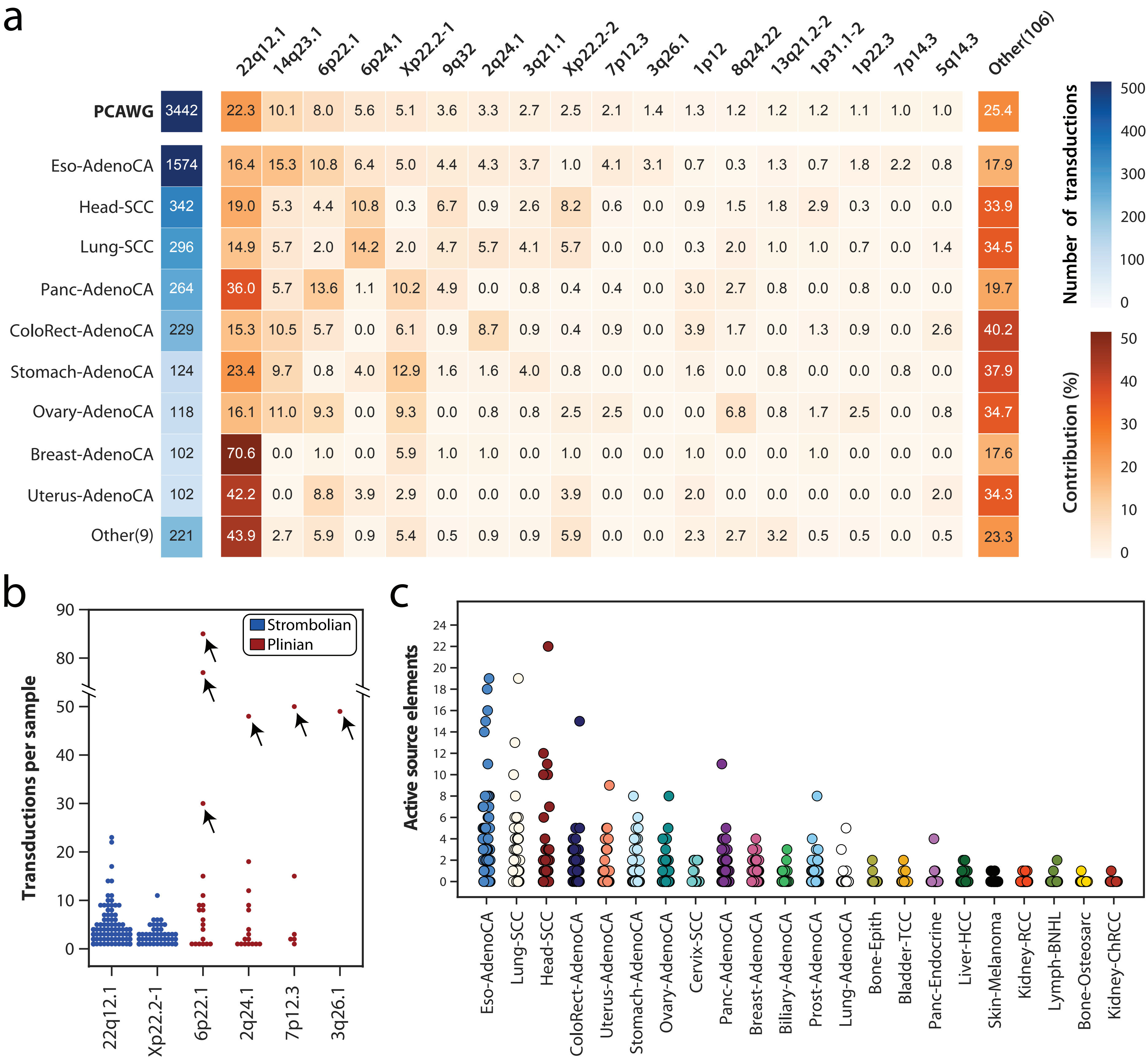

a
L1 density
Replication time
Expression
L1 insertions / 10 Mb
200
150
100
50
0
Late
Early
Replication time
Expression level
Fraction
1.00
0.75
0.50
Euchr.
Heterochr.
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
b
Spearman's ρ = 0.7; P ~ 0
Late
Early
Replication time
1000
600
200
0
5
10
15
20
25
30
Number L1 insertions / Mb
c
Spearman's ρ = -0.69; P ~ 0
log10 (Expression level)
1
2
3
4
5
6
7
0
5
10
15
20
25
30
Number L1 insertions / Mb
d
P = 9.032e-237
L1 insertions / Mb
0
2
4
6
8
10
Repressed
Transcribed
e
Expected
Observed
Ratio (Obs/Exp)
proportion of events
0
0.1
0.2
0.3
0.4
0.5
Ratio
0.0
1.0
2.0
<1
<3
<10
<30
<100
<300
<1000
>1000
Gene expression (FPKM)

Figure 3



		22q12.1	14q23.1	6p22.1	6p24.1	Xp22.2-1	9q32	2q24.1	3q21.1	Xp22.2-2	7p12.3	3q26.1	1p12	8q24.22	13q21.2-2	1p31.1-2	1p22.3	7p14.3	5q14.3	Other(106)
PCAWG	3442	22.3	10.1	8.0	5.6	5.1	3.6	3.3	2.7	2.5	2.1	1.4	1.3	1.2	1.2	1.2	1.1	1.0	1.0	25.4
Eso-AdenoCA	1574	16.4	15.3	10.8	6.4	5.0	4.4	4.3	3.7	1.0	4.1	3.1	0.7	0.3	1.3	0.7	1.8	2.2	0.8	17.9
Head-SCC	342	19.0	5.3	4.4	10.8	0.3	6.7	0.9	2.6	8.2	0.6	0.0	0.9	1.5	1.8	2.9	0.3	0.0	0.0	33.9
Lung-SCC	296	14.9	5.7	2.0	14.2	2.0	4.7	5.7	4.1	5.7	0.0	0.0	0.3	2.0	1.0	1.0	0.7	0.0	1.4	34.5
Panc-AdenoCA	264	36.0	5.7	13.6	1.1	10.2	4.9	0.0	0.8	0.4	0.4	0.0	3.0	2.7	0.8	0.0	0.8	0.0	0.0	19.7
ColoRect-AdenoCA	229	15.3	10.5	5.7	0.0	6.1	0.9	8.7	0.9	0.4	0.9	0.0	3.9	1.7	0.0	1.3	0.9	0.0	2.6	40.2
Stomach-AdenoCA	124	23.4	9.7	0.8	4.0	12.9	1.6	1.6	4.0	0.8	0.0	0.0	1.6	0.0	0.8	0.0	0.0	0.8	0.0	37.9
Ovary-AdenoCA	118	16.1	11.0	9.3	0.0	9.3	0.0	0.8	0.8	2.5	2.5	0.0	0.0	6.8	0.8	1.7	2.5	0.0	0.8	34.7
Breast-AdenoCA	102	70.6	0.0	1.0	0.0	5.9	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	17.6
Uterus-AdenoCA	102	42.2	0.0	8.8	3.9	2.9	0.0	0.0	0.0	3.9	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	2.0	34.3
Other(9)	221	43.9	2.7	5.9	0.9	5.4	0.5	0.9	0.9	5.9	0.0	0.0	2.3	2.7	3.2	0.5	0.5	0.0	0.5	23.3

Figure 4

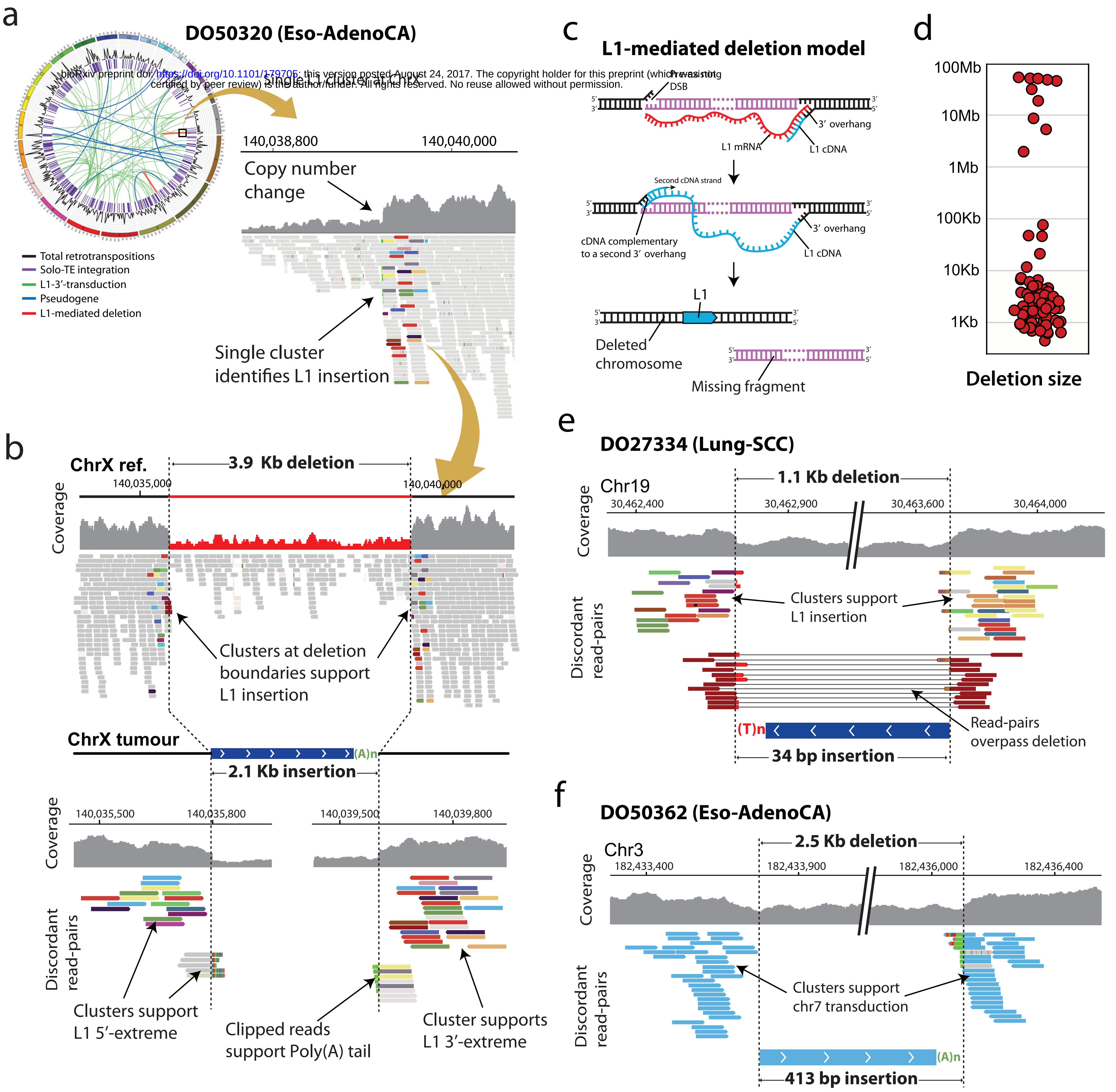

a
DO50320 (Eso-AdenoCA)
Single L1 cluster at ChrX
140,038,800
140,040,000
Copy number change
Single cluster identifies L1 insertion
Total retrotranspositions
Solo-TE integration
L1-3′-transduction
Pseudogene
L1-mediated deletion
b
ChrX ref.
3.9 Kb deletion
140,035,000
140,040,000
Coverage
Clusters at deletion boundaries support L1 insertion
ChrX tumour
(A)n
2.1 Kb insertion
140,035,500
140,035,800
140,039,500
140,039,800
Coverage
Discordant read-pairs
Clusters support L1 5′-extreme
Clipped reads support Poly(A) tail
Cluster supports L1 3′-extreme
c
L1-mediated deletion model
Pre-existing DSB
5′
3′
3′ overhang
L1 mRNA
L1 cDNA
Second cDNA strand
3′ overhang
cDNA complementary to a second 3′ overhang
L1 cDNA
L1
Deleted chromosome
Missing fragment
d
100Mb
10Mb
1Mb
100Kb
10Kb
1Kb
Deletion size
e
DO27334 (Lung-SCC)
1.1 Kb deletion
Chr19
30,462,400
30,462,900
30,463,600
30,464,000
Coverage
Discordant read-pairs
Clusters support L1 insertion
Read-pairs overpass deletion
(T)n
34 bp insertion
f
DO50362 (Eso-AdenoCA)
2.5 Kb deletion
Chr3
182,433,400
182,433,900
182,436,000
182,436,400
Coverage
Discordant read-pairs
Clusters support chr7 transduction
(A)n
413 bp insertion

Figure 5

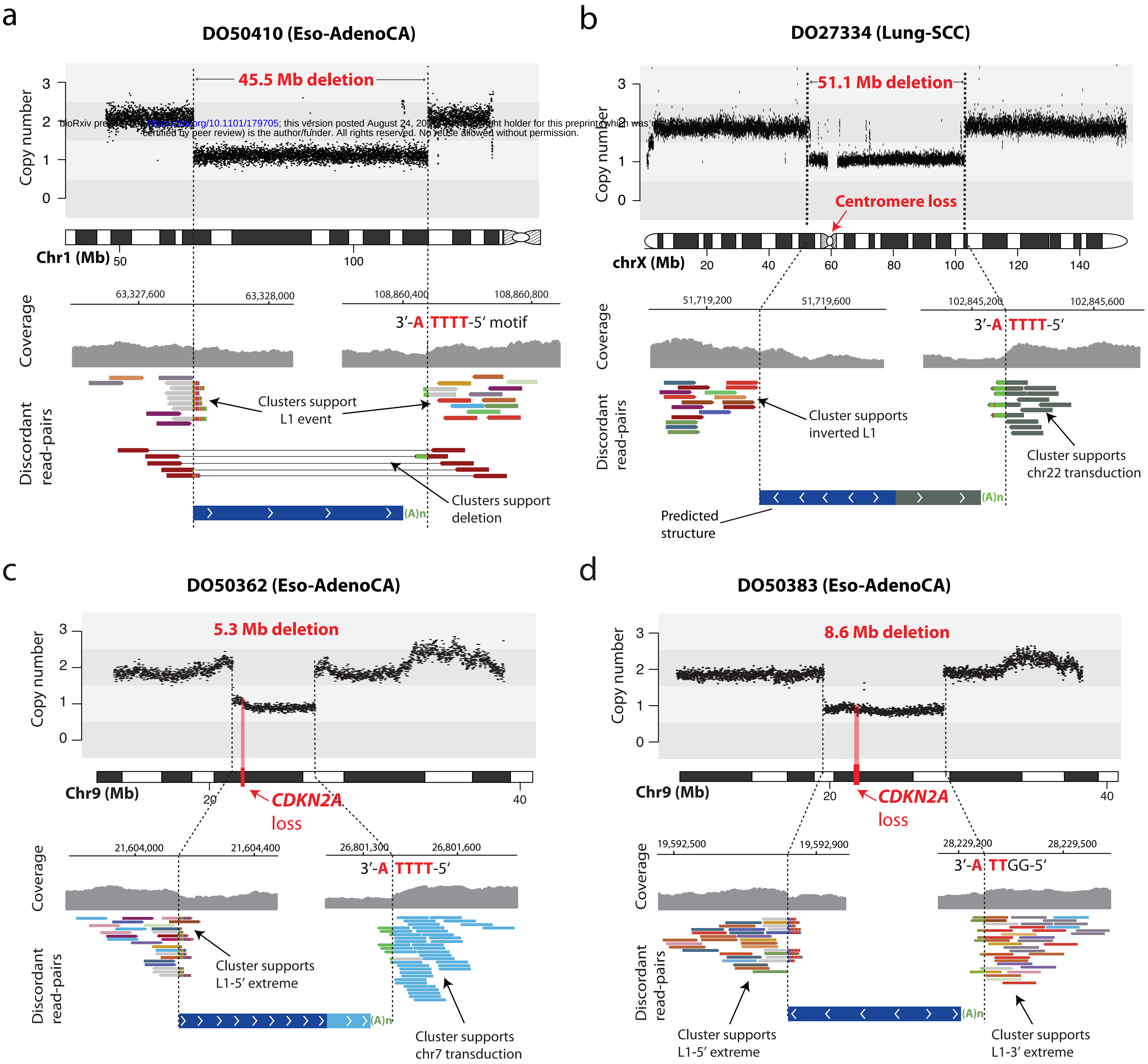

a
DO50410 (Eso-AdenoCA)
45.5 Mb deletion
Copy number
3
2
1
0
Chr1 (Mb)
50
100
63,327,600
63,328,000
108,860,400
108,860,800
Coverage
3′-A TTTT-5′ motif
Discordant read-pairs
Clusters support L1 event
Clusters support deletion
(A)n
b
DO27334 (Lung-SCC)
51.1 Mb deletion
Copy number
3
2
1
0
Centromere loss
chrX (Mb)
20
40
60
80
100
120
140
51,719,200
51,719,600
102,845,200
102,845,600
Coverage
3′-A TTTT-5′
Discordant read-pairs
Cluster supports inverted L1
Cluster supports chr22 transduction
Predicted structure
(A)n
c
DO50362 (Eso-AdenoCA)
5.3 Mb deletion
Copy number
3
2
1
0
Chr9 (Mb)
20
40
CDKN2A loss
21,604,000
21,604,400
26,801,300
26,801,600
Coverage
3′-A TTTT-5′
Discordant read-pairs
Cluster supports L1-5′ extreme
Cluster supports chr7 transduction
(A)n
d
DO50383 (Eso-AdenoCA)
8.6 Mb deletion
Copy number
3
2
1
0
Chr9 (Mb)
20
40
CDKN2A loss
19,592,500
19,592,900
28,229,200
28,229,500
Coverage
3′-A TTGG-5′
Discordant read-pairs
Cluster supports L1-5′ extreme
Cluster supports L1-3′ extreme
(A)n

Figure 6

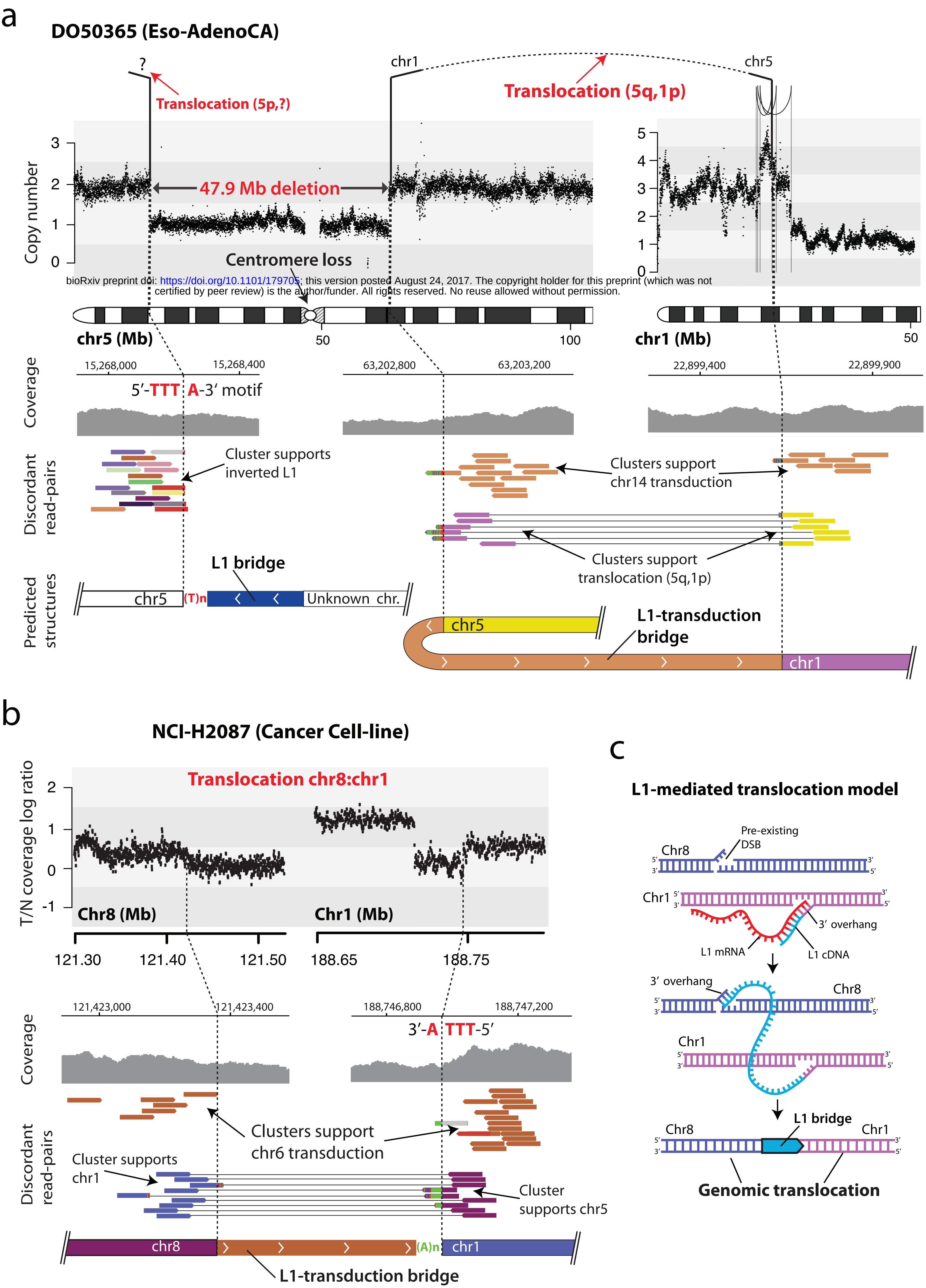
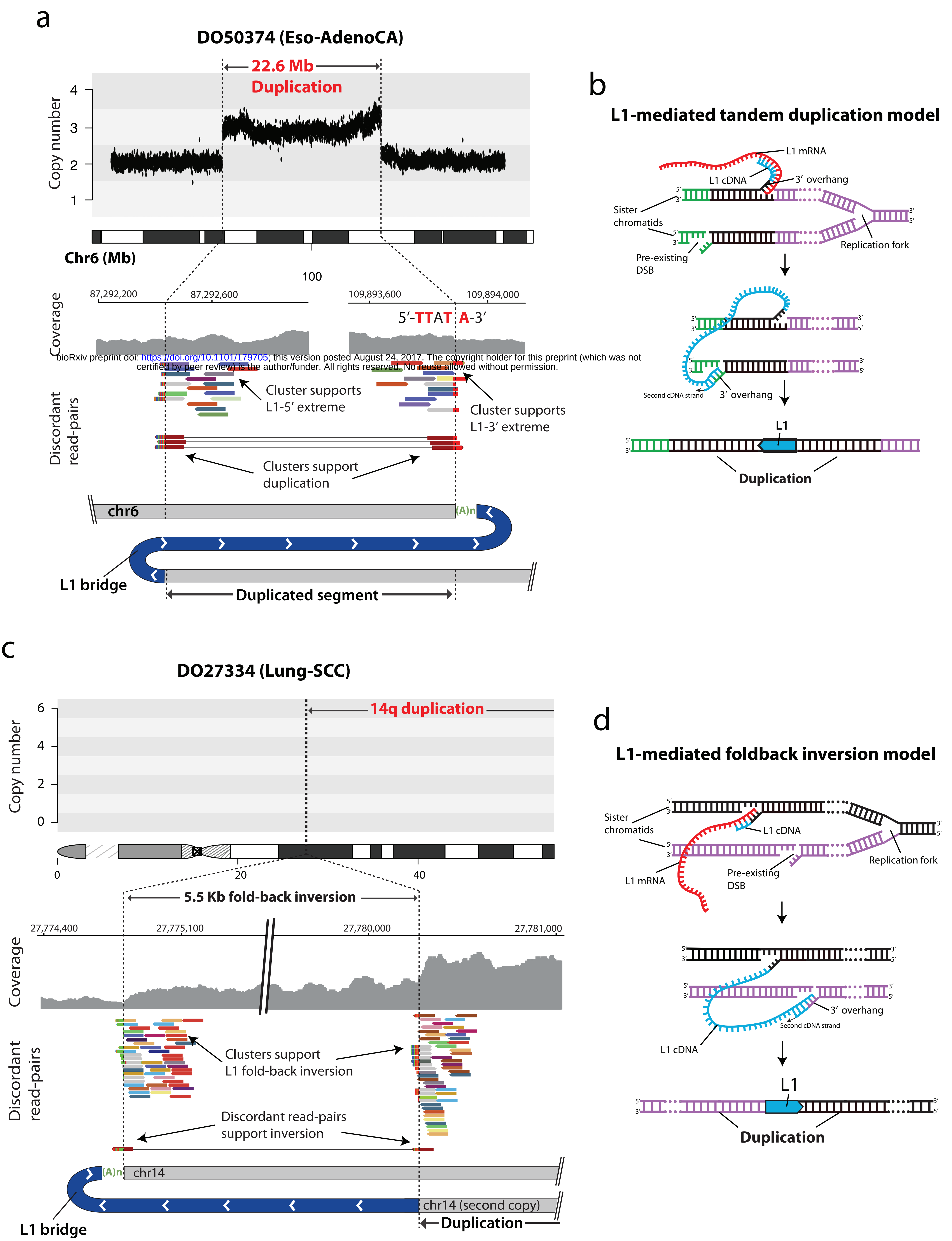
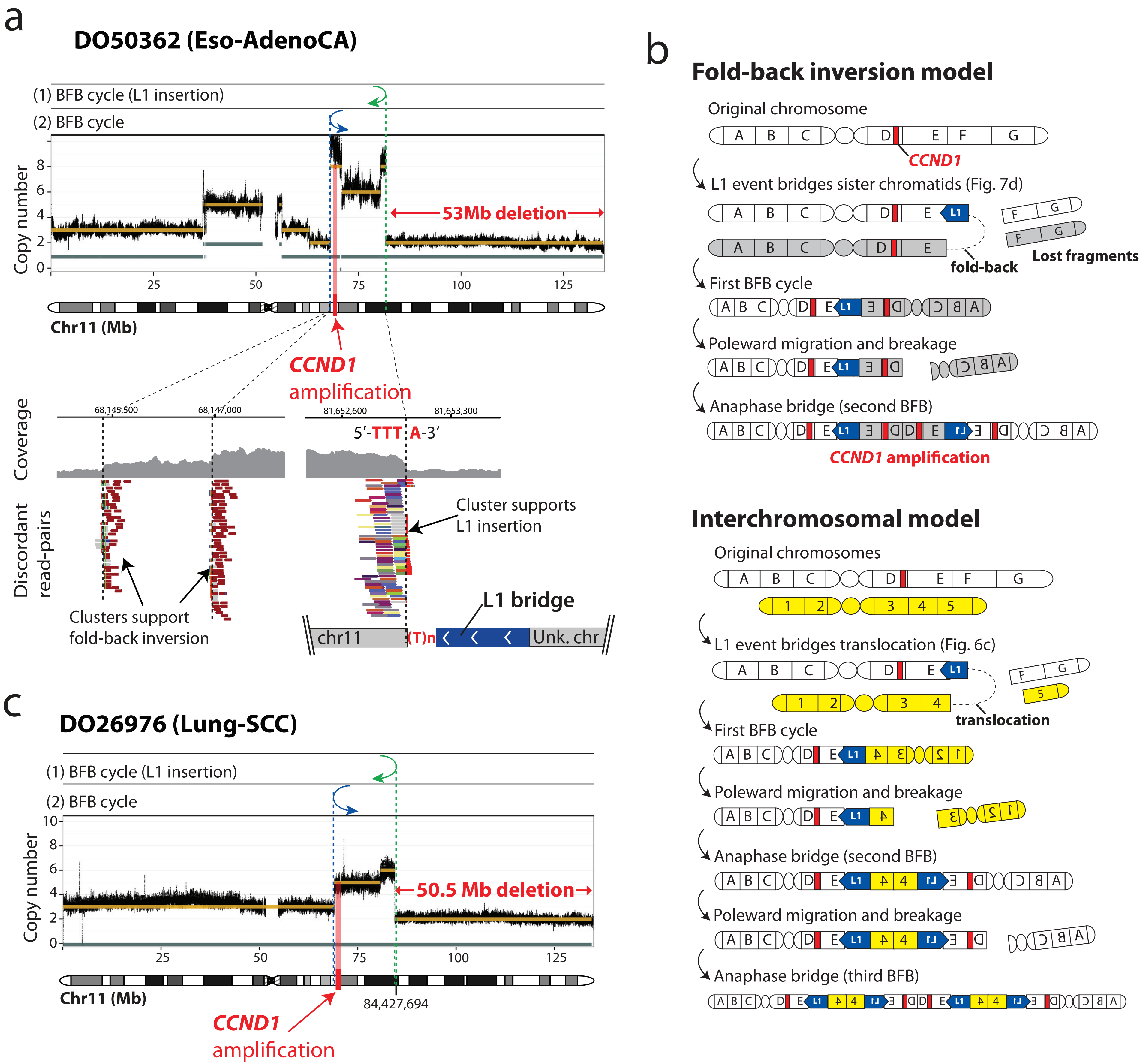

a
DO50365 (Eso-AdenoCA)
?
Translocation (5p,?)
chr1
chr5
Translocation (5q,1p)
Copy number
3
2
1
0
47.9 Mb deletion
Centromere loss
5
4
3
2
1
0
chr5 (Mb)
50
100
chr1 (Mb)
50
Coverage
15,268,000
15,268,400
5′-TTT A-3′ motif
63,202,800
63,203,200
22,899,400
22,899,900
Discordant read-pairs
Cluster supports inverted L1
Clusters support chr14 transduction
Clusters support translocation (5q,1p)
Predicted structures
L1 bridge
chr5
(T)n
Unknown chr.
chr5
L1-transduction bridge
chr1
b
NCI-H2087 (Cancer Cell-line)
Translocation chr8:chr1
T/N coverage log ratio
2
1
0
-1
Chr8 (Mb)
Chr1 (Mb)
121.30
121.40
121.50
188.65
188.75
121,423,000
121,423,400
188,746,800
188,747,200
3′-A TTT-5′
Coverage
Discordant read-pairs
Clusters support chr6 transduction
Cluster supports chr1
Cluster supports chr5
chr8
(A)n
chr1
L1-transduction bridge
C
L1-mediated translocation model
Pre-existing DSB
Chr8
5′
3′
Chr1
3′ overhang
L1 mRNA
L1 cDNA
3′ overhang
Chr8
Chr1
L1 bridge
Chr8
Chr1
Genomic translocation

Figure 7



a
DO50374 (Eso-AdenoCA)
22.6 Mb
Duplication
Copy number
4
3
2
1
Chr6 (Mb)
100
Coverage
87,292,200
87,292,600
109,893,600
109,894,000
5′-TTAT A-3′
Discordant read-pairs
Cluster supports L1-5′ extreme
Cluster supports L1-3′ extreme
Clusters support duplication
chr6
(A)n
L1 bridge
Duplicated segment
b
L1-mediated tandem duplication model
L1 mRNA
L1 cDNA
3′ overhang
Sister chromatids
Replication fork
Pre-existing DSB
Second cDNA strand
3′ overhang
L1
Duplication
c
DO27334 (Lung-SCC)
14q duplication
Copy number
6
4
2
0
0
20
40
5.5 Kb fold-back inversion
27,774,400
27,775,100
27,780,000
27,781,000
Coverage
Discordant read-pairs
Clusters support L1 fold-back inversion
Discordant read-pairs support inversion
(A)n
chr14
chr14 (second copy)
L1 bridge
Duplication
d
L1-mediated foldback inversion model
Sister chromatids
L1 cDNA
Replication fork
L1 mRNA
Pre-existing DSB
3′ overhang
Second cDNA strand
L1 cDNA
L1
Duplication

Figure 8


a
DO50362 (Eso-AdenoCA)
(1) BFB cycle (L1 insertion)
(2) BFB cycle
Copy number
0
2
4
6
8
25
50
75
100
125
53Mb deletion
Chr11 (Mb)
CCND1
amplification
Coverage
Discordant read-pairs
68,145,500
68,147,000
81,652,600
81,653,300
5′-TTT A-3′
Cluster supports L1 insertion
Clusters support fold-back inversion
L1 bridge
chr11
(T)n
Unk. chr
b
Fold-back inversion model
Original chromosome
A
B
C
D
E
F
G
CCND1
L1 event bridges sister chromatids (Fig. 7d)
L1
fold-back
Lost fragments
First BFB cycle
Poleward migration and breakage
Anaphase bridge (second BFB)
CCND1 amplification
Interchromosomal model
Original chromosomes
1
2
3
4
5
L1 event bridges translocation (Fig. 6c)
translocation
First BFB cycle
Poleward migration and breakage
Anaphase bridge (second BFB)
Poleward migration and breakage
Anaphase bridge (third BFB)
c
DO26976 (Lung-SCC)
(1) BFB cycle (L1 insertion)
(2) BFB cycle
Copy number
0
2
4
6
8
10
25
50
75
100
125
50.5 Mb deletion
Chr11 (Mb)
84,427,694
CCND1
amplification