The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection

Tania Sultana, Dominic van Essen, Oliver Siol, Marc Bailly-Bechet, Claude Philippe, Amal Zine El Aabidine, Léo Pioger, Pilvi Nigumann, Simona Saccani, Jean-Christophe Andrau, et al.

Version of Record: https://www.sciencedirect.com/science/article/pii/S1097276519301479
Manuscript_0b904f39cec156ecddaa32b0f207c690

The landscape of L1 retrotransposons in the human genome is shaped by pre-insertion sequence biases and post-insertion selection

Tania Sultana[1,6,7], Dominic van Essen[1,7], Oliver Siol[2], Marc Bailly-Bechet[3], Claude Philippe[1], Amal Zine El Aabidine[4], Léo Pioger[4], Pilvi Nigumann[1], Simona Saccani[1], Jean-Christophe Andrau[4], Nicolas Gilbert[2,5], Gael Cristofari[1,*]

[1] Université Côte d'Azur, Inserm, CNRS, Institute for Research on Cancer and Aging of Nice (IRCAN), Nice, France
[2] Institut de Génétique Humaine, University of Montpellier, CNRS, Montpellier, France
[3] Université Côte d'Azur, INRA, CNRS, ISA, France
[4] Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France
[5] Institut de Médecine Régénératrice et de Biothérapie, INSERM U1183, CHU Montpellier, Montpellier, France
[6] Present address: Department of Biochemistry and Molecular Biology, University of Dhaka, Dhaka, Bangladesh

[7] These authors contributed equally

* Correspondence and lead contact: gael.cristofari@univ-cotedazur.fr

Keywords: LINE-1; L1; LINE1; transposable elements; mobile genetic element; transposon; transposition; retrotransposon; retrotransposition; integration; insertion; integration site preference; chromatin states; DNA replication; selection; genomic profiling

SUMMARY

L1 retrotransposons are transposable elements and major contributors of genetic variation in humans. Where L1 integrates into the genome can directly impact human evolution and disease. Here, we experimentally induced L1 retrotransposition in cells and mapped integration sites at nucleotide resolution. At local scales, L1 integration is mostly restricted by genome sequence biases and the specificity of the L1 machinery. At regional scales, L1 shows a broad capacity for integration into all chromatin states, in contrast with other known mobile genetic elements. However, integration is influenced by the replication timing of target regions, suggesting a link to host DNA replication. The distribution of new L1 integrations differs from those of pre-existing L1 copies, which are significantly reshaped by natural selection. Our findings reveal that the L1 machinery has evolved to efficiently target all genomic regions, and underline a predominant role for post-integrative processes on the distribution of endogenous L1 elements.

INTRODUCTION

Transposable elements are present in almost all species and significantly contribute to shaping host genome structure and function (Chuong et al., 2017). In humans, the only autonomously-active family is the long interspersed element-1 (LINE-1 or L1) (Kazazian and Moran, 2017). Our genome contains approximately 500,000 copies of this non-LTR-retrotransposon, occupying 17% of the genome (Kazazian and Moran, 2017). However, only ~100 L1 copies are still retrotransposition-competent, all of them belonging to the youngest and human-specific L1HS subfamily (Brouha et al., 2003). Each individual also has L1 copies not present in the reference genome, which contribute to as much as 20% of all structural variants in humans (Mir et al., 2015; Sudmant et al., 2015). Many of these highly polymorphic elements (in terms of presence or absence) are active and can produce new insertions (Beck et al., 2010; Gardner et al., 2017; Philippe et al., 2016; Scott et al., 2016; Tubio et al., 2014). Retrotransposition is not restricted to the germline - leading to inheritable genetic variations and occasionally to novel genetic diseases - but can also drive somatic genome rearrangements during embryogenesis, neural development, and in many cancers (Burns, 2017; Faulkner and Garcia-Perez, 2017; Hancks and Kazazian, 2016).

Intact L1 elements, approximately 6 kb long, are transcribed from an internal promoter and encode two major proteins, ORF1p and ORF2p, required for L1 retrotransposition (Moran et al., 1996). ORF2p is a combined endonuclease (EN) and reverse transcriptase (RT) (Feng et al., 1996; Mathias et al., 1991). After L1 nuclear import, ORF2p EN activity nicks the genomic DNA at a loosely defined consensus sequence (5'-TTTT/A-3' or variants of that sequence) (Cost and Boeke, 1998; Feng et al., 1996; Jurka, 1997). Then, the liberated T-rich 3' end anneals to the L1 RNA poly(A) and is extended by ORF2p RT activity to synthesize first strand L1 cDNA (Cost et al., 2002; Doucet et al., 2015; Kulpa and Moran, 2006; Luan et al., 1993; Monot et al., 2013). Subsequent steps - which are less-well defined - then result in creation of a new genomically-inserted dsDNA L1 copy, flanked by short (4-16bp) target-site duplications. Of note, the majority of insertions are 5’ truncated due to DNA repair pathways involved in insertion resolution or host defense (Coufal et al., 2011; Suzuki et al., 2009; Zingler et al., 2005). Despite a pronounced cis-preference for its own RNA, L1 is also responsible for the trans-mobilization of non-autonomous retrotransposons, such as Alu or SVA sequences, and cellular mRNAs, following a similar mechanism (Kazazian and Moran, 2017).

L1 elements carry a number of cis regulatory sequences (sense and antisense promoters, cryptic splice sites, polyadenylation signals), they can transduce 5' and 3' sequences from the donor locus to the site of insertion potentially leading to exon or cis-regulatory sequence shuffling, and they can attract chromatin or DNA modification complexes leading to altered epigenetic patterns (Cordaux and Batzer, 2009; Denli et al., 2015; Kaer and Speek, 2013; Liu et al., 2018; Tubio et al., 2014; Walter

et al., 2016). Thus, L1 insertion can considerably remodel gene structure and networks in a very short evolutionary time frame (Cordaux and Batzer, 2009; Denli et al., 2015). Additionally, expression and mobilization of inserted L1 copies is restricted to cell-type dependent permissive loci (Deininger et al., 2017; Gardner et al., 2017; Philippe et al., 2016; Scott et al., 2016; Tubio et al., 2014). Altogether, where L1 integrates into the host genome dictates both its genomic impact and the ability of the novel copy to be subsequently expressed and re-mobilized. Therefore, elucidating L1 target site selection is critical to understanding genome evolution and somatic genome plasticity in cancer or aging.

Although L1 and Alu elements share a common integration machinery and similar target site consensus motifs, they are distributed in differing chromosomal regions within the human genome (Gilbert et al., 2002; Korenberg and Rykowski, 1988; Lander et al., 2001; Wagstaff et al., 2012). L1 elements accumulate in AT-rich isochores, while Alu sequences are rather enriched in GC-rich isochores. These observations suggest that the distributions of L1 and Alu may be drastically (and differently) reshaped after integration, by recombination, purifying selection, and possibly other processes (Pavlicek et al., 2001). Direct analysis of L1 genomic integration, and comparison with the landscape of existing copies, is essential to understand the interplay between L1 retrotransposon and its observed distribution within the host genome. *De novo* L1 insertion site maps based on Sanger sequencing have provided mechanistic insight about L1 replication but only limited information on target site preference due to the low numbers of integration sites recovered (Gilbert et al., 2002; 2005; Symer et al., 2002).

Here, we assessed whether L1 can integrate homogeneously throughout the genome, or whether any genomic features or properties might favor or restrict L1 integration. To this end, we induced *de novo* L1 retrotransposition in cultured cells by transfecting a plasmid-borne active L1 element, and we mapped novel L1 target sites by a dedicated deep-sequencing approach (Philippe et al., 2016). We compared the new integration sites with a large collection of publicly available genomic datasets and with the distribution of existing endogenous L1 copies. Our data confirm the primordial role of the local DNA sequence at the target site and reveal how pre-existing biases in the genomic distribution of L1 target motifs skew the profile of new integration events. Overall, we find that new L1 insertions broadly target all the regions of the human genome, being insensitive to chromatin organization or transcriptional activity, although with a bias for early-replicating genomic domains. This distribution markedly differs from that of endogenous L1 elements, and we find that this difference predominantly results from evolutionary selection rather than from L1-induced chromatin changes.

RESULTS

ATLAS-seq can detect *de novo* engineered L1 retrotransposition in cultured cells

To facilitate the genomic characterization of pre-integration sites *in silico*, we screened cell lines well-characterized by the ENCODE project for their ability to sustain high levels of retrotransposition (K562, GM12878, HeLa S3, MCF-7, HepG2, IMR90) using the assay described in Figure 1A. Among them, HeLa S3 cells were the most permissive to L1 retrotransposition (0.1-1% of transfected cells) and were chosen to obtain a large number of independent L1 insertions. We induced retrotransposition from a plasmid-borne active L1 element expressed from its natural promoter to avoid saturating any cellular factors involved in L1 target site selectivity (Figure 1A). The L1 construct also contains a retrotransposition reporter based on the neomycin-resistance gene (Neo^R) in its 3' untranslated region, allowing us to discriminate new copies from endogenous ones and the possibility to select cells with new insertions, since it only becomes functional upon transcription, splicing, reverse transcription and integration (Moran et al., 1996). To locate new engineered L1 insertions in the genome of cultured cells, we modified ATLAS-seq, a technique originally developed to profile endogenous L1 elements (Philippe et al., 2016). We mapped 1565 *de novo* L1 target sites from 28 independent pools of HeLa S3 cells selected with G418 (referred as *L1 neo* hereafter), as well as a smaller set of 184 insertions obtained from pools of cells without any selection for retrotransposition cassette expression (referred as *L1 neo-unsel* hereafter), which were used in subsequent downstream analyses (Figure S1 and Table S2).

New engineered L1 insertions detected by ATLAS-seq display the expected hallmarks of L1 retrotransposition. First, we observe a poly(dA) tract at the junction between L1 and its 3' flanking sequence (Figure 1B). Since the poly(dA) sequence is not encoded in the plasmid DNA, this feature enables us to discriminate *bona fide* retrotransposition events from random plasmid integration or chimeric reads (see Methods). Second, sequence analysis of pre-integration sites reveals a consensus motif consistent with a canonical L1 endonuclease-mediated cleavage and with the subsequent annealing of the L1 mRNA poly(A) tail to an extended T-tract to promote reverse transcription priming (Figure 1C) (Feng et al., 1996; Jurka, 1997; Monot et al., 2013). The L1 target motif was virtually identical whether insertions were recovered with G418 selection or not (Figure S1C). Although the pre-integration site sequence is often a non-perfect match, we nevertheless found that this motif is present at the vast majority of the mapped insertion sites (Figure 1D). To summarize, ATLAS-seq can detect *de novo* engineered L1 insertions in cultured cells, and the local DNA motif is a primary determinant of L1 integration site selection.

L1 preferentially inserts into sites with low nucleosome occupancy

In vitro, target DNA structure and assembly into nucleosomes can impact L1 EN activity (Cost et al., 2001). To examine whether nucleosome occupancy influences L1 local integration site selection *in vivo*, we examined publicly available MNase-seq data from HeLa S3 at and around pre-integration sites (Lacoste et al., 2014). Because the L1 machinery preferentially targets a specific AT-rich motif (Figure 1C-D), we computationally constructed three distinct types of control datasets, each comprising the same number of sites as the experimental data (1565): (i) 1000 purely random control datasets (termed *Random*); (ii) 1000 base-composition-matched control datasets (termed *BMC*), also generated randomly but with a base-composition around the target site (±5 bp) identical to the experimental dataset; and (iii) 1000 motif-matched control datasets (termed *MMC*) consisting of a random collection of sites with an L1 target motif. L1 target sites are significantly depleted in nucleosomes (Figure 2, left). This depletion can be largely explained by low nucleosome occupancy at AT-rich sequences, since it is also observed for the BMC and the MMC (Figure 2, middle panels) but not for the Random control (Figure 2, right), consistent with previous observations indicating that these sequences disfavor nucleosome positioning (Valouev et al., 2011). We confirmed the trend of nucleosome depletion at AT-rich regions characteristic of L1 target sites using data obtained with other cell lines (Descostes et al., 2014; Schwartz et al., 2018) (Figure S2). Insertions can still occur at regions in which nucleosomes are more dense (lower part of the heat maps) implying that nucleosomal DNA is not refractory to L1 integration *per se*. Thus, L1 preferentially inserts into nucleosome-depleted DNA primarily due to sequence context.

L1 integration sites are dispersed across all chromosomes with few hotspots

HeLa cells possess an abnormal karyotype and are hypertriploid (Adey et al., 2013). To analyze the genomic distribution of L1 insertions, we corrected for aneuploidy and local copy number variations (CNV), using low-coverage whole genome sequencing (WGS) of the HeLa S3 stock used in our retrotransposition assays (~1.6X; Figure 3C, CNV track).

We find that the chromosomal distribution of new L1 insertions sites largely reflects chromosome size and copy number (Figure 3A). Endogenous L1 copies are enriched in the X chromosome, where they have been proposed to contribute to X chromosome inactivation (Bailey et al., 2000; Lyon, 1998) (see also Figure 3C, endogenous L1 tracks). However, new L1 insertions are not significantly enriched in the X chromosome under our experimental conditions, suggesting that the evolutionary accumulation of L1 in the X chromosome results either from post-integration selection or requires a stably inactivated X (Xi) chromosome (like many cancer cell lines, HeLa cells do not express the Xist RNA and only contain activated X chromosomes (Kawakami et al., 2004)). New insertions are over-represented in chromosome 1, indicating that L1 integration is not perfectly random (Figure 3A). We

next examined the overall spacing of *de novo* L1 insertions (Figure 3B), and compared it to each simulated control dataset, or to 1000 random sampling of 1565 reads from HeLa S3 WGS (Random, BMC, and MMC controls, see above; and WGS). New L1 insertions are more clustered than expected, particularly at inter-insertion site distances in the 10kb-1Mb range. This clustering is observed irrespective of the control dataset used, indicating that it does not result from an uneven distribution of AT-rich sequences, from clusters of L1 target motifs in the human genome, or from aneuploidy of the HeLa S3 cell line. To further test whether some genomic locations are preferred, we scanned the genome by 0.1, 0.5, 1, 5 or 10 Mb-windows for *de novo* L1 insertions. Using 10 Mb-bins, we detect only a single insertional hotspot in chromosome 1 (overlapping 1p31-1p32.1 cytobands, Figure 3C and Table S3), consistent with the over-representation of insertions in chromosome 1 (Figure 3A). Two additional hotspots on chromosomes 5 and 12 are detected when smaller bin sizes are considered (Figure S3 and Table S3). These apparent hotspots do not arise from karyotype alterations since our analysis corrects for CNV. None of the detected hotspots correspond to regions with a significantly enrichment of endogenous elements or L1 target motifs (endogenous L1 and MMC tracks, Figure 3C). Recurrent L1-mediated inherited insertions at the same nucleotide position in the *BTK*, *NF1*, *F9* or *APC* genes have previously been found, suggesting that these loci might also represent L1 integration hotspots (Hancks and Kazazian, 2016). However, none of these genes were hit in our experiments. In summary, the distribution of new L1 insertions measurably deviates from homogeneity, and some genomic locations are favored over others, independently of the number of L1 target motif sequences or of existing instances of previous L1 insertions.

To model L1 target site selection at different genomic scales, we analyzed the densities of L1 target motifs in differently-sized bins surrounding L1 integration sites. At a given scale, if an L1 element inserts into available target motifs independently of other factors, then the resulting integration sites will be enriched for bins with higher numbers of target motifs; on the other hand, if an L1 element chooses between potential target regions independently of the number of motifs, then the resulting integration sites will mirror the genomic distribution of motif densities. At the kilobase scale, L1 integration sites are highly enriched for elevated local L1 target motif densities (Figure S3B), and closely-fit a model of unbiased selection between local motifs. However, at larger scales, L1 integration deviates significantly from this pattern, and at scales exceeding 10-100 kb it is best modeled by selection of genomic regions independently from the presence or number of target motifs (Figure S3B). Thus, consistent with the large-scale clustering observed in Figure 3B, L1 elements appear to localize to genomic regions of >10-100 kb independently of motif density; and at smaller scales within these, they insert near-randomly among available target motifs. We therefore next considered whether the choice of genomic target regions is dictated by particular chromatin features or processes.

L1 integration is only modestly influenced by local chromatin states and histone marks

To identify potential regulatory features of L1 integration, we measured the association of new L1 insertion profiles with chromatin segmentation states and histone modifications obtained from the ENCODE project (Ernst and Kellis, 2012; Hoffman et al., 2013). For comparison, we also analyzed simulated background datasets (Random, BMC and MMC), existing endogenous L1HS-Ta found in HeLa S3 cells (Philippe et al., 2016) (referred as *L1 endo*), and previously published genome-wide integration profiles of retroviruses and other transposable elements (Gogol-Döring et al., 2016; LaFave et al., 2014; Wang et al., 2007). *De novo* L1 insertions show a modest but significant association with weak enhancer and weak transcription chromatin segments (Figure 4A), and with histone modifications that are characteristic of enhancers (H3K4me1) or transcription over gene bodies (H3K36me3) (Figure 4C). However, the levels of these associations are low (excess overlaps represent only 2-3% of all L1 insertions; see Figure 4B), and it is noteworthy that L1 retrotransposons show the least deviation from randomness, as compared to other transposable elements and retroviruses (Figure 4 and Figure S4A). The profile obtained with *de novo* insertions recovered in the absence of G418 selection (L1 neo-unsel) is similar, although we could only test the association with the most abundant chromatin segments due to the smaller dataset size (Figure S4B). The L1 integration profile is in sharp contrast to that observed for murine leukemia virus (MLV) and PiggyBac (PB), which strongly accumulate in chromatin regions related to strong enhancers, promoters and transcription start sites (Gogol-Döring et al., 2016; Sultana et al., 2017) (Figure 4 and Figure S4A). L1 also strongly differs from human immunodeficiency virus (HIV) and Sleeping Beauty (SB), which rather integrate into transcribed regions with different levels of preference strength (Figure 4 and Figure S4A), consistent with previous reports (Gogol-Döring et al., 2016; Sultana et al., 2017).
To ensure that the chromatin segmentation model does not obscure any more-significant associations between L1 insertion and any individual genomic feature(s), we also compared L1 insertion sites to each of the entire ENCODE collection of 346 publicly-available ChIP-seq, DNase-seq, and FAIRE-seq datasets, as well as to annotated features from the UCSC genome browser. In line with the results above, *de novo* L1 insertions are under-represented in GC-rich regions (which disfavor the AT-rich insertion motif), but exhibit negligible or low associations with almost all assayed genomic features (which encompass transcription factors, histone variants and modifications, chromatin enzymes, and biochemical properties of DNA and chromatin; Figure S5A-C), and the few significant associations - which primarily correspond to features typical of enhancers (DNaseI hypersensitive sites and H3K4me1) - are nevertheless characterized by modest z-scores, which are far below the levels of associations exhibited by other transposons or viruses. We validated the

(modest) association with H3K4me1 using ChIP-seq with the same cell stock used in our assays (Figure S5D-E).

The absence of a strong association between L1 insertion and any chromatin features also contrasts markedly with the measured specificities of a variety of mammalian (DNase-I), bacterial (CviPII nickase, micrococcal nuclease, Dam methyltransferase), and phage (Tn5 transposase) enzymes on human chromatin (Figure S4C). Every tested enzyme activity exhibits moderate-to-strong associations with a range of chromatin segments, which are consistently higher than the strongest associations measured for L1 insertion. In conclusion, the ability of L1 to consistently insert into most or all genomic regions with similar efficiency represents a unique and highly-developed property - rather than the default expected activity for a chromatin-templated process - and may contribute to the evolutionary success of L1 elements and their abundance in the human genome.

De novo L1 insertions and endogenous copies associate with distinct chromatin segments

For most detectable associations, *de novo* and endogenous L1HS-Ta insertions show almost reciprocal enrichment and depletion, suggesting that post-integrative phenomena can alter the relative distribution of L1 elements among the different chromatin states or genomic regions. Such post-integrative mechanisms might include purifying selection or L1-mediated epigenetic alterations (see below). To test further this possibility, we measured the association of endogenous primate-specific L1 elements of increasing age (from L1HS, *i.e.* L1PA1, to L1PA17) with the different chromatin segments (Figure 4D and Figure S6). Although *de novo* L1 insertions are modestly enriched in chromatin segments annotated as weak enhancers (EnhWk) or weakly transcribed (TxWk), endogenous L1HS to L1PA5 exhibit increasing levels of depletion from these regions. Conversely, *de novo* insertions are neither enriched nor depleted in segments annotated as quiescent (Quies), yet reference L1HS to L1PA3-PA5 elements show increasing levels of enrichment in these low-activity regions. L1 loci as old as 20 Myr (the estimated age of the L1PA5 family) (Khan et al., 2006) show evidence for ongoing changes in their associations with genome functions. Beyond this age, the strengths of these positive or negative associations is stable or progressively reduced. Collectively, these results indicate that integration of L1 elements exhibits a remarkable **in**difference to known genomic features in comparison to other mobile genetic elements, and that the genomic association of endogenous L1 copies with particular chromatin segments largely reflects post-integrative processes rather than integration site preference.

It was recently proposed that L1 preferentially accumulates into older L1 elements in the brain (Jacob-Hirsch et al., 2018), in contrast to previous reports (Szak et al., 2002). We did not find strong evidence for preferred integration into existing transposable elements or repeats, besides AT-rich

low-complexity repetitive DNA (Figure S5B). We conclude that if nested L1 insertions are indeed enriched in particular scenarios, this rather results from post-integrative processes.

De novo L1 insertions mirror the unbalanced target motif distribution in transcribed regions

The modest enrichment of new L1 integrants in chromatin segment representative of weak transcription prompted us to directly test their potential association with genes and transcription. Annotated protein coding genes cover approximately 45% of the human genome (GENCODE v19, from transcription start to termination sites, including introns, see Methods). *De novo* engineered L1 insertions are not significantly enriched in genes, in contrast to those from other well-characterized DNA transposons or retroviruses (Figure 5A). This pattern also differs from that of endogenous L1HS-Ta elements already present in Hela S3 cells, which are significantly depleted from genic regions (36%), likely due to post-integrative negative selection, as previously observed for reference L1 elements (Medstrand et al., 2002; Smit, 1999). Although, the median expression level of genes with new L1 insertions is higher than for the Random and BMC datasets, it is not significantly different from that of genes containing L1 target site motifs (MMC, Figure 5B). Moreover, the slight association between higher gene expression levels and L1 insertions is much less pronounced than observed for other transposable elements or retroviruses, in particular HIV, known to preferentially integrate into highly expressed genes (Sultana et al., 2017).

L1 retrotransposons inserted within genes and recorded in the human reference genome occur more frequently in the opposite orientation with respect to their enclosing gene (Medstrand et al., 2002; Smit, 1999; Zhang et al., 2011). Since a majority of disease-causing L1 insertions are in the sense orientation relative to the disrupted gene (Hancks and Kazazian, 2016), a generally accepted explanation is that sense insertions are more likely to be detrimental and counter-selected after integration. We found that both endogenous and *de novo* insertions are significantly biased toward antisense insertions, although this effect is more pronounced for endogenous L1HS-Ta copies (Figure 5C and 5E). However, the distribution of L1 target motifs (MMC) in genes is also highly skewed toward the antisense orientation and parallels the orientation bias of *de novo* L1 insertions (Figure 5D and 5E). To summarize, L1 orientation bias in genes results from an unbalanced distribution of L1 target site motifs between the transcribed and non-transcribed strands, leading to a significant bias of orientation at the time of integration, and this effect is further strengthened by purifying selection after integration.

L1 integration is influenced by host DNA replication

Existing L1 elements present in human or murine genomes have been shown to be enriched in late replicating domains, an observation correlated with the accumulation of L1 elements in GC-poor isochores (Buckley et al., 2017; Hansen et al., 2010; Hiratani et al., 2008). This enrichment is particularly evident for more recent, primate- and human-specific L1 subfamilies (Figure S7A). To directly investigate whether this reflects preferential L1 insertion into late-replicating regions, we computed the association of new L1 insertions with replication timing, and with other genomic properties linked to DNA replication such as origins of replication and replication fork directionality (RFD), using publicly available datasets.

We observe that new L1 insertions occur preferentially within early-replicating regions of the genome (Figure 6A, left, and 6B). This represents the strongest association with any genomic feature analyzed in this study, and is sufficient to explain the detectable skew of L1 insertions in other large-scale domains (including lamin-associated domains (LADs) and topological A and B domains; see Figure S5A). Integration sites recovered without selecting for the reporter cassette are also enriched in early replicating regions and their replication timing markedly differs from that of endogenous L1 elements, although this is less pronounced than for selected events (Figure S7A). Neither local base-composition, nor the density of L1 target motifs at the insertion sites (BMC and MMC control datasets, Figure 6B), nor the increased copy number of already-replicated regions (Figure S7B), are responsible for this association. We also did not find evidence of enrichment at replication origins detected by short nascent strand sequencing in HeLa cells (SNS-seq, Figure 6B, left) (Besnard et al., 2012).

We next investigated whether the orientation of *de novo* L1 insertions is influenced by the directionality of the replication machinery. Replication fork directionality (RFD) was obtained from publicly available Okazaki fragment sequencing (OK-seq) data in HeLa cells (Petryk et al., 2016) (see Figure 6 legend). Positive strand insertions (*i.e.*, initial endonuclease cut occurring at the bottom strand) are highly enriched when the replication fork is moving leftward, and vice versa (Figure 6A, right; Figure 6C). Notably, the distribution of genomic L1 target motifs exhibits a near-identical bias (Figure 6C), mirroring the A/T-skew between leading and lagging strands which has been described in most genomes (Huvet et al., 2007; Touchon et al., 2005). We conclude that the biased orientation of new L1 integration events toward replication fork is driven by an intrinsic property of the human genome, similarly to what we observed for transcription.

Overall, the associations with host DNA replication represent the most-pronounced departures of *de novo* L1 insertions from homogeneity, providing a first hint that these two different processes may be mechanistically linked.

Post integration selection vs L1-mediated instruction

Our results indicate that new L1 insertions and endogenous copies are associated with distinct chromatin states and replication timing (see above, Figure 4 and Figure 6). Two broad post-integrative mechanisms could contribute to these divergences: L1 insertions could be subject to evolutionary selection - so that instances that negatively affect fitness are purged, or the presence of L1 insertions might itself instruct genomic changes that alter the local chromatin state or function. To discriminate between these scenarios, we analyzed 1634 natural L1HS insertions detected in 12 different human cell-lines (Philippe et al., 2016). We first identified the set of private insertions that are each present only in a single cell-line (Figure 7A). The set of private insertions specific to HeLa S3 cells exhibits associations with chromatin states and replication timing that match those of fixed, non-polymorphic L1HS insertions, and differ from those of *de novo* insertions, confirming the influence of post-integrative effects on these insertion sites (Figure 7C, compare 'present' with 'fixed' and 'L1 neo' sites). We therefore compared the chromatin states at private insertion sites that are specific for HeLa S3 ('present' sites) with those specific for other cell-lines ('absent' sites, which in HeLa S3 are unused and represent a pre-insertion configuration) (Figure 7B). Any instructive effects that could be driven by the presence of L1 would not occur in HeLa S3 cells at these sites. Nevertheless, we found that the distribution of chromatin states at 'absent' sites was significantly different from those of random genomic locations or unselected new insertions, and instead closely-matched that of endogenous insertion sites at which L1 is present. This rules out L1-dependent instruction as a major determinant (Figure 7C), and is consistent with a dominant role for evolutionary selection in the association of endogenous L1 copies with the quiescent chromatin state and in their depletion from other annotated chromatin segments. Similarly, the disparity between the L1 insertion preference for early-replicating regions and the enrichment of endogenous L1 copies in late-replicating regions is primarily driven by post-integrative selection, rather than L1-driven alterations in replication timing, since natural insertion sites from multiple other cell-lines that are unused in HeLa S3 cells exhibit the same bias for late-replication in Hela S3 (Figure 7B).
Our data do not exclude that specific L1-derived sequences may contribute to modifying some aspects of local chromatin, and indeed this has been described (Liu et al., 2018; Walter et al., 2016). Nonetheless, they clearly indicate that the dominant process that shapes the distribution of recent L1 insertions in the human genome is evolutionary selection, and that the influences of initial target-site preference or of instructive chromatin changes are largely concealed when analyzing endogenous copies of human-specific L1s.

DISCUSSION

In this study, we characterized a set of 1565 L1 target sites, recovering almost an order of magnitude more insertions than the combined total from all previous studies (Gilbert et al., 2002; 2005; Symer et al., 2002) (Figure S1 and Table S2). The retrotransposition assay using engineered L1 elements in cultured cells faithfully recapitulates many properties of naturally occurring insertion (Rangwala and Kazazian, 2009), but has also limitations. First, the reporter cassette inserted in L1 3' UTR could affect its integration pattern. However, genomic sequences of similar size and downstream of active L1s are often naturally mobilized, through 3' transduction (Kazazian and Moran, 2017). Second, the use of a selectable marker to enrich for cells with integration events could influence the distribution of functional regions that can be detected as target sites. We do not favor this possibility since a smaller subset of new L1 insertions recovered without any selection in this study exhibit identical properties to G418-selected insertions, and we do not observe enrichment for detected L1 insertions in expressed genes. The recovery approach used here only provides information on the 3' junctions and thus on the initial steps of target-primed reverse transcription. It is possible that a minor fraction of the integration events was resolved through recombination pathways as previously observed (Gilbert et al., 2005), but this would not influence our conclusions since the initial targeting would be unaffected. Similarly, recombination with existing copies, as previously described for limited cases of Alu insertions (Roy et al., 2000), is unlikely to bias our analysis since *de novo* insertions are not enriched at the locations of any other classes of endogenous L1 (Figure S5) and only 8 out 1565 (0.5%) were found within an existing L1HS element (Table S2).

Finally, our strategy of analyzing *de novo* L1 integration patterns in a single, well-studied cell line was designed to benefit from the availability of a broad range of genomic and epigenomic datasets. Although these data generally come from the same cell line, they derive from independent cell stocks cultured in conditions that might differ from those used in our L1 retrotransposition assay. Despite this limitation, we could corroborate our findings using multiple external datasets (Figures S2, S5C), and we validated the most-significant detected association with a histone mark (H3K4me1; Figures 4C, S5C) using data generated from our own cell stocks and culture conditions (Figure S5D-E). However, it is conceivable that some aspects of L1 transposition may also be influenced by cell line-specific factors, which our analysis could overlook.

At a local scale the distribution of *de novo* L1 insertions is driven by target sequence specificity

Almost all recovered insertions occur at pre-integration sites bearing a consensus AT-rich L1 target motif (Figure 1C-D) and at local scales (100bp-1kb) L1 insertions are consistent with a model of unbiased selection between motifs (Figure S3). Pre-existing biases in the distribution of this motif has a measurable impact on new insertions. The L1 target motif contains multiple AA-dinucleotides that

are naturally disfavored by nucleosomes (Valouev et al., 2011) leading to low nucleosome occupancy at sites of L1 insertions (Figure 2 and Figure S2). Similarly, the orientation bias that we detect for new L1 integration sites relative to transcription and replication directionalities (Figure 5 and Figure 6) can be explained by the known unequal distribution of A/T-content between the two DNA strands of the human genome (Green et al., 2003; Huvet et al., 2007; Touchon et al., 2005).

Endogenous genic L1 elements are more frequently oriented antisense relative to their surrounding gene, a bias ascribed to the detrimental effects of sense-oriented insertions (Hancks and Kazazian, 2016; Zhang et al., 2011). A recent study proposed that this bias could also result from the activity of the transcription-coupled repair pathway during retrotransposition, although the small number of *de novo* insertions analyzed limited the statistical power of this analysis (Servant et al., 2017). Our findings reveal that the unbalanced distribution of L1 target motifs on the transcribed *vs* non-transcribed strands is sufficient to account for the orientation bias of novel insertion events, without any need to invoke cellular host defense mechanisms, and that this bias is further strengthened by evolutionary selection (Figure 5). We note that favoring the non-transcribed strand could naturally serve to limit collisions between L1 reverse transcription and host transcription that might otherwise activate DNA damage repair pathways.

De novo L1 insertions are preferentially oriented facing in the opposite direction to replication fork progression (Figure 6). Given the directionality of target-primed reverse transcription, this observation implies that L1 generally initiates retrotransposition in the strand that serves or will serve as a template for lagging strand synthesis or in the neo-synthesized leading strand. Again, this property primarily results from the unbalanced distribution of L1 target motifs between leading and lagging strands, a consequence of compositional strand asymmetries and A/T-skew associated with replication and observed in most species (Huvet et al., 2007; Touchon et al., 2005). Nevertheless, it is possible that this intrinsic replication-associated strand asymmetry may have been exploited by L1 to facilitate steps required for integration resolution, including second-strand synthesis and ligation (see below).

At a regional scale *de novo* L1 insertions are minimally-influenced by chromatin state and activity

Almost all transposable elements and retroviruses studied until now exhibit highly-biased integration into host genomes (Sultana et al., 2017). However, by systematically examining a large variety of annotated genomic features, we found that L1 integration is remarkably insensitive to described chromatin states and to gene organization (Figure 4 and Figure 5). This nearly-neutral pattern of L1 insertion is starkly different to those of other well-characterized transposable elements (SB, PB) or retroviruses (MLV, HIV) or to other enzymatic processes that act on DNA as a substrate (Figure S4),

and strongly implies that this is a highly-specific and selected property of the L1 machinery (and hints to a possibility that it may even act on non-chromatinized DNA; see below).

At large genomic scales, L1 integration is influenced by the replication timing of the target sites

At scales between 10 kb and 1 Mb, L1 insertions are more clustered than expected (Figure 3), and localization to large-scale regions occurs independently of fine-scale target motif choice (Figure S3). Accordingly, the most prominent, but also unanticipated, feature of *de novo* L1 insertions in our experiments is their preferential targeting of early-replicating regions of the genome (Figure 6 and Figure S7): these occur in megabase-scale domains (Hansen et al., 2010), and their enrichment for *de novo* L1 insertions suggests a possible link between L1 integration and host DNA replication. Although L1 is capable of retrotransposition in non-dividing cells, cell cycle progression and S-phase seem to promote its mobilization (Kubo et al., 2006; Macia et al., 2017; Mita et al., 2018; Xie et al., 2013), and several factors acting at replication forks (PCNA, MCM proteins, TOP1, PARP1, and RPA1) have been shown to interact with L1 ORF2p (Goodier, 2016; Liu et al., 2018; Pizarro and Cristofari, 2016; Taylor et al., 2018; 2013). Collectively, these observations are consistent with a model whereby the L1 machinery is recruited to replication forks by host factors in dividing cells, and that early-replicating regions may be preferentially targeted due to limiting ORF2p levels (Alisch et al., 2006; Taylor et al., 2013). Due to the unbalanced density of L1 target motifs between the two replicating DNA strands, L1 favors initiation of reverse transcription from the cleaved lagging strand template DNA, where it could hijack host enzymatic activities and DNA structures involved in lagging strand synthesis to prime L1 second strand synthesis and resolve integration. Such a model would also be consistent with the low influence of chromatin states on L1 insertion site preference, since local chromatin organization at the replication fork is expected to be heavily disturbed. Nevertheless, the ability of L1 to retrotranspose in both dividing and non-dividing cells, suggests that this model might represent an opportunistic pathway for L1 insertion, rather than a strict requirement. Our work is consistent with an independent and parallel study (Flasch *et al.* 2019), which detected association of new L1 insertions with early replication domains in a subset of cell types, although also with late-replicating domains using other cell lines and experimental conditions. Further studies will be needed to understand the conditions that lead to preference for early- or late-replicating regions.

The distribution of existing L1 insertions is dominated by post-integration selection

Our characterization of new L1 insertions allowed us to examine the respective contributions of insertion site selection and post-integrative events to the observed distribution of L1 elements within germline and somatic genomes. We found that the landscape of endogenous L1 copies differs

significantly from that of new insertions, being enriched in quiescent chromatin segments and late-replicating regions, and depleted from genes. It thereby more closely resembles the profile of insertions recovered from tumors, which accumulate in heterochromatic regions (Bryner et al., 2017; Tubio et al., 2014). Somatic L1 insertions in the brain were previously reported to preferentially accumulate in enhancers (Upton et al., 2015), which is reminiscent to the modest association with weak enhancers we observe for *de novo* insertions, suggesting that post-integrative processes operate more drastically in tumors than in the normal brain. By analyzing a large collection of natural insertion sites from a panel of 12 different cell-lines, we show that the predominant post-integrative process that models the distribution of endogenous L1s is purifying selection, and not L1-dependent chromatin alterations (Figure 7). Recombinational deletion among distant L1 or Alu copies has the potential to deplete retrotransposon copies from large and distinct genomic compartments, and in conjunction with natural selection, this could contribute to the differential distributions of L1 and Alu (Han et al., 2008; Sen et al., 2006). We conclude that mapping *de novo* insertions, rather than existing insertions that accumulated in the genome over a long period of time or were selected by clonal expansion during tumor growth, is essential to separate insertion site preference from evolutionary selection.

In sum, our data reveal the unique properties of the L1 machinery which allow this transposable element to broadly target most regions of the human genome. They also highlight how the landscape of insertions is driven by pre-existing aspects of genomic sequence and how it is reshaped throughout evolution by selection. In a more general way, our work illustrates the co-adaptive trajectories of transposable elements with their host.

ACKNOWLEDGMENTS

We are grateful to C. Baudoin and IRCAN Genomics Core Facility for sequencing (supported by FEDER, Région Provence Alpes Côte d'Azur, Conseil Départemental 06, ITMO Cancer Aviesan (plan cancer) and Inserm), and to O. Croce and IRCAN bioinformatics service for computing resources. We are grateful to the ENCODE Consortium for RNA-seq, ChIP-seq and Repli-seq data. This work was supported by a joint grant to SS and GC from the Fondation pour la Recherche Médicale (FRM-DEP20131128533), by grants to GC from the European Research Council (ERC-2010-StG 243312, Retrogenomics), the Agence Nationale de la Recherche (LABEX SIGNALIFE, ANR-11-LABX-0028-01; RETROMET, ANR-16-CE12-0020), the Cancéropôle PACA (Projet Emergence), CNRS (GDR 3546), and the University Hospital Federation (FHU) OncoAge. TS was supported by a joint Erasmus Mundus Mobility with Asia (EMMA) fellowship between Université Côte d'Azur (France) and University of

Dhaka (Bangladesh). Work in the laboratory of NG was supported by the Agence Nationale de la Recherche (ANR-12-BSV6-0003, RETROGENO).

AUTHOR CONTRIBUTIONS

NG and GC conceived the study. OS and NG designed and performed the retrotransposition assays. TS and CP optimized the ATLAS-seq procedure and prepared the libraries. PN contributed to early ATLAS-seq technical developments. TS, DvE, MBB and GC designed and conducted the computational analyses. TS, LP, AZEA and JCA designed and conducted the nucleosome occupancy analyses. DvE and SS provided computational tools and performed additional experiments. NG and SS provided critical review of work. TS, DvE and GC drafted the manuscript. DvE and GC revised the manuscript. GC supervised the project.

DECLARATION OF INTERESTS

Oliver Siol is currently an employee of TES-AMM (computer material shop).

REFERENCES

Adey, A., Burton, J.N., Kitzman, J.O., Hiatt, J.B., Lewis, A.P., Martin, B.K., Qiu, R., Lee, C., and Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature *500*, 207–211.

Alisch, R.S., Garcia-Perez, J.L., Muotri, A.R., Gage, F.H., and Moran, J.V. (2006). Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev *20*, 210–224.

Bailey, J.A., Carrel, L., Chakravarti, A., and Eichler, E.E. (2000). Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. Proc. Natl. Acad. Sci. U.S.a. *97*, 6634–6639.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. Cell *141*, 1159–1170.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., and Lemaitre, J.-M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat. Struct. Mol. Biol. *19*, 837–844.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U.S.a. *100*, 5280–5285.

Bryner, D., Criscione, S., Leith, A., Huynh, Q., Huffer, F., and Neretti, N. (2017). GINOM: A statistical framework for assessing interval overlap of multiple genomic features. PLoS Comput Biol *13*, e1005586.

Buckley, R.M., Kortschak, R.D., Raison, J.M., and Adelson, D.L. (2017). Similar Evolutionary Trajectories for Retrotransposon Accumulation in Mammals. Genome Biol Evol *9*, 2336–2353.

Burns, K.H. (2017). Transposable elements in cancer. Nat Rev Cancer *17*, 415–424.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet *18*, 71–86.

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

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

Cost, G.J., Golding, A., Schlissel, M.S., and Boeke, J.D. (2001). Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res *29*, 573–577.

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

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Marchetto, M.C.N., Muotri, A.R., Mu, Y., Carson, C.T., Macia, A., Moran, J.V., and Gage, F.H. (2011). Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proc. Natl. Acad. Sci. U.S.a. *108*, 20382–20387.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188–1190.

Deininger, P., Morales, M.E., White, T.B., Baddoo, M., Hedges, D.J., Servant, G., Srivastav, S., Smither, M.E., Concha, M., deHaro, D.L., et al. (2017). A comprehensive approach to expression of L1 loci. Nucleic Acids Res *45*, e31–e31.

Denli, A.M., Narvaiza, I., Kerman, B.E., Pena, M., Benner, C., Marchetto, M.C.N., Diedrich, J.K., Aslanian, A., Ma, J., Moresco, J.J., et al. (2015). Primate-specific ORF0 contributes to retrotransposon-mediated diversity. Cell *163*, 583–593.

Descostes, N., Heidemann, M., Spinelli, L., Schüller, R., Maqbool, M.A., Fenouil, R., Koch, F., Innocenti, C., Gut, M., Gut, I., et al. (2014). Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells. Elife *3*, e02105.

Doucet, A.J., Wilusz, J.E., Miyoshi, T., Liu, Y., and Moran, J.V. (2015). A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. Mol. Cell *60*, 728–741.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Faulkner, G.J., and Garcia-Perez, J.L. (2017). L1 Mosaicism in Mammals: Extent, Effects, and Evolution. Trends Genet *33*, 802–816.

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

Fenouil, R., Descostes, N., Spinelli, L., Koch, F., Maqbool, M.A., Benoukraf, T., Cauchy, P., Innocenti, C., Ferrier, P., and Andrau, J.-C. (2016). Pasha: a versatile R package for piling chromatin HTS data. Bioinformatics *32*, 2528–2530.

Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., 1000 Genomes Project Consortium, and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res *27*, 1916–1929.

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. (2005). Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol. Cell. Biol. *25*, 7780–7795.

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

Gogol-Döring, A., Ammar, I., Gupta, S., Bunse, M., Miskey, C., Chen, W., Uckert, W., Schulz, T.F., Izsvák, Z., and Ivics, Z. (2016). Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4(+) T Cells. Mol. Ther. *24*, 592–606.

Goodier, J.L. (2016). Restricting retrotransposons: a review. Mob DNA *7*, 16.

Green, P., Ewing, B., Miller, W., Thomas, P.J., NISC Comparative Sequencing Program, and Green, E.D. (2003). Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet *33*, 514–517.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812.

Han, K., Lee, J., Meyer, T.J., Remedios, P., Goodwin, L., and Batzer, M.A. (2008). L1 recombination-associated deletions generate human genomic variation. Proc. Natl. Acad. Sci. U.S.a. *105*, 19366–19371.

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

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. U.S.a. *107*, 139–144.

Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.-W., Lyou, Y., Townes, T.M., Schübeler, D., and Gilbert, D.M. (2008). Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol *6*, e245.

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res *41*, 827–841.

Huvet, M., Nicolay, S., Touchon, M., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., and Thermes, C. (2007). Human gene organization driven by the coordination of replication and transcription. Genome Res *17*, 1278–1285.

Jacob-Hirsch, J., Eyal, E., Knisbacher, B.A., Roth, J., Cesarkas, K., Dor, C., Farage-Barhom, S., Kunik, V., Simon, A.J., Gal, M., et al. (2018). Whole-genome sequencing reveals principles of brain retrotransposition in neurodevelopmental disorders. Cell Res. *28*, 187–203.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. U.S.a. *94*, 1872–1877.

Kaer, K., and Speek, M. (2013). Retroelements in human disease. Gene *518*, 231–241.

Kawakami, T., Zhang, C., Taniguchi, T., Kim, C.J., Okada, Y., Sugihara, H., Hattori, T., Reeve, A.E., Ogawa, O., and Okamoto, K. (2004). Characterization of loss-of-inactive X in Klinefelter syndrome and female-derived cancer cells. Oncogene *23*, 6163–6169.

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

Khan, H., Smit, A., and Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res *16*, 78–87.

Korenberg, J.R., and Rykowski, M.C. (1988). Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. Cell *53*, 391–400.

Kubo, S., Seleme, M.D.C., Soifer, H.S., Perez, J.L.G., Moran, J.V., Kazazian, H.H., and Kasahara, N. (2006). L1 retrotransposition in nondividing and primary human somatic cells. Proc. Natl. Acad. Sci. U.S.a. *103*, 8036–8041.

Kulpa, D.A., and Moran, J.V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat. Struct. Mol. Biol. *13*, 655–660.

Lacoste, N., Woolfe, A., Tachiwana, H., Garea, A.V., Barth, T., Cantaloube, S., Kurumizaka, H., Imhof, A., and Almouzni, G. (2014). Mislocalization of the centromeric histone variant CenH3/CENP-A in human cells depends on the chaperone DAXX. Mol. Cell *53*, 631–644.

LaFave, M.C., Varshney, G.K., Gildea, D.E., Wolfsberg, T.G., Baxevanis, A.D., and Burgess, S.M. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res *42*, 4257–4269.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., H omer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1.G.P.D.P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liu, N., Lee, C.H., Swigut, T., Grow, E., Gu, B., Bassik, M.C., and Wysocka, J. (2018). Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature *553*, 228–232.

Luan, D.D., Korman, M.H., Jakubczak, J.L., and Eickbush, T.H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell *72*, 595–605.

Lyon, M.F. (1998). X-chromosome inactivation: a repeat hypothesis. Cytogenet. Cell Genet. *80*, 133–137.

Macia, A., Widmann, T.J., Heras, S.R., Ayllon, V., Sanchez, L., Benkaddour-Boumzaouad, M., Muñoz-Lopez, M., Rubio, A., Amador-Cubero, S., Blanco-Jimenez, E., et al. (2017). Engineered LINE-1 retrotransposition in nondividing human neurons. Genome Res *27*, 335–348.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, pp–10.

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

Medstrand, P., van de Lagemaat, L.N., and Mager, D.L. (2002). Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res *12*, 1483–1495.

Mir, A.A., Philippe, C., and Cristofari, G. (2015). euL1db: the European database of L1HS retrotransposon insertions in humans. Nucleic Acids Research *43*, D43–D47.

Mita, P., Wudzinska, A., Sun, X., Andrade, J., Nayak, S., Kahler, D.J., Badri, S., Lacava, J., Ueberheide, B., Yun, C.Y., et al. (2018). LINE-1 protein localization and functional dynamics during the cell cycle. Elife *7*, 210.

Monot, C., Kuciak, M., Viollet, S., Mir, A.A., Gabus, C., Darlix, J.-L., and Cristofari, G. (2013). The specificity and flexibility of l1 reverse transcription priming at imperfect T-tracts. PLoS Genet. *9*, e1003499.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H. (1996). High frequency retrotransposition in cultured mammalian cells. Cell *87*, 917–927.

Pavlicek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J.V., and Bernardi, G. (2001). Similar integration but different stability of Alus and LINEs in the human genome. Gene *276*, 39–45.

Petryk, N., Kahli, M., d'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C.-L., and Hyrien, O. (2016). Replication landscape of the human genome. Nat Commun *7*, 10208.

Philippe, C., Vargas-Landin, D.B., Doucet, A.J., van Essen, D., Vera-Otarola, J., Kuciak, M., Corbin, A., Nigumann, P., and Cristofari, G. (2016). Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. Elife *5*, 166.

Pizarro, J.G., and Cristofari, G. (2016). Post-Transcriptional Control of LINE-1 Retrotransposition by Cellular Host Factors in Somatic Cells. Front Cell Dev Biol *4*, 14.

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

Ramos Pittol, J.M., Oruba, A., Mittler, G., Saccani, S., and van Essen, D. (2018). Zbtb7a is a transducer for the control of promoter accessibility by NF-kappa B and multiple other transcription factors. PLoS Biol *16*, e2004526.

Rangwala, S.H., and Kazazian, H.H. (2009). The L1 retrotransposition assay: a retrospective and toolkit. Methods *49*, 219–226.

Roy, A.M., Carroll, M.L., Nguyen, S.V., Salem, A.H., Oldridge, M., Wilkie, A.O., Batzer, M.A., and Deininger, P.L. (2000). Potential gene conversion and source genes for recently integrated Alu elements. Genome Res *10*, 1485–1495.

Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics *20*, 3246–3248.

Schwartz, U., Németh, A., Diermeier, S., Exler, J.H., Hansch, S., Maldonado, R., Heizinger, L., Merkl, R., and Längst, G. (2018). Characterizing the nuclease accessibility of DNA in human cells to map higher order structures of chromatin. Nucleic Acids Res *154*, 515.

Scott, E.C., Gardner, E.J., Masood, A., Chuang, N.T., Vertino, P.M., and Devine, S.E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res *26*, 745–755.

Sen, S.K., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P.A., Dyer, M., Cordaux, R., Liang, P., and Batzer, M.A. (2006). Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet *79*, 41–53.

Servant, G., Streva, V.A., and Deininger, P.L. (2017). Transcription coupled repair and biased insertion of human retrotransposon L1 in transcribed genes. Mob DNA *8*, 18.

Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev *9*, 657–663.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Sultana, T., Zamborlini, A., Cristofari, G., and Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet *18*, 292–308.

Suzuki, J., Yamaguchi, K., Kajikawa, M., Ichiyanagi, K., Adachi, N., Koyama, H., Takeda, S., and Okada, N. (2009). Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition. PLoS Genet. *5*, e1000461.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327–338.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol. *3*, research0052.

Taylor, M.S., Altukhov, I., Molloy, K.R., Mita, P., Jiang, H., Adney, E.M., Wudzinska, A., Badri, S., Ischenko, D., Eng, G., et al. (2018). Dissection of affinity captured LINE-1 macromolecular complexes. Elife *7*, 210.

Taylor, M.S., Lacava, J., Mita, P., Molloy, K.R., Huang, C.R.L., Li, D., Adney, E.M., Jiang, H., Burns, K.H., Chait, B.T., et al. (2013). Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. Cell *155*, 1034–1048.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform *14*, 178–192.

Touchon, M., Nicolay, S., Audit, B., Brodie, E.-B.B.O., d'Aubenton-Carafa, Y., Arneodo, A., and Thermes, C. (2005). Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins. pp. 9836–9841.

Tubio, J.M.C., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas, C.P., Zamora, J., Raine, K., et al. (2014). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343–1251343.

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sánchez-Luque, F.J., Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M., et al. (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell *161*, 228–239.

Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *474*, 516–520.

Wagstaff, B.J., Hedges, D.J., Derbes, R.S., Campos Sanchez, R., Chiaromonte, F., Makova, K.D., and Roy-Engel, A.M. (2012). Rescuing Alu: Recovery of New Inserts Shows LINE-1 Preserves Alu Activity through A-Tail Expansion. PLoS Genet. *8*, e1002842.

Walter, M., Teissandier, A., Pérez-Palacios, R., and Bourc'his, D. (2016). An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. Elife *5*, R87.

Wang, G.P., Ciuffi, A., Leipzig, J., Berry, C.C., and Bushman, F.D. (2007). HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Res *17*, 1186–1194.

Wei, W., Morrish, T.A., Alisch, R.S., and Moran, J.V. (2000). A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. Anal Biochem *284*, 435–438.

Xie, Y., Mates, L., Ivics, Z., Izsvák, Z., Martin, S.L., and An, W. (2013). Cell division promotes efficient retrotransposition in a stable L1 reporter cell line. Mob DNA *4*, 10.

Zhang, Y., Romanish, M.T., and Mager, D.L. (2011). Distributions of transposable elements reveal hazardous zones in mammalian introns. PLoS Comput Biol *7*, e1002046.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zingler, N., Willhoeft, U., Brose, H.-P., Schoder, V., Jahns, T., Hanschmann, K.-M.O., Morrish, T.A., Löwer, J., and Schumann, G.G. (2005). Analysis of 5“ junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5-”end attachment requiring microhomology-mediated end-joining. Genome Res *15*, 780–789.

MAIN FIGURE TITLES AND LEGENDS

Figure 1 - *De novo* L1 integration events show the typical hallmarks of L1 retrotransposition.

(**A**) Experimental workflow. *De novo* L1 retrotransposition was induced by transfecting HeLa S3 cells with a plasmid-borne L1 element containing a neomycin-based retrotransposition marker (see Methods). Cells with new L1 insertions were selected by G418 (insertions referred as *L1 neo*) or recovered after only selecting for plasmid uptake by hygromycin (insertions referred as *L1 neo-unsel*). L1 3' junctions with host DNA were amplified and barcoded (bc) by a modified ATLAS-seq protocol, multiplexed, sequenced by Ion Torrent (400 bp single-end reads). Reads were mapped on the human reference genome (hg19).

(**B**) Genome browser view of ATLAS-seq read coverage (*top*) and alignment (*bottom*) on chromosome 7 supporting an antisense L1 insertion in the *ANKMY2* gene (vertical red lines). Mapped reads contain a soft-clipped region absent from the reference genome corresponding to the inserted L1 copy with its poly(dA) tail. The latter originates from the reverse transcription of L1 mRNA poly(A) tail and is a hallmark of L1 retrotranposition. Here, the putative endonuclease cleavage site (5'-TCTC/AA-3', black boxed) is followed by a poly(T) stretch, as expected for an antisense L1 insertion. ATLAS-seq reads are in opposite orientation relative to the inserted L1 element.

(**C**) Sequence logo representing the consensus motif detected at L1-neo pre-integration sites, consistent with a canonical L1 endonuclease mediated cleavage (5'-TTTT/A-3', cleaved strand) and with the subsequent annealing of the L1 mRNA poly(A) tail to an extended T-tract to promote reverse transcription priming.

(**D**) Distribution of motif scores based on the position weight matrix (PWM) shown in (C) for observed (L1 neo, blue) or random (red) datasets. Motif scores are binned by 0.5 and the curves represent Gaussian fits. The L1 target motif (as defined in panel C) is present at the vast majority of the mapped insertion sites in contrast to sequences found at random sites ($p<1e-16$, Wilcoxon test).

See also Figure S1.

Figure 2 - Preferential integration of *de novo* L1 insertions into sites with low nucleosome occupancy.

Heat maps and average profiles of nucleosome occupancy around pre-integration (L1 neo) or control sites (MMC, BMC and Random) as measured by MNase-seq in HeLa S3 cells [data from (Hansen et al., 2010)]. For each heat map independently, loci were sorted by increasing signal in the central 150 bp-window. Average nucleosome densities were calculated either globally (bottom graphs, blue) or by separating them in three classes of equal size, after ranking them by increasing levels of nucleosome density at the insertion or control sites (green graphs; low, middle and high nucleosome occupancy, from top to bottom). The high nucleosome occupancy class indicates that nucleosome presence *per se* does not prevent L1 integration. Global profiles of controls on the bottom represent the average of 50 datasets. MMC, motif-matched control; BMC, base-composition-matched control.

See also Figure S2.

Figure 3 - L1 integration sites are not uniformly dispersed in the human genome.

(**A**) L1 integration sites are dispersed in all chromosomes. Each dot represents a chromosome. The number of L1 integration sites is highly correlated (Pearson correlation) with HeLa S3 whole genome sequencing (WGS) coverage of each chromosomes (number of reads). Chromosome 1 is more frequently targeted than expected (red dot, two-tailed binomial test, FDR-adjusted p=0.010).

(**B**) Spacing between *de novo* L1 insertions. Empirical cumulative distribution of target site spacing (bp) was compared with those of *in silico* generated random datasets. The increased clustering of *de novo* L1 insertions (L1 neo) is evident at inter-insertion distances between $5x10^4$ and 10^6 bp, when compared to random positions (Random), base-composition- or motif-matched controls (BMC and MMC), or to random positions weighted by

copy-number in our HeLa S3 stock (WGS) (two-sample Kolmogorov-Smirnov test; these four curves are almost superimposed in the graph).

(**C**) L1 integration hotspots in 10Mb bins. From inside to outside tracks: **CNV** (brown), relative copy number variation calculated from low-pass WGS of the HeLa S3 stock used in this study (arbitrary units, modal n=3); **L1HS**, **L1PA**, and **L1** tracks (from dark to light blue), count of endogenous L1HS, L1PA or total L1 elements, respectively. The L1 family is the highest classification level analyzed here, and other subfamilies are nested in the following order: L1 > L1PA (primate-specific) > L1HS (human-specific). **L1 neo** (black), count of observed L1 insertion sites. The overlaid blue line represents the average number of *potential* L1 target sites (average count of motif-matched control sites per bin, 1565 MMC sampled 1000 times). Red bars indicate hotspots (Poisson distribution corrected for CNV, FDR-adjusted $p < 0.05$).

See also Figure S3 and Table S3.

Figure 4 - Novel L1 insertions are only modestly influenced by chromatin states compared to other transposable elements and their distribution is rapidly altered post-integration.

(**A**) Association of various transposable elements and retroviruses with ENCODE chromatin states. The 18 chromatin states were defined by ENCODE based on a combination of several histone marks (ChromHMM). Heat map displays the magnitudes of overlaps with each chromatin state, expressed as z-scores (see methods, blue for depletion and red for enrichment). Publicly available *de novo* insertion datasets previously obtained for other classes of elements were analyzed in parallel (Gogol-Döring et al., 2016; LaFave et al., 2014; Wang et al., 2007). Endogenous L1s (L1 endo) correspond to the 3' ends of endogenous L1HS-Ta elements present in the HeLa-S3 genome (Philippe et al., 2016). Chromatin segmentation data were obtained from ENCODE in HeLa S3 (for L1 and simulated datasets) or K562 cells (for SB, HIV, PB and MLV). BMC, MMC and Random are *in silico*-simulated control datasets (see Figure 2 legend). SB, Sleeping Beauty; PB, PiggyBac; MLV, Murine Leukemia Virus; HIV, Human Immunodeficiency Virus.

(**B**) Proportion of insertions falling in each chromatin segment for *de novo* insertions compared to endogenous L1HS-Ta copies and to a Random dataset.

(**C**) Association of various transposable elements and retroviruses with histone modifications. Heat map displays z-scores of observed overlaps with various histone ChIP-seq peak obtained by ENCODE in HeLa S3 (for L1 and simulated datasets) or K562 cells (for SB, HIV, PB and MLV). Color-scale and legends are as in (A).

(**D**) The tempo of primate-specific L1 association with a subset of functional chromatin segments. EnhWk, weak enhancer; TxWeak, weak transcription; Quies, quiescent chromatin segments. L1HS is equivalent to L1PA1. L1 are sorted by increasing age (Khan et al., 2006). Curves represent loess-smoothed profiles of z-scores across primate-specific L1 subfamilies. Solid and dashed red lines indicate the z-score values of L1 neo and endogenous L1 elements as a whole, respectively.

See also Figure S4, Figure S5, and Figure S6.

Figure 5 - *De novo* L1 insertions are not enriched in genes but their orientation toward genes is unbalanced.

(**A**) Proportion of various transposable elements and retroviruses inserted in genes. Values and error bars for simulated controls represent the average percentage of 1000 datasets ± s.d. The dashed line at 45% indicates the proportion of the human genome encompassed by annotated protein coding genes (GENCODE v19). A red bullet indicates a significant difference with this proportion ($p<0.001$), while a grey bullet indicates no significant difference ($p\geq0.05$) in a two-tailed binomial test with FDR correction (FDR<0.05).

(**B**) Expression levels of genes containing insertions. Each dot represents the median expression level of genes with an insertion in each of the 1000 simulated datasets, overlaid with their median and interquartile range

(red). On the right, the median expression levels of genes with experimental insertions are indicated. With the exception of endogenous L1HS-Ta (L1 endo), all tested elements are inserted in genes with significantly higher levels of expression than expected randomly (L1 neo *vs* Random, p=0.012; L1 neo *vs* BMC, p=0.032; all others p<0.002). However, the median expression level of genes with novel L1 insertions does not significantly differ from those containing L1 target motifs (L1 neo *vs* MMC, p=0.204). p-values are two-tailed estimates from empirical distributions. TPM, transcripts per million. Poly(A)$^+$ RNA-seq data in HeLa S3 were obtained from ENCODE.

(**C**) Orientation of L1 insertions relative to genes. Both *de novo* (L1 neo) and endogenous L1 insertions occur significantly more frequently in antisense direction with respect to gene transcription (two-tailed binomial test).

(**D**) and (**E**) Biased *de novo* L1 insertion orientation results from an asymmetric distribution of L1 target motifs in transcribed *vs* non-transcribed strand. A/S ratio, ratio of insertions in antisense *vs* sense orientation relative to genes. Each dot represents the mean A/S ratio in each of the 1000 simulated datasets, overlaid with their median and interquartile range (red). On the right, the A/S ratios for experimental datasets are indicated. Since Random, BMC, SB, PB and HIV datasets were unstranded, orientations were assigned randomly to each insertion and their average ratio is expected to be equal to 1. (E) A/S ratios comparisons between experimental and control datasets. p-values are two-tailed estimates from empirical A/S ratio distributions.

See Figure 4A for dataset definitions.

Figure 6 - Influence of host DNA replication on sites and orientation of L1 integration.

(**A**) Expected distribution (histogram) and observed (line) average signal at L1 pre-insertion loci for wavelet-smoothed repli-seq signal (Replication timing, left), origins of replication (SNS-seq peak height; ORI, middle), and replication fork directionality signal (measured by OK-seq; RFD, right). The expected distributions represent the distribution of the mean signal of 1565 randomly-selected genomic positions, repeated 900 times. Repli-seq signal ranges from 0 to 100, high values representing early replication; RFD signal ranges from -1 (fork always travelling right-to-left) to 1 (fork always travelling left-to-right); regions with RFD=0 have equal chance of having a fork in either direction. The observed RFD was calculated separately for (+, red) and (-, blue) strand insertions. Data are from (Besnard et al., 2012; Hansen et al., 2010; Petryk et al., 2016).

(**B**) Replication timing at L1 pre-integration sites. Plot represent the Log2 enrichment (mean observed/expected ± s.e.m.) of the repli-seq signal for each cell cycle fraction. Repli-seq data for HeLa S3 cells are from (Hansen et al., 2010).

(**C**) Orientation of L1 insertions relative to replication fork directionality (RFD). Depicted is the fraction of positive (red) or negative (blue) strand *de novo* L1 insertions (plain lines) or MMC sites (dashed lines) for genomic regions binned according to RFD. Positive strand insertions (*i.e.*, initial EN cut occurring at the bottom strand) are highly enriched when replication fork is moving leftward, and vice versa. The density of L1 target motifs (MMC) parallels this orientation bias (dashed lines). Curves were obtained by loess-smoothing. OK-seq data for HeLa cells are from (Petryk et al., 2016).

See also Figure S7.

Figure 7 - The distribution of chromatin states and replication timing at endogenous L1HS copies is shaped predominantly by selection, and not by L1-driven instruction.

(**A**) Cartoon illustrating the phylogeny of distinct classes of recent L1HS insertions that are detected in the genomes of 12 cell-lines, including HeLa S3 cells (Philippe et al., 2016). The repertoire of endogenous L1HS copies in each cell-line consists of a mixture of ancestral, fixed L1 insertions shared by all cell-lines, polymorphic

insertions present in several cell-lines, and private L1 insertions specific to a single cell-line (blue). The meta-genotype of all existing insertion sites comprises sites that are used in a particular cell-line (here illustrated for HeLa S3) and that consequently have an L1 copy present (black), as well as unused sites at which L1 is absent in that cell-line (white). Comparison of the genomic features associated with each subset allows the effects of selection to be distinguished from the effects of L1-driven alterations.

(**B**) The association of genomic features with endogenous L1HS copies can be driven by two distinct post-integrative mechanisms: instruction (left) or selection (right). The predicted outcome of each model is depicted. The grey zones indicate a hypothetical genomic feature associated with endogenous L1HS elements. In the instructive model, L1 insertions directly induce this genomic feature at the site of integration (rounded arrow). In this scenario, HeLa S3-private insertions (black arrowheads) are enriched in the genomic features present in HeLa S3, while sites of private insertions absent from HeLa S3 are not (white arrowheads). In the selective model, insertions - or the individuals that carry them - that occur outside the feature are eliminated (black crosses). Under this hypothesis, all private insertion sites have undergone equivalent selection, whether present or absent from HeLa S3, and so show a similar profile of association with genomic features (note that most private insertions likely occurred in the ancestral germline, and so would have since been subject to selection in the context of genomic features present in multiple different cell types).

(**C**) Association of L1HS subsets with distinct ENCODE chromatin segments (EnhWk, weak enhancer; TxWeak, weak transcription; Quies, quiescent chromatin segments) or replication timing in HeLa S3 cells. Dendrograms summarize the pairwise similarities of chromatin features associated with each set of insertion sites, and squares indicate the branchpoint of the expected distribution based on random genomic placement (z-score=0, *rand*). These associations are consistent with the selective model and not with the instructive model, described in (B). See (A) for dataset definitions. Z-scores were corrected for sample sizes.

STAR*METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents may be directed to, and will be fulfilled by, the corresponding author, Gael Cristofari (Gael.Cristofari@unice.fr).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

HeLa S3 cells were obtained from ECACC (distributed by Sigma-Aldrich) and maintained in a tissue culture incubator (37°C at a 5% CO2 level) in Dulbecco's modified Eagle medium (DMEM) containing 4.5 g/L D-Glucose, 110 mg/L Sodium Pyruvate, and supplemented with 10% FBS, 100 U/mL penicillin, and 100 µg/mL streptomycin. Growth medium was also supplemented with 862 mg/mL L-Alanyl-L-Glutamine (Glutamax). Cell cultures tested negative for mycoplasma infection using the MycoAlert Mycoplasma Detection Kit (Lonza). Cell line authenticity was verified by multiplex STR analysis (PowerPlex 21 PCR system, Promega, assays performed by Eurofins Genomics as a service provider) and comparison with the DSMZ database (https://www.dsmz.de/services/services-human-and-animal-cell-lines/online-str-analysis.html).

METHOD DETAILS

Plasmid constructs

The plasmid pOS24 consists in a full length and retrotransposition-competent human L1 (L1.3 clone, Genbank accession number L19088), containing the *mneoI* retrotransposition reporter cassette (Figure 1A and (Moran et al., 1996)). The engineered L1 was cloned into a modified pCEP4 backbone (Invitrogen) in which the CMV promoter and SV40 polyadenylation sequence were removed. Hence, L1 is expressed from its natural internal promoter (in the 5'UTR) and transcription is terminated by its natural polyadenylation signal. The *mneoI* cassette is expressed from an SV40 promoter and allows discriminating *de novo* copies from the thousands of endogenous L1 elements already present in the human genome. Neomycin (G418)-resistance was used to enrich cells containing new retrotransposition events (L1 neo). A hygromycin-resistance marker is also present in the plasmid backbone and has been used to select transfected cells in the absence of G418 selection (L1 neo-unsel).

Oligonucleotides

Oligonucleotides, described in Table S1, were synthesized by Integrated DNA Technologies (IDT, Coralville, IA).

L1 retrotransposition assays

The cultured cell retrotransposition assay was conducted as described previously (Wei et al., 2000). Briefly, 2×10^5 cells/well were plated in 6-well plates. The next day, cells were transfected with 1 µg of plasmid DNA and 3 µL of Lipofectamine 2000 (Life Technologies) diluted in 200 µL of Opti-MEM (Life Technologies). Medium was replaced with fresh medium after 5 hr. For retrotransposition assays in T75 and T175 flasks, 2×10^6 and 5×10^6 cells were plated, respectively, and the amount of each reagent was adapted accordingly to the plate surface. Two days post-transfection, medium was supplemented with G418 (Life Technologies) at 400 µg/mL to select for retrotransposition events. The media was changed daily. After 10 days of selection, surviving cells in one well per batch of retrotransposition assay was washed with 1X Phosphate-Buffered Saline (PBS), fixed, and stained with crystal violet to visualize colonies. If stained cells surviving the G418 selection were present in the well, cells were collected from other wells. Genomic DNA was extracted using a QiaAmp DNA Blood mini kit (Qiagen). In parallel, HeLa S3 cells were plated in 6-well plates and transfected with 0.5 µg of the same plasmid and phrGFP (Stratagene). Three days post-transfection, cells were subjected to flow cytometry and the transfection efficiency was determined based on the number of GFP-positive cells by FACS.

Library preparation and high-throughput sequencing

Mechanical fragmentation, end-repair, A-tailing, and adapter ligation. Libraries were prepared as described previously (Philippe et al., 2016). Briefly, 1µg of genomic DNA was sonicated for 6-12 cycles (5s on/90s off) at 4˚C with a Bioruptor sonicator (Diagenode), to reach an average fragment size of 1200 bp. DNA ends were repaired using the End-It DNA End-Repair Kit (Epicentre, Madison, WI). A-tailing of the repaired blunt ends was performed with Klenow Fragment (3'-to-5' exo-, New Englands Biolabs, Ipswich, MA) following the manufacturer's protocol. Adapter and dummy oligonucleotides were ligated to the A-tailed DNA. Between each of the above steps, DNA was purified with Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) using a 0.8:1 ratio of beads to DNA solution (v/v) and DNA was quality-controlled by Bioanalyzer 2100 (DNA high sensitivity kit, Agilent Technologies, Santa Clara, CA).

Library preparation by suppression PCR. Junctions of novel L1 3' end and genome were selectively enriched by suppression PCR. To reduce PCR stochasticity, the ligated genomic DNA of each sample was amplified in 8 independent parallel reactions of 40 µL each, containing 20 ng of ligated genomic

DNA under the following cycling conditions: 1 cycle at 95˚C for 4 min; followed by 30 cycles at 95˚C for 30 s, 68˚C for 30 s, and 72˚C for 1 min; and a final extension step at 72˚C for 7 min. Primers are described in Table S1. Each primer pair contains the trP1 and Ion A sequences, to be used for subsequent Ion Torrent library quantification and Ion Torrent sequencing. PCR products from the 8 reactions corresponding to the same population of clones were pooled. For some libraries prepared from unselected cells, linker-ligated genomic DNA was digested for 1h at 37°C with BamHI prior to suppression PCR (RD=restriction-digested, Figure S1). This step was intended to limit amplification from plasmid DNA, since this enzyme cuts downstream of L1 polyadenylation signal, but it did not consistently improve unselected insertion recovery (Figure S1).

Library preparation in emulsion. Some of the libraries were amplified by digital droplet PCR (ddPCR) in parallel to the PCR method described above to reduce stochasticity in PCR amplification and increase library complexity. We reasoned that massive partitioning of template DNA into 20,000 droplets would allow capturing late and rare retrotransposition events and also minimize over-amplification of insertions. ddPCR amplification was done using QX200 ddPCR EvaGreen supermix from Bio-Rad under the following cycling conditions: 1 cycle at 95˚C for 5 min; followed by 30 cycles at 95˚C for 30 s, 64˚C for 1 min; followed by signal stabilization of 1 cycle at 4˚C for 5 min, 1 cycle at 90˚C for 5 min (Bio-Rad's C1000 touch thermal cycle). For each sample, 9 parallel reactions were performed. Droplets of one reaction were read with a QX200TM droplet reader and analyzed with the QuantaSoft software to control amplification. Droplets from the remaining 8 reactions were pooled and amplified DNA was extracted from the droplets by the chloroform extraction method according to manufacturer's protocol.

Size selection. Pooled amplicons from either PCR method were subjected to double size selection to retain amplicons ranging between 300 and 450 bp by two consecutive Agencourt AMPure XP bead purification steps using beads-to-DNA ratios of 0.6:1 and 0.7:1, respectively. The supernatant of the first bead purification using bead:DNA ratio of 0.6:1 contains DNA fragments larger than 300 bp. This supernatant is applied to a second selection step with a bead:DNA ratio of 0.7:1 (*i.e.*, addition of 0.1X beads to the supernatant), where fragments smaller than 450 bp are bound to the beads and subsequently eluted. To eliminate any traces of primers, a last step of purification using beads to DNA ratio of 1:1 was performed.

Library quantification. Each library was quantified for copy number using a quantitative PCR-based assay (library quantification kit for Ion Torrent, Kappa Biosystems, Wilmington, MA). Average amplicon length was quantified by Bioanalyzer 2100 (DNA high sensitivity kit, Agilent Technologies, Santa Clara, CA). Library concentration was deduced from library average length and copy number.

Ion Torrent PGM sequencing. For sequencing, three to five libraries were pooled in equimolar amounts (final concentration of 20 pM). Emulsion PCR and enrichment for positive Ion Sphere

Particles (ISPs) was performed on the Ion OneTouch 2 and ES enrichment modules, respectively, using the Ion PGM Template OT2 400 Kit (Life Technologies), and sequenced on the Ion Torrent PGM, using the Ion PGM Sequencing 400 Kit and Ion 318 v2 Chips (Life Technologies), according to the manufacturer's protocols.

Integration site mapping

Ion Torrent sequencing reads were processed and mapped to the reference human genome (hg19) in order to locate *de novo* L1 insertion sites, using a modified ATLAS-seq pipeline (Philippe et al., 2016), summarized below.

FASTQ files were demultiplexed according to the sample-specific barcode using cutadapt (https://github.com/marcelm/cutadapt). Reads from each barcoded library were then trimmed using cutadapt to remove barcodes, ATLAS-seq primers, and adapters. Trimmed reads were mapped to the hg19 human reference genome using the Burrows-Wheeler Aligner (BWA) program with the 'mem' algorithm allowing soft-clipping (Li and Durbin, 2010). Mapped reads were filtered to remove secondary alignment and ambiguously mapped reads (MAPQ<20) using SAMtools (Li et al., 2009). Soft-clipped reads with a polyA or polyT at the junction were recovered and insertion sites were called, based on the soft-clipped position of each read. Within a sample, PCR duplicate reads were removed with Picard tools (http://broadinstitute.github.io/picard, MarkDuplicates function), keeping only the longest representative read. Reads were considered redundant if they started from the same linker position, which corresponds to the initial genomic DNA break during sonication. Identical insertion sites from independent pool of cells which were sequenced together were merged into clusters using BEDtools (Quinlan and Hall, 2010). In a given run, an insertion falling at the same nucleotide position as - or less than 500 bp - from another insertion but supported by less non-redundant reads was considered as a mapping artefact or index hoping and were eliminated. Thus, recurrent insertions at the same nucleotide were only considered when found in two independent samples sequenced in two distinct runs. Finally, insertions falling in ENCODE blacklisted regions of hg19 genome assembly (DAC Blacklisted Regions and Duke Excluded Regions, downloaded from UCSC Genome Brower Tables) were removed. Altogether, we obtained 1565 target loci and 1647 insertions (includes recurrent insertions at the same target locus), from 28 samples sequenced in 8 independent Ion Torrent runs on 318 Chips (Figure S1 and Table S2). Note that each insertion is defined by a 2-bp interval spanning the integration point (0-based coordinates in hg19).

Motif analysis at the sites of integration

The genomic DNA sequence flanking the integration site was extracted with BEDtools and we used weblogo 3.5.0 to generate the motif at the insertion site between positions -3 and +8 (Figure 1 and Figure S1). The corresponding position-weighted matrix (PWM) was calculated and subsequently used to infer the score of each individual recovered integration site. The motif score is defined as the sum of the log-odds probabilities of the nucleotides found at each position, with an equiprobable frequency of each nucleotide.

Generation of controls datasets

We generated three distinct control datasets:

- a random control (Random): this dataset was obtained by randomizing 1000 times the experimental dataset. Orientation was picked arbitrarily.
- a base composition-matched control (BMC): this dataset was generated by randomly shuffling 1000 times the experimental dataset as for the 'Random' dataset but choosing insertion sites so that the distribution of base composition (%AT) around the integration sites (± 5bp) of all randomized insertions sites is identical to the distribution of the experimental dataset. Orientation was picked arbitrarily.
- a motif-matched control (MMC): this dataset was created by picking up randomly 1000x1565 genomic locations with a motif matching the L1 motif position-weighted matrix. A site was considered as matching if its score was above a threshold of 4.96256, which corresponds to 1 hit per kb assuming equal probability of each base, as estimated by the TFMpvalue package. Motifs are asymmetric, and orientation was maintained.

The BMC and MMC datasets were generated to test whether associations observed between some genomic features and L1 insertions indirectly results from a biased base- or motif-composition of these genomic features. Of note, all three types of control datasets were filtered as the experimental datasets. In other words, we excluded DAC blacklisted and Duke excluded regions, as well as unassembled contigs, gaps, alternate assemblies, and the Y chromosome.

Processing of published transposable element and retrovirus insertion datasets

All datasets were reformatted as 2 bp-interval .bed files and their genomic coordinates were converted by the UCSC LiftOver tool to hg19, when needed. As for experimental and control datasets, we excluded DAC blacklisted and Duke excluded regions, as well as unassembled contigs, gaps, alternate assemblies, and the Y chromosome. Sleeping beauty (SB) and PiggyBac (PB) datasets were published in (Gogol-Döring et al., 2016) and downloaded from GEO (GSE58744). We only kept libraries prepared from sonicated DNA. The available data were unstranded, but we arbitrarily

assigned a positive orientation to comply with BED6 format specifications. Murine Leukemia Virus (MLV) datasets were published in (LaFave et al., 2014) and were downloaded from the NHRGI website (https://research.nhgri.nih.gov/software/GeIST/download.shtml). Human Immunodeficiency Virus (HIV) datasets were published in (Wang et al., 2007) and were downloaded from the author website (http://microb230.med.upenn.edu/ucsc/hiv.wig.bed). The data were unstranded, but we arbitrarily assigned a positive orientation to comply with BED6 format specifications. Endogenous L1HS-Ta elements present in HeLaS3 cells, and called throughout this work 'L1 endo', were obtained from (Philippe et al., 2016).

Nucleosome density

We used publicly available micrococcal-nuclease (MNase-seq) data generated in HeLa S3 to calculate nucleosome occupancy at insertion sites (GSM1053817) (Lacoste et al., 2014) (Figure 2) and confirmed the obtained results with publicly available HeLa and Raji datasets (Descostes et al., 2014; Schwartz et al., 2018). Briefly, reads were mapped to the hg19 human reference genome with bowtie2 (v2.1.0) (Langmead and Salzberg, 2012), aligned reads were compressed in bam format using samtools (Li et al., 2009), and signal pile up was obtained with the PASHA R package (Fenouil et al., 2016). Data were processed using the paired-end approach, binned by 10 bp. Pile-up (PCR) artefacts were removed when more than 1 read aligned exactly at the same position every 7 millions of aligned reads as previously described (Fenouil et al., 2016). Data were then scaled using the number of reads and the signal (normAndSubstractWIG). To generate heat maps, each locus representing ±2.5 kb surrounding the integration site was divided into 1000 equal bins and the signal of each bin was transformed by the asinh function. Loci were sorted by increasing MNase-seq signal in the central ±75bp window surrounding the integration sites, and visualized with Java Treeview (Saldanha, 2004). Metaprofiles were generated following the same approach, but without the asinh normalization, either for the whole set of insertions, or after dividing the ordered sets of the heat maps in 3 equal groups. In Figure 2, heat maps and subgroup profiles were prepared with a single dataset, but global profiles for the control datasets represent the average of 50 randomization.

H3K4me1 and whole genome sequencing

Chromatin extraction, chromatin immunoprecipitation (ChIP) DNA purification and sequencing were performed using standard techniques as previously described (Ramos Pittol et al., 2018). Briefly, cells were fixed with 1% formaldehyde for 15 minutes at room temperature and washed extensively in ice-cold PBS. Nuclei were released by incubation for 5 minutes in ice-cold L1 buffer (50 mM Tris pH 8, 2 mM EDTA, 0.1% NP40, 10% glycerol) followed by 5 minutes centrifugation at 1000 xg, and lysed in

L2 buffer (50 mM Tris pH 8, 5 mM EDTA, 1% SDS). Chromatin in the supernatant was fragmented to a size range of approximately 300-700 bp using a tip sonicator, and insoluble debris was removed by centrifugation. A sample of this chromatin was used as an input control for ChIP, and for whole-genome sequencing. Chromatin was diluted 10-fold in DB buffer (50 mM Tris pH 8, 5 mM EDTA, 200 mM NaCl, 0.5% NP40), pre-cleared with protein-A sepharose for 1 hour, and incubated with anti-H3K4me1 antibodies (Abcam ab8895) at a concentration of 2 µg/mL overnight at 4°C. Chromatin was immunoprecipitated for 30 minutes using 10 µL/ml protein-A sepharose, and sequentially washed 6x with ice-cold NaCl wash buffer (20 mM Tris pH 8, 2mM EDTA, 500 mM NaCl, 1% NP40, 0.1% SDS) followed by 3x with ice-cold TE (50 mM Tris pH 8, 2 mM EDTA). Immunoprecipitated chromatin was released by incubation at room temperature in buffer EB (50 mM Tris pH8, 2 mM EDTA, 2% SDS), and cross-links were reversed by overnight incubation at 65°C. ChIP and input DNA were both purified using Qiagen MinElute PCR purification kits, and DNA was re-fragmented for sequencing to achieve a mean size of 300 bp using a water-bath sonicator. Input and ChIP DNA were subjected to Illumina NextSeq 500 paired-end sequencing. Sequencing reads were aligned to the human reference genome (hg19) using bowtie with options -v 2 -a -m 5 --maxins 2000 --tryhard (Langmead and Salzberg, 2012), and statistically excess reads mapping to the exact same fragment (representing likely PCR artefacts) were removed. For analysis of H3K4me1 ChIP-seq, enriched peaks were identified using MACS 1.4 with options -p 1e-4 --nomodel --shiftsize=150 --keep-dup=all (Zhang et al., 2008), using input DNA as background.

Chromosomal distribution

To assess chromosomal distribution of L1 insertions, all data were corrected for copy number variation, genome gaps, and mapping artefacts by normalization with HeLa S3 whole genome sequencing coverage, either for entire chromosomes or for selected bins. For hotspot detection, we calculated for each bin an expected cumulated Poisson probability to obtain the observed number of insertions or more, based on the observed average insertion rate, and corrected this probability by: (i) the WGS coverage value; and (ii) the L1 target motif density. Correction for multiple testing was ensured using an FDR<0.05. The coverage correction was done by scaling the expected number of insertions, *i.e.* the Poisson process mean, by the relative WGS coverage of each bin relative to the average WGS coverage. The L1 target motif density was approximated by computing the average per bin density of the 1000 MMC control datasets (see above), and then was used as the mean of the Poisson process instead of the average insertion rate in correction. Hotspots were detected with only the correction for coverage and with both corrections. Circular plots were generated using the circlize R packageCircular plots were generated using the circlize R package (Gu et al., 2014).

Enrichment of genomic features

To allow a fair comparison of the associations of integration profiles (both actual and simulated) with a wide range of genomic features, we used a statistical approach in which we generate a large number of controlled *in silico* randomizations of each dataset, and we express the magnitudes of each association as a z-score, which reflects the number of standard deviations by which the measured similarity of any pair of datasets differs from the similarity expected by chance. This approach does not require any *a priori* assumptions about the abundance, distribution or resolution of tested genomic features, and accounts for feature clustering at all scales. Importantly, z-scores for pairwise associations between unrelated datasets can be directly compared to each other, and indicate their relative levels of similarity.

To identify genomic features that are significantly associated with pre-integration sites, we compared the observed magnitude of overlaps between the dataset of L1 integrations and datasets describing each genomic feature to the expected distribution of magnitudes of overlap between the pair of datasets according to a null hypothesis of no relationship between them. This was empirically estimated by calculating the magnitudes of overlaps between a series of randomized versions of each dataset. Randomized datasets were constructed by randomly permuting the genomic positions of every feature, in such a way that that their scores, sizes, relative distributions and inter-regional structure at all scales is preserved. This strategy ensures that any structure or clustering in the distribution of genomic features (either observed or hidden) does not bias the estimate of association between datasets. Each dataset in a compared pair was generally permuted 30 times, providing 30x30=900 pairwise comparisons to calculate the expected random distribution of overlaps. The significance of each overlap is expressed as a Z-score, calculated as the number of standard deviations by which the observed overlap between datasets differs from the overlap level expected by chance. The Z-score calculated in this way is not biased by the absolute magnitude of observed overlaps, nor by the size of either dataset (either the number of features or the level of genomic coverage, and it is thereby suitable for comparing the significance of associations between genomic datasets with widely varying sizes and resolutions. Control datasets (random, BMC, MMC) were generated 100 times independently, and the median Z-scores were computed. Published integration datasets (SB, PB, HIV, MLV) were randomly subsampled to 1565 sites 100 times independently and median Z-scores were computed for the subsamples, to allow straightforward comparison of Z-scores to the L1 insertions sites without requiring correction for different sampling levels. For datasets with smaller sampling levels (unselected L1 insertions [Figure S4B] and absent/present private endogenous insertions [Figure 7]) Z-scores were corrected for subsampling by scaling by the square root of the sample size.

Relation between gene expression and integration sites

Gene expression levels in HeLa S3 cells were obtained from ENCODE (HeLa S3 polyA+ RNA-seq, ENCFF622JEX) as transcript per millions (TPM), using all known protein-coding genes annotated in GENCODE Genes v19. Proportion of genes in the genome, count of genic insertions and of each orientation was computed with BEDTools (Quinlan and Hall, 2010) after subtracting DAC blacklisted and Duke excluded regions, as well as unassembled contigs, gaps, alternate assemblies, and the Y chromosome.

Relation between replication and integration sites

Replication timing at L1 insertion sites, and in control datasets, was analyzed by calculating the enrichment in the repli-seq signal at each location compared to the genome-wide average signal, for each cell-cycle fraction. To correct for the elevated abundance of already-replicated DNA during S-phase (Figure S7), we multiplied the enrichments of control datasets by the calculated mean amount of DNA at each locus (based on their repli-seq profiles), assuming that each of the 6 repli-seq cell-cycle fractions encompasses an equal proportion of S-phase. Thus, the relative abundance of each locus is calculated as $1+(^{11}/_{12})f_1+(^{9}/_{12})f_2+(^{7}/_{12})f_3+(^{5}/_{12})f_4+(^{3}/_{12})f_5+(^{1}/_{12})f_6$, where f_1 to f_6 denote the proportion of the normalized repli-seq signal at each locus in each of the 6 cell-cycle fractions (so, for example, the mean abundance of hypothetical locus which replicates exclusively in the earliest sorted cell-cycle fraction is calculated as $1+^{11}/_{12}$).
Replication fork directionality was analyzed by calculating the ratio of forward- to reverse-strand insertions, or controls, that occur at genomic sites with a RFD score from OK-seq in each of 20 equal bins spanning the range from -1 to +1.

Testing selection vs instruction models

Private (present in only one cell line; n=744) and fixed (present in all analyzed cell lines; n=336) insertions were identified among 1633 L1HS-Ta mapped and annotated insertions from 12 different cell lines (Philippe et al., 2016), and private insertions were separated into those that are present (n=73) or absent (n=671) in HeLa S3 cells. The Z-score corresponding to the enrichment or depletion of each set of insertions at selected genomic features was determined as described above, with correction for the size of each sample. Pairwise Z-scores, indicating the measure of dissimilarity between sets of insertions, were calculated for each genomic feature by permuting the combined elements of each pair of sets 1000 times, and were used for hierarchical clustering.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests were performed in R and are explicitly stated in each Figure legend.

DATA AND SOFTWARE AVAILABILITY

ATLAS-seq data were submitted to the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-6933 (L1 neo insertions), E-MTAB-7644 (L1 neo-unsel insertions) and E-MTAB-7643 (H3K4me1 ChIP-seq). The genomic locations of *de novo* L1 insertions are provided in Table S2. The scripts written to call L1 insertions from ATLAS-seq data and to generate the control datasets, are available at https://github.com/retrogenomics/iss, as well as useful annotation files used in the course of this study.

SUPPLEMENTAL INFORMATION

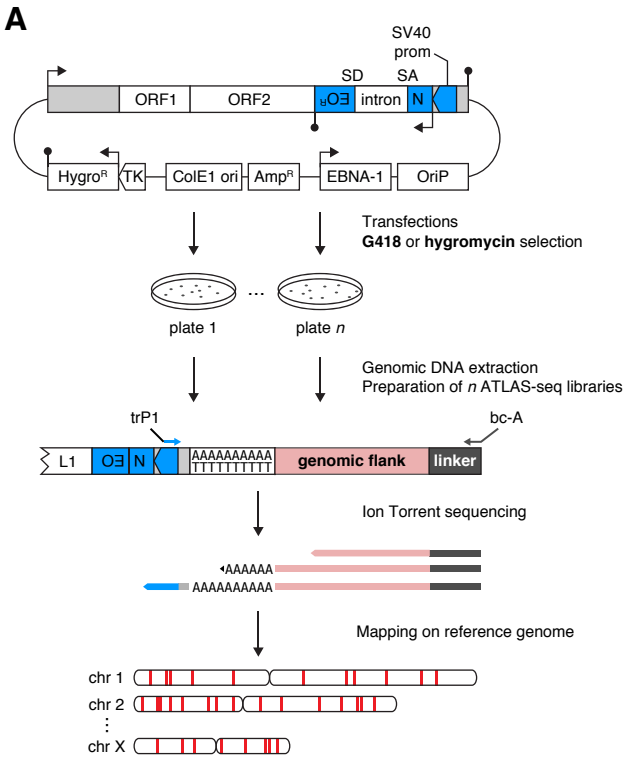
Supplemental Information includes 7 figures and 3 tables and can be found with this article online.

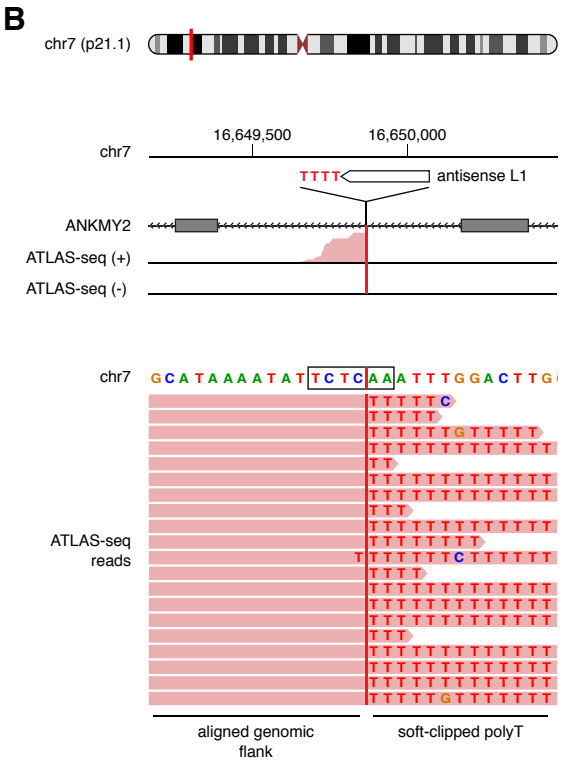
SUPPLEMENTAL DATA

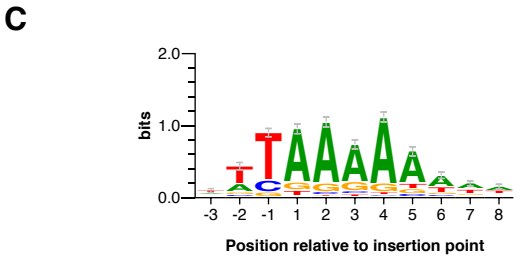
Table S1 - DNA oligonucleotides used in this study. Related to the STAR Methods section.

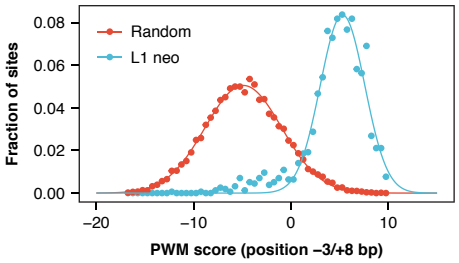
Table S2 - Genomic coordinates and characteristics of all L1 insertions recovered. Related to Figure 1.

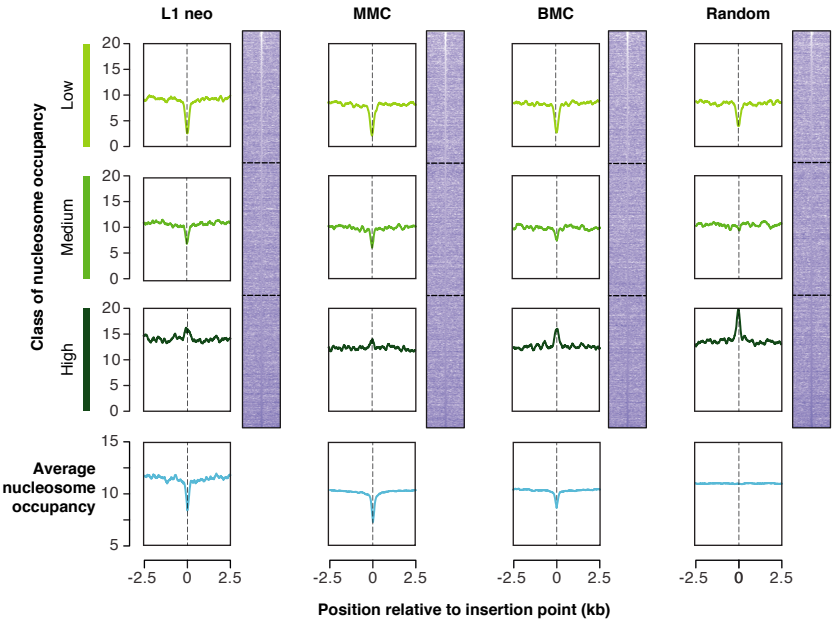
Table S3 - Hotspots of L1 insertions. Related to Figure 3.

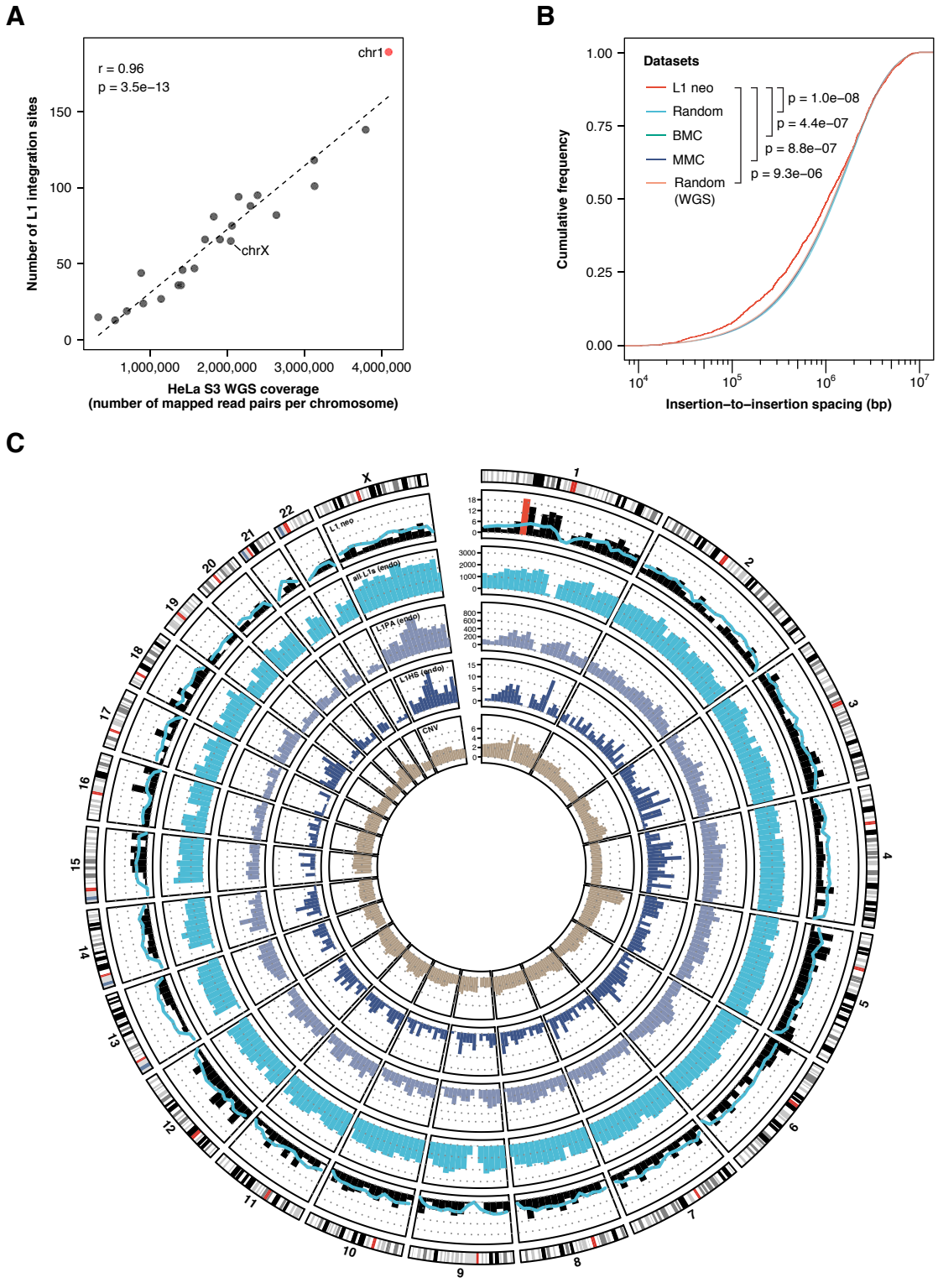

A
SV40
prom
SD
SA
ORF1
ORF2
intron
N
Hygro^R
TK
ColE1 ori
Amp^R
EBNA-1
OriP
Transfections
G418 or hygromycin selection
plate 1
plate n
Genomic DNA extraction
Preparation of n ATLAS-seq libraries
trP1
bc-A
L1
N
genomic flank
linker
Ion Torrent sequencing
Mapping on reference genome
chr 1
chr 2
chr X

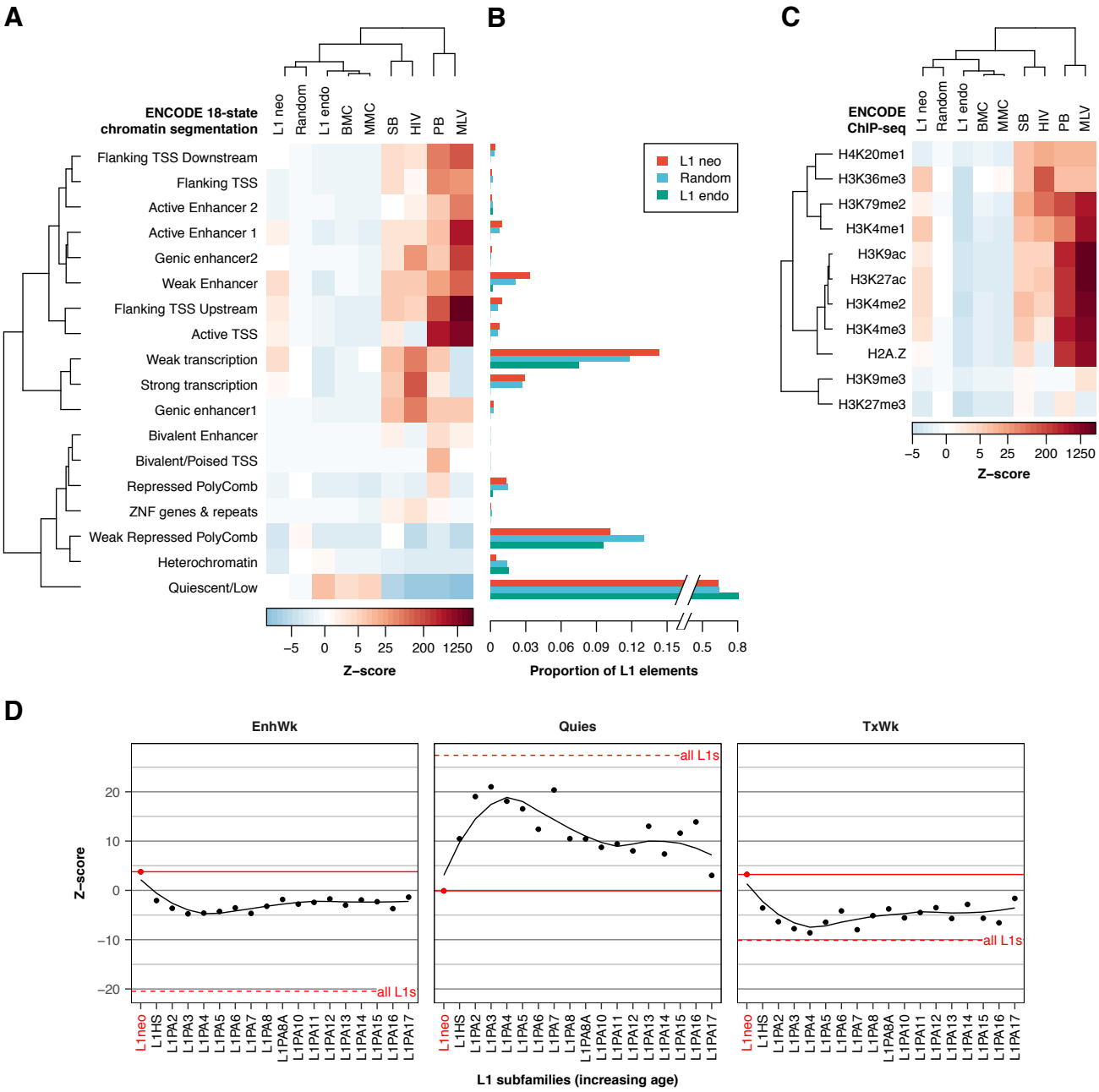

B
chr7 (p21.1)
chr7
16,649,500
16,650,000
TTTT
antisense L1
ANKMY2
ATLAS-seq (+)
ATLAS-seq (-)
chr7
ATLAS-seq
reads
aligned genomic
flank
soft-clipped polyT


C
bits
2.0
1.0
0.0
-3 -2 -1 1 2 3 4 5 6 7 8
Position relative to insertion point

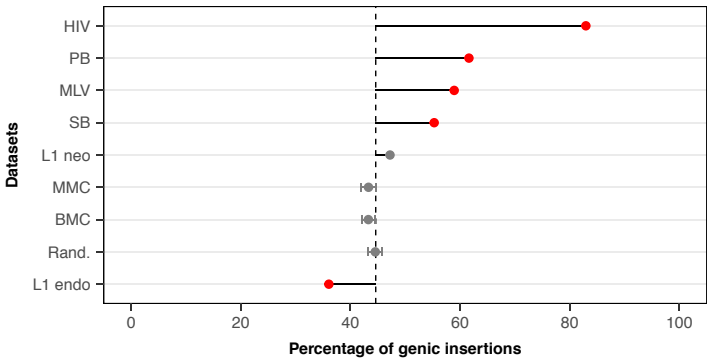

D
Random
L1 neo
Fraction of sites
0.08
0.06
0.04
0.02
0.00
−20
−10
0
10
PWM score (position −3/+8 bp)


L1 neo
MMC
BMC
Random
Class of nucleosome occupancy
Low
Medium
High
20
15
10
5
0
Average nucleosome occupancy
15
10
5
-2.5
0
2.5
Position relative to insertion point (kb)

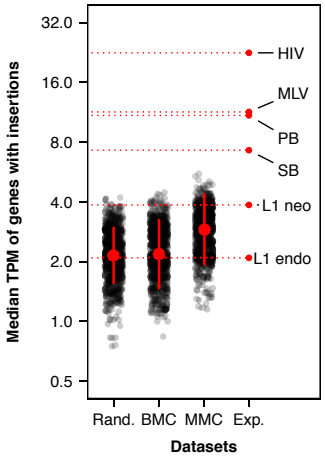

A
chr1
r = 0.96
p = 3.5e−13
chrX
Number of L1 integration sites
0
50
100
150
1,000,000
2,000,000
3,000,000
4,000,000
HeLa S3 WGS coverage
(number of mapped read pairs per chromosome)
B
Datasets
L1 neo
Random
BMC
MMC
Random
(WGS)
p = 1.0e−08
p = 4.4e−07
p = 8.8e−07
p = 9.3e−06
Cumulative frequency
0.00
0.25
0.50
0.75
1.00
Insertion−to−insertion spacing (bp)
C
L1 neo
all L1s (endo)
L1PA (endo)
L1HS (endo)
CNV


A
ENCODE 18-state chromatin segmentation
L1 neo
Random
L1 endo
BMC
MMC
SB
HIV
PB
MLV
Flanking TSS Downstream
Flanking TSS
Active Enhancer 2
Active Enhancer 1
Genic enhancer2
Weak Enhancer
Flanking TSS Upstream
Active TSS
Weak transcription
Strong transcription
Genic enhancer1
Bivalent Enhancer
Bivalent/Poised TSS
Repressed PolyComb
ZNF genes & repeats
Weak Repressed PolyComb
Heterochromatin
Quiescent/Low
−5 0 5 25 200 1250
Z−score
B
L1 neo
Random
L1 endo
0 0.03 0.06 0.09 0.12 0.15 0.5 0.8
Proportion of L1 elements
C
ENCODE ChIP-seq
H4K20me1
H3K36me3
H3K79me2
H3K4me1
H3K9ac
H3K27ac
H3K4me2
H3K4me3
H2A.Z
H3K9me3
H3K27me3
−5 0 5 25 200 1250
Z−score
D
EnhWk
Quies
TxWk
Z-score
20
10
0
−10
−20
all L1s
L1neo
L1HS
L1PA2
L1PA3
L1PA4
L1PA5
L1PA6
L1PA7
L1PA8
L1PA8A
L1PA10
L1PA11
L1PA12
L1PA13
L1PA14
L1PA15
L1PA16
L1PA17
L1 subfamilies (increasing age)

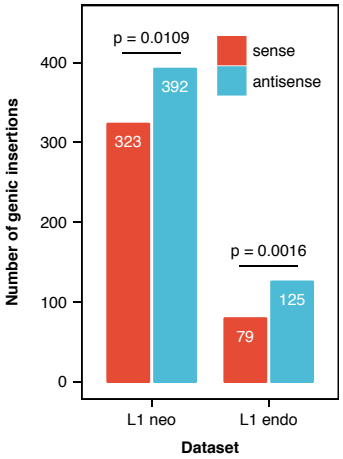
A



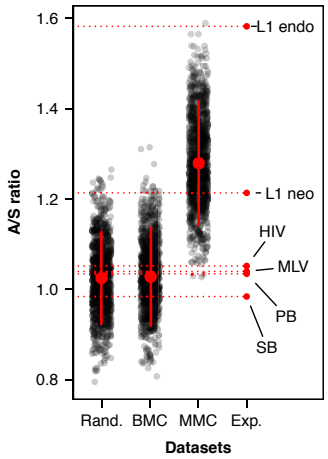
B



C

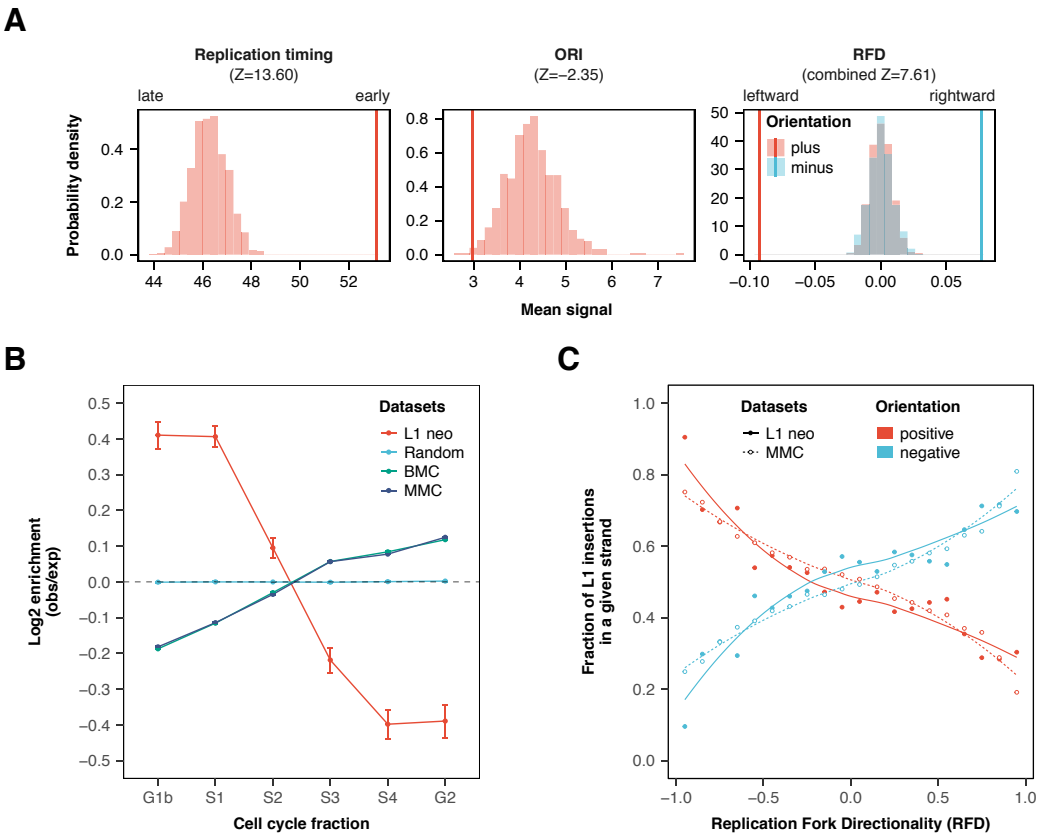


D

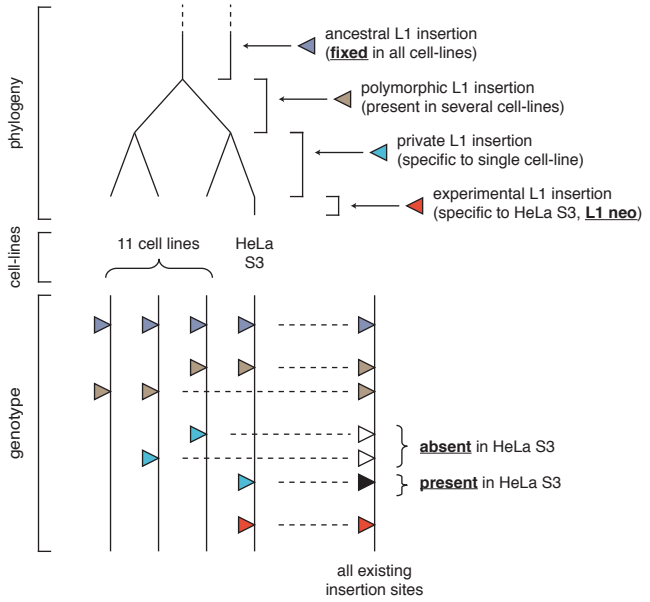


E

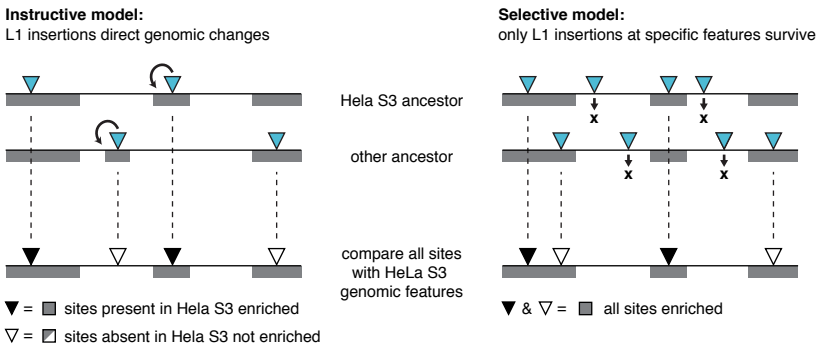
Datasets	A/S ratio	Rand.	BMC	MMC
L1 neo	1.21	0.024	0.028	n.s.
L1 endo	1.58	<0.001	<0.001	0.002
PB	1.03	n.s.	n.s.	n.s.
SB	0.98	n.s.	n.s.	n.s.
MLV	1.04	n.s.	n.s.	n.s.
HIV	1.05	n.s.	n.s.	n.s.


A
Replication timing
(Z=13.60)
late
early
Probability density
ORI
(Z=−2.35)
RFD
(combined Z=7.61)
leftward
rightward
Orientation
plus
minus
Mean signal
B
Datasets
L1 neo
Random
BMC
MMC
Log2 enrichment
(obs/exp)
G1b
S1
S2
S3
S4
G2
Cell cycle fraction
C
Datasets
L1 neo
MMC
Orientation
positive
negative
Fraction of L1 insertions
in a given strand
Replication Fork Directionality (RFD)

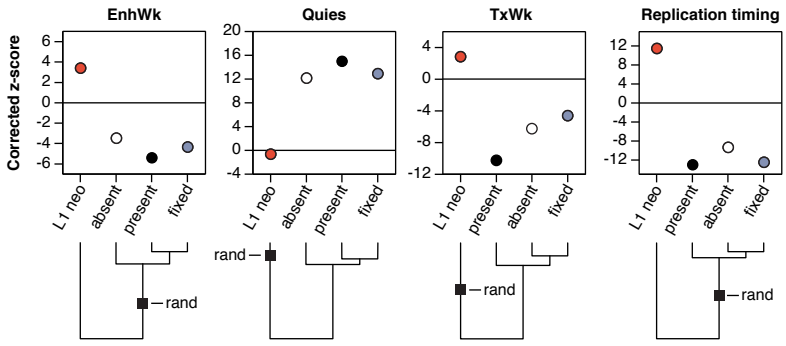
A


phylogeny
cell-lines
genotype
ancestral L1 insertion
(fixed in all cell-lines)
polymorphic L1 insertion
(present in several cell-lines)
private L1 insertion
(specific to single cell-line)
experimental L1 insertion
(specific to HeLa S3, L1 neo)
11 cell lines
HeLa S3
absent in HeLa S3
present in HeLa S3
all existing insertion sites

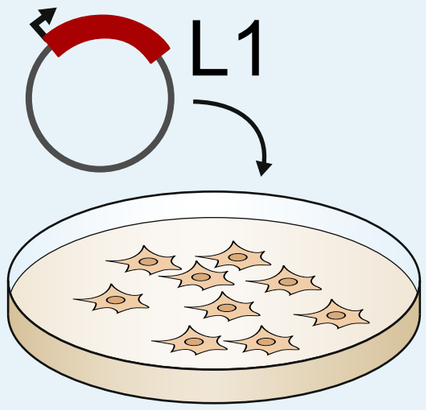
B

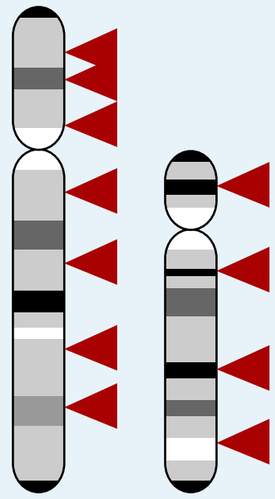

Instructive model:
L1 insertions direct genomic changes
Selective model:
only L1 insertions at specific features survive
Hela S3 ancestor
other ancestor
compare all sites with HeLa S3 genomic features
▼ = sites present in Hela S3 enriched
▽ = sites absent in Hela S3 not enriched
▼ & ▽ = all sites enriched

C


EnhWk
Quies
TxWk
Replication timing
Corrected z-score
L1 neo
absent
present
fixed
rand

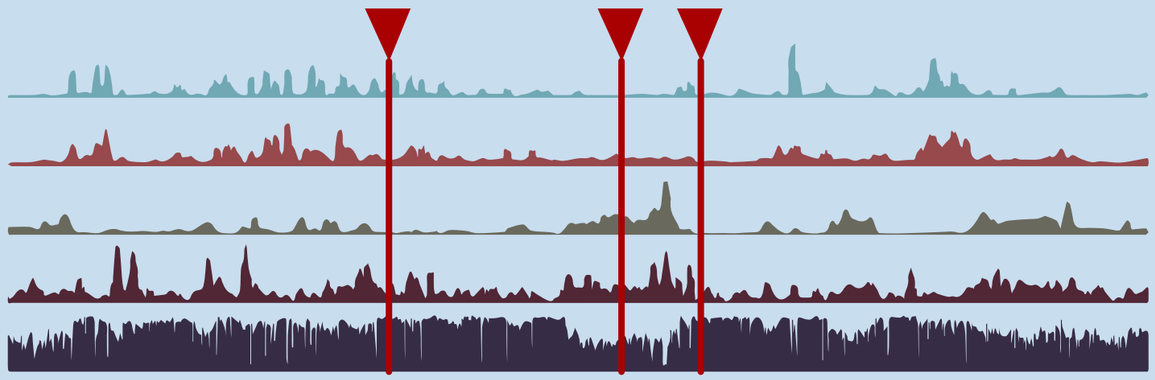
Genome-wide L1 retrotransposon insertion profiling



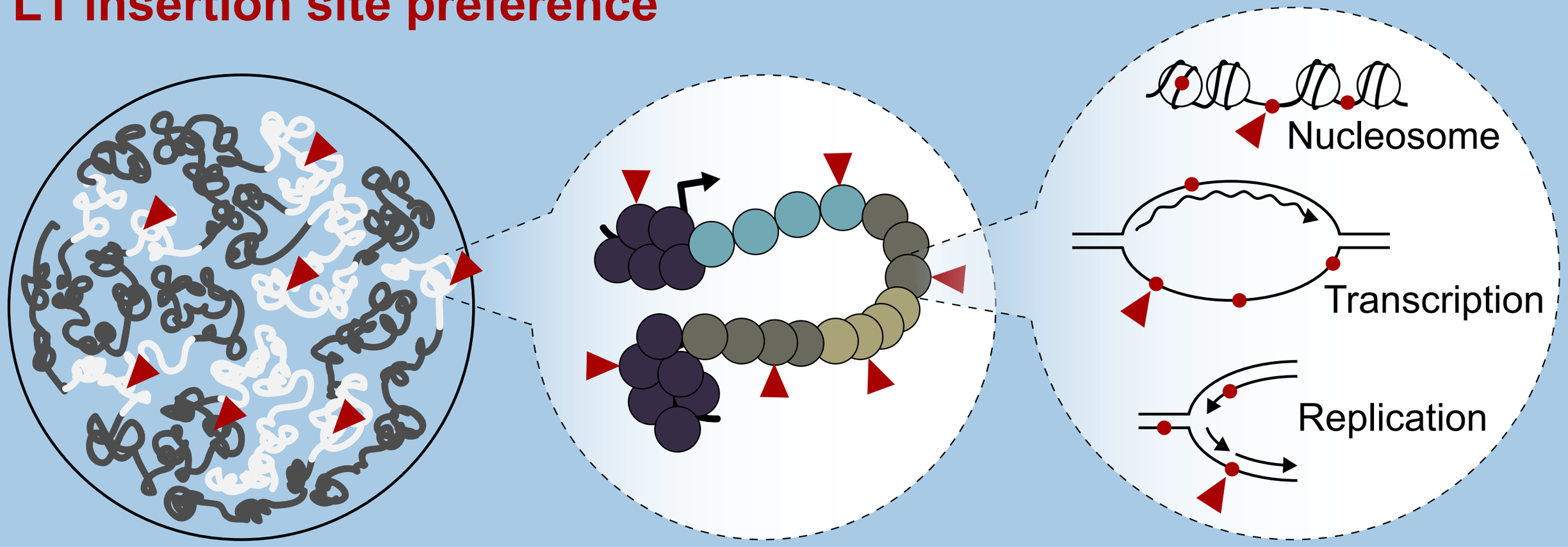


human reference genome

Integration with public data



L1 insertion site preference



Mb **kb** **bp**

Replication timing

All chromatin states

Strand-specific L1 target motif (•)