G-quadruplex structures within the 3′ UTR of LINE-1 elements stimulate retrotransposition

Aleksandr B Sahakyan[1,2,4], Pierre Murat[1,2,4], Clemens Mayer[1,2] & Shankar Balasubramanian[1–3]

Long interspersed nuclear elements (LINEs) are ubiquitous transposable elements in higher eukaryotes that have a significant role in shaping genomes, owing to their abundance. Here we report that guanine-rich sequences in the 3′ untranslated regions (UTRs) of hominoid-specific LINE-1 elements are coupled with retrotransposon speciation and contribute to retrotransposition through the formation of G-quadruplex (G4) structures. We demonstrate that stabilization of the G4 motif of a human-specific LINE-1 element by small-molecule ligands stimulates retrotransposition.

Transposable genetic elements (TEs) are substantial components of eukaryotic and prokaryotic genomes[1]. In humans, for example, ~45% of the genome consists of TEs, and, owing to their contribution to genomic instability through *de novo* insertion and recombination events, TEs are responsible for a number of genetic disorders and cancers[2,3]. LINE-1 (L1) elements are non–long terminal repeat (non-LTR) retrotransposons that duplicate through a reverse-transcribed RNA intermediate and integrate at new genomic loci. Active, full-length L1 elements—the only autonomous non-LTR retrotransposons in primate genomes—are ~6 kb long and contain a 5′ UTR with an RNA polymerase II promoter, two open reading frames (ORF1 and ORF2, encoding an RNA-binding protein and a reverse transcriptase with endonuclease activity, respectively), and a 3′ UTR that harbors a guanine-rich sequence together with a polyadenylation signal and terminates with an adenine-rich tail[4] (**Fig. 1a**). Unlike LINE elements in plants[5] and in some metazoans[6,7], in which conserved stem-loop secondary structures at the 3′ tail control *de novo* insertions by recruiting the ORF2-encoded protein[8], mammalian 3′ UTRs of L1 elements are thought to lack such *cis*-regulating elements. Indeed, early studies could not confirm a functional role of the 3′ UTRs of the L1 elements found in human and primate genomes[9]. Although the 3′ UTRs of primate L1 elements are devoid of canonical secondary structures, they contain conserved guanine-rich sequences (**Fig. 1a**) with the ability to fold *in vitro* into G4 structures[10,11]. G4s are noncanonical secondary structures formed by guanine-rich nucleic acids and stabilized by the stacking of guanine tetrads held together by Hoogsteen base pairing[12]. Recent works showed that putative G4 sequences are present and conserved in specific parts of TEs in plants and humans[13,14]. Interestingly, young L1 remnants were found to be enriched in detectable G4-forming sequences[11]. On the basis of these observations, we hypothesized a functional role for G4-forming sequences within L1 elements and carried out a study to evaluate the contribution of the L1 3′ UTR to retrotransposition activity.

RESULTS

G-rich sequences are a hallmark of young L1 retrotransposons

To elucidate a potential functional role of the conserved G4 motif in L1 evolution, we first devised a computational approach to identify the G4 sequences that stem from L1 elements (LQSs; Online Methods). We recovered all sequences from the human genome that matched the definition $G_{3+}N_{1-12}G_{3+}N_{1-12}G_{3+}N_{1-12}G_{3+}$ (where N is any base including G), which is characteristic of potential quadruplex sequences (PQSs; 703,091 sequences recovered)[15]. By counting the frequencies of distinct PQSs, we identified the most abundant LQS, with 2,503 occurrences, and used it as a reference sequence for identifying all members of the LQS family with shared origin (Online Methods, **Supplementary Note 1**, **Supplementary Table 1** and **Supplementary Fig. 1a–c**). In total, the analysis identified 3,228 unique LQSs in 15,724 genomic locations, with all but 5 present in the 3′ UTRs of L1 remnants (**Supplementary Fig. 1d**). This highly preferential colocalization of the identified sequences with L1 elements demonstrates that our computational approach is robust for identifying members of the LQS family.

Having assigned each LQS to its corresponding L1 subfamily (**Supplementary Table 2**), we then sought to identify original G4 sequences present in the respective viable states. To identify the L1 subfamilies for which time-accumulated random substitutions left enough sequence preference to allow inference of the original LQS, we introduced a G4 diversity index (DI^G4), defined as the ratio of unique PQSs to the total number of PQSs found in a given L1 subfamily. A DI^G4 value of 1 means that all PQSs found in a given L1 subfamily are different, and thus are affected by different substitutions. Smaller values indicate lower diversity, at which the original PQSs can still be differentiated by their abundance. DI^G4 values increase with L1 subfamily age, which demonstrates the erosive effect of time-accumulated random substitutions (**Fig. 1b**). Conversely, for the youngest lineage of L1 elements, specific to hominoid apes and

Figure 1 G-rich sequences are a hallmark of young L1 retrotransposons. (**a**) Sequence organization of the L1 element, showing the position of the G-rich sequence. (**b**) Change in the DI^{G4} index as a function of the divergence age of the source L1 subfamilies. The gray shaded area highlights the region where the DI^{G4} index is low—and thus where the time-accumulated substitutions have not interfered too much with sequences of the corresponding age—and where the original G-rich sequence can be robustly identified. (**c**) The most frequently occurring G4 sequences found within the 3′ UTR of primate L1 elements and the associated number of occurrences in the human genome (in brackets), together with a multiple sequence alignment plot highlighting the core nucleotide substitutions. Source data for **b** are available online.

Figure 2 The sequence variation of L1-3′-UTR-encoded G4 is coupled with L1 subspeciation. (**a**) Distribution of the 5′ start positions of all the PQSs found in the remnants of different L1 subfamilies. The positions are referenced in relation to the variable 3′ end of L1 remnants. The numbers in brackets show the overall percentage of the identified L1 subfamily remnants that contain a PQS and the percentage of all the PQSs that are still close to their original sequence (the LQS definition). (**b**) Substitution-based tree of L1-originated PQSs. We built the tree so as to account for the L1 subfamilies with at least five LQSs found in their remnants, using the neighbor-joining method based on the multiple sequence alignment (**Fig. 1c**). The subfamily of the retrotransposon, the number of instances in the given branch, the recovered original sequences of the most recent hominoid lineage and the substitution cascade along the L1PA5-to-L1Hs speciation events are reported. (**c**) The gradual decrease in original G4-motif stability in the younger hominoid-specific L1 elements, as assessed by G4 scores inferred from G4-seq experiments (Online Methods). Each individual box in the box plot represents the distribution of G4 scores from all instances of the corresponding G4Ai (i = 1–5) intact sequence (with varying flanks) in the human genome covered by the G4-seq experiment, and indicates the median (central line) and the first and third quartiles of the distribution (box edges). The whiskers indicate 1.5 times the interquartile range; points represent outliers. From left to right in the plot (G4A5 to G4A1), the sample sizes for **c** were 125, 807, 2,945, 765 and 505, and the P values are 7.73×10^{-1}, 7.68×10^{-16}, 2.96×10^{-15} and 2.49×10^{-4} (one-tailed Mann–Whitney nonparametric test; **Supplementary Note 1**). ns, not statistically significant ($P > 0.05$); ***$P < 0.001$. Source data for **c** are available online.

humans, DI^{G4} values were smaller and allowed for the robust identification of the original G4 sequences for the L1PA5–2 and L1Hs subfamilies (**Supplementary Table 3** and **Supplementary Fig. 1e**). Multiple sequence alignment of the most frequent PQSs in each L1 subfamily revealed a well-defined pattern of nucleotide substitutions (**Fig. 1c**). Together, these observations identify G4s in the 3′ UTRs of L1PA5–2 and L1Hs elements in their viable states (**Supplementary Note 1**). Unlike in younger L1 subfamilies, for which the remnants allowed the identification of the original G4 sequences, time-accumulated nucleotide substitutions substantially deteriorated older L1 remnants (**Fig. 2a**), impeding the robust inference of the original G-rich sequences in older L1 subfamilies.

Mutations in 3′-UTR G4 motifs are coupled with L1 speciation
Using multiple sequence alignment of the most frequent LQSs, we were able to construct a substitution-based tree and robustly state the relatedness of sequences in the L1 subfamily (**Fig. 2b**). Strikingly, not only did the resulting tree correctly recapitulate the age hierarchy of the hominoid lineage of L1s, but it also identified a cascade of

single-nucleotide substitutions in the original G4 sequences of different L1 subfamilies. This pattern of nucleotide substitutions demonstrates that mutations in the G4 motif are coupled with L1 subspeciation (**Fig. 2b**). By using a recently developed sequencing technique for the genome-wide assignment of G4 stabilities (G4-seq)[16], we further identified a gradual decrease of G4-formation propensity (Online
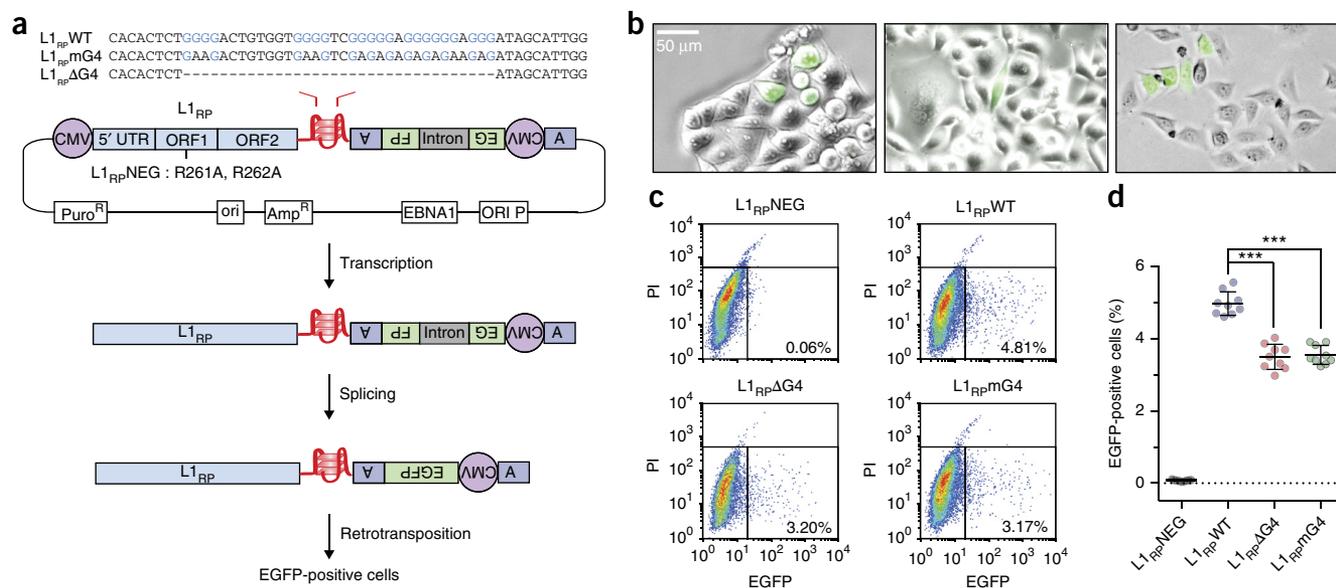
Figure 3 The G4 motif encoded within the 3′ UTR of the human $L1_{RP}$ element contributes to its retrotransposition. (**a**) Schematic representation of the L1 retrotransposition assay used in this study. The 3′ end of the human $L1_{RP}$ element in the pCEP4 episomal expression vector encodes an EGFP retrotransposition indicator cassette. The cassette consists of a backward copy of EGFP driven by a CMV promoter and interrupted by a self-splicing intron in the same transcriptional orientation as the $L1_{RP}$ RNA. EGFP-positive cells arise only if the $L1_{RP}$ transcript is reverse-transcribed, integrated into chromosomal DNA and expressed from its own CMV promoter. $pL1_{RP}WT$ expresses the $L1_{RP}$ element encoding the conserved G-rich sequence in its 3′ UTR. $pL1_{RP}WT$ and $pL1_{RP}\Delta G4$ display, respectively, a mutated and a deleted quadruplex motif. $pL1_{RP}NEG$ expresses an inactivated $L1_{RP}$ element. After transfection of an EGFP-tagged $L1_{RP}$ element and antibiotic selection, retrotransposition was detected by EGFP fluorescence under UV light. (**b**) Fluorescence microscopy images of HeLa cells 14 d after transfection with the $pL1_{RP}WT$ construct. Scale bar applies to all images. (**c**) FACS profiles of HeLa cells transfected with $pL1_{RP}NEG$, $pL1_{RP}WT$, $pL1_{RP}G4$ or $pL1_{RP}mG4$. Deletion or mutation of the G4 motif within the 3′ UTR of the $L1_{RP}$ element resulted in an ~30% decrease of retrotransposition activity compared to the wild-type element. ORF1 mutation ($L1_{RP}NEG$) resulted in no detectable retrotransposition events. (**d**) Relative transposition efficiency of the four elements, as assessed by percentage of EGFP-positive cells by FACS analysis (data represent mean ± s.d.; $n = 9$ independent transfection experiments). ***$P < 0.001$, two-tailed unpaired t test. Source data for **d** are available online.

Methods and **Supplementary Note 1**) with the emergence of younger L1 subfamilies (**Fig. 2c**). This observation suggests that active L1 elements bearing more stable G4 motifs in their 3′ UTRs are subject to negative selection, which is supported by the reported fitness cost of L1 activity in humans[17].

Overall, our computational analysis demonstrates that G4 sequences in the 3′ UTR of L1 elements are structural features conserved across different L1 subfamilies in which single-nucleotide substitutions throughout evolution cause a gradual decrease in G4 stability that is coupled with L1 subspeciation and the number of insertions. These observations are consistent with a functional role of G-rich sequences in 3′ UTRs of L1 elements, specifically through the formation of G4 structures. Notably, L1 subfamilies harboring more stable G4s had higher genomic copy numbers (**Supplementary Table 3**), consistent with a link between G4 stability and L1 retrotransposition activity (discussed below).

G4-motif alteration modulates L1Hs mobility in cultured cells
To test this hypothesis, we assessed the effect of G4 formation and stability on the activity of an L1 element. We used an episomal system[18,19] (**Fig. 3a**) to monitor the frequency of retrotransposition of the $L1_{RP}$ element, a human TE inserted into the *RP2* gene of a person with retinitis pigmentosa. The $L1_{RP}$ element harbors the L1Hs-specific G4 sequence, which folds into a G4 structure at both DNA and RNA levels *in vitro* as judged by UV, circular dichroism and ^{1}H NMR spectroscopic analyses (**Supplementary Fig. 2**). For the retrotransposition assay, the $L1_{RP}$ element was inserted into a pCEP4 vector and was constitutively expressed by a CMV promoter (**Fig. 3a**). Additionally, the

plasmid comprised a retrotransposition indicator cassette that consists of a reverse copy of EGFP under the same promoter and interrupted by a self-splicing intron in the same transcriptional orientation as the $L1_{RP}$ RNA. This arrangement ensures that EGFP expression becomes activated only upon successful $L1_{RP}$ retrotransposition. We transfected HeLa cells with $pL1_{RP}WT$—a construct expressing the unmodified TE—and cultured them for 14 d under antibiotic selection. At the end of that 14-d period, we were able to detect EGFP production in cells by fluorescence microscopy (**Fig. 3b,c**). We performed quantitative assessment of $L1_{RP}$ retrotransposition efficiency by counting the resulting number of EGFP-positive cells via FACS analysis. After $pL1_{RP}WT$ transfection, 4.96% ± 0.32% (mean ± s.d.; $n = 9$ independent transfection experiments) of the cells were found to be EGFP positive (**Fig. 3d**). In contrast, a construct expressing a functionally inactive $L1_{RP}$ (harboring two missense mutations in ORF1) did not yield any detectable retrotransposition events under similar conditions (0.07% ± 0.02% EGFP-positive cells (mean ± s.d.; $n = 9$ independent transfection experiments)).

We interrogated the role of the L1Hs-specific G-rich sequence by assessing the activities of $L1_{RP}$ variants in which this sequence was either deleted or mutated (**Supplementary Fig. 3**). Deletion of the G-rich motif ($L1_{RP}\Delta G4$) resulted in an ~30% decrease of retrotransposition activity (3.46% ± 0.32% EGFP-positive cells (mean ± s.d.; $n = 9$ independent transfection experiments)) compared to that in $L1_{RP}WT$ (**Fig. 3c,d**). This result indicates that the G-rich sequence contributes positively to the retrotransposition of the $L1_{RP}$ element. Mutation of the G-rich motif ($L1_{RP}mG4$; nine G-to-A mutations to disrupt G4 formation while preserving the

Figure 4 Stabilization of the G4 motif in the 3′ UTR stimulates human L1$_{RP}$ mobility. (**a**) Structure of the quadruplex ligands PDS, PhenDC3, 12459 and PDC12 together with the negative control PDC20. (**b**) Effect of small-molecule treatments on the relative transposition efficiency of the L1$_{RP}$WT and L1$_{RP}$ΔG4 elements, as assessed by FACS analysis (data represent mean ± s.d.; $n = 3$ independent transfection experiments). (**c**) Concentration-dependent stimulation of retrotransposition in L1$_{RP}$WT, but not L1$_{RP}$ΔG4, by the small molecule PDC12 (data represent mean ± s.d.; $n = 3$ independent transfection experiments). ns, not statistically significant ($P > 0.05$); **$P < 0.01$, ***$P < 0.001$; two-tailed unpaired t test. Source data for **b** and **c** are available online.

purine content of the sequence) resulted in a similar decrease in retrotransposition activity (3.48% ± 0.24% EGFP-positive cells (mean ± s.d.; $n = 9$ independent transfection experiments); **Fig. 3c,d**), thus suggesting that the G4 structural motif, rather than the sequence, modulates L1$_{RP}$ retrotransposition. Taken together, our results suggest that the conserved G4 motif found in the 3′ UTRs of hominoid L1 elements contributes to their activity. Because retrotransposition of TEs is a multistep mechanism, we postulate that the conserved G4 motif is one of the *cis*-regulatory elements within L1 retrotransposons that modulate their activity and influence the selection and evolution of hominoid L1s. Given that the presence of G4 structures was found to increase retrotransposition activity, factors that promote the formation of and/or stabilize the L1 G4 folded structures would be expected to stimulate retrotransposition.

Stabilization of the 3′-UTR G4 motif stimulates L1Hs mobility

To assess the effect of L1Hs G4 structural stabilization on retrotransposition, we investigated the effect of small-molecule G4 ligands on L1$_{RP}$ activity. We used four different small molecules that stabilize G4 structures (**Fig. 4a**). Pyridostatin (PDS)[20], PhenDC3 (ref. 21) and 12459 (ref. 22) have all been reported to affect cellular pathways in a manner consistent with G4 stabilization. The fourth small molecule, PDC12, was recently identified in our laboratory through a cellular screening assay. Biophysical characterization confirmed that all four small molecules stabilized the L1Hs G4 structure in both DNA and RNA (**Supplementary Fig. 4**), and we measured their cell-growth inhibition properties to establish a subcytotoxic dosing range (**Supplementary Fig. 5a**). To evaluate the potential of these ligands to modulate retrotransposition, we transfected HeLa cells with either pL1$_{RP}$WT or pL1$_{RP}$ΔG4 and subsequently allowed them to grow in the presence

of a sub-GI$_{50}$ dose (where GI$_{50}$ is the drug concentration causing a 50% reduction in cell proliferation) of each compound. As before, we assessed the retrotransposition efficiencies of both TEs by FACS analysis. All G4-stabilizing molecules were found to stimulate the retrotransposition frequency of L1$_{RP}$WT but not L1$_{RP}$ΔG4 (**Fig. 4b** and **Supplementary Fig. 5a–d**). It is noteworthy that the control molecule PDC20, which is structurally related to PDC12 but unable to stabilize L1Hs quadruplexes *in vitro* (**Fig. 4b** and **Supplementary Fig. 5a–d**), did not have an effect on the activity of either construct. Conversely, the heavy metal salt HgS, known to stimulate human L1 mobility[23], was found to increase the frequency of retrotransposition of both L1$_{RP}$WT and L1$_{RP}$ΔG4 elements (**Fig. 4b** and **Supplementary Fig. 5a–d**). This latter result indicates that deletion of the G4 motif does not impede modulation of L1$_{RP}$ activity by other mechanisms. Lastly, we assessed the activity of cells transfected with L1$_{RP}$WT or L1$_{RP}$ΔG4 when incubated with increasing concentrations of PDC12 (up to 40 μM). We observed a concentration-dependent increase of L1$_{RP}$WT activity (approximately two-fold at higher concentrations), whereas no significant stimulation ($P > 0.05$) was observed for L1$_{RP}$ΔG4 at any ligand concentration tested (**Fig. 4c** and **Supplementary Fig. 5e–g**). Taken together, these data support the hypothesis that the formation and stabilization of the G4 structural motif encoded within the 3′ UTR of the L1$_{RP}$ element stimulate its activity. Furthermore, observations from our computational analysis (**Supplementary Table 3**), in which L1 subfamilies harboring more stable G4s had higher genomic copy numbers, are in line with this hypothesis.

DISCUSSION

Herein we have presented computational and experimental evidence to support the idea that stable G4 formation in the 3′ UTR of L1 elements

contributes to hominoid L1 selection and mobility. Although the contribution of G4 motifs to L1 mobility is moderate, it is worth noting that retrotransposition events are generally infrequent (~10^{-1} events per generation[24]) and that L1 elements have coevolved with the human genome over millions of years[25]. Hence, the G4s may have notably contributed to the selection and accumulation of L1 elements and, therefore, might constitute a substantial part of the forces that shape the human genome.

Interestingly, the nonautonomous TEs that rely on the L1 machinery for retrotransposition, such as Alu and SVA elements[11], and retroviral LTRs, such as in the HIV genome[26], display conserved G4-forming sequences. This observation suggests that G4 motifs may be drivers of gene copy-number variation and horizontal gene transfer. Hence G4s ought to be considered to contribute to the forces that drive genome evolution. The molecular mechanism by which G4 motifs modulate L1 mobility could involve L1 retrotransposition at DNA (replication, recombination or transcription regulation) and/or RNA (reverse transcription, stabilization and transport of L1 mRNA) levels, and exploring such possibilities in detail in future work could offer useful insights.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS
All authors contributed to the concepts and design of the research. A.B.S. carried out the computational analyses. P.M. and C.M. carried out the retrotransposition experiments. All authors interpreted the data. A.B.S. and P.M. wrote the manuscript with contributions from all authors. S.B. supervised the research.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Cordaux, R. & Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
2. Kazazian, H.H. Jr. *Mobile DNA: Finding Treasure in Junk* (Pearson Education, 2011).
3. Levin, H.L. & Moran, J.V. Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* **12**, 615–627 (2011).
4. Erwin, J.A., Marchetto, M.C. & Gage, F.H. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat. Rev. Neurosci.* **15**, 497–506 (2014).
5. Ohshima, K. RNA-mediated gene duplication and retroposons: retrogenes, LINEs, SINEs, and sequence specificity. *Int. J. Evol. Biol.* **2013**, 424726 (2013).
6. Kajikawa, M. & Okada, N. LINEs mobilize SINEs in the eel through a shared 3′ sequence. *Cell* **111**, 433–444 (2002).
7. Takahashi, H. & Fujiwara, H. Transplantation of target site specificity by swapping the endonuclease domains of two LINEs. *EMBO J.* **21**, 408–417 (2002).
8. Hayashi, Y., Kajikawa, M., Matsumoto, T. & Okada, N. Mechanism by which a LINE protein recognizes its 3′ tail RNA. *Nucleic Acids Res.* **42**, 10605–10617 (2014).
9. Moran, J.V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917–927 (1996).
10. Howell, R. & Usdin, K. The ability to form intrastrand tetraplexes is an evolutionarily conserved feature of the 3′ end of L1 retrotransposons. *Mol. Biol. Evol.* **14**, 144–155 (1997).
11. Lexa, M. *et al.* Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics* **15**, 1032 (2014).
12. Bochman, M.L., Paeschke, K. & Zakian, V.A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13**, 770–780 (2012).
13. Kejnovsky, E. & Lexa, M. Quadruplex-forming DNA sequences spread by retrotransposons may serve as genome regulators. *Mob. Genet. Elements* **4**, e28084 (2014).
14. Kejnovsky, E., Tokan, V. & Lexa, M. Transposable elements and G-quadruplexes. *Chromosome Res.* **23**, 615–623 (2015).
15. Huppert, J.L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33**, 2908–2916 (2005).
16. Chambers, V.S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33**, 877–881 (2015).
17. Boissinot, S., Davis, J., Entezam, A., Petrov, D. & Furano, A.V. Fitness cost of LINE-1 (L1) activity in humans. *Proc. Natl. Acad. Sci. USA* **103**, 9590–9594 (2006).
18. Ostertag, E.M., Prak, E.T., DeBerardinis, R.J., Moran, J.V. & Kazazian, H.H. Jr. Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic Acids Res.* **28**, 1418–1423 (2000).
19. Farkash, E.A., Kao, G.D., Horman, S.R. & Prak, E.T. Gamma radiation increases endonuclease-dependent L1 retrotransposition in a cultured cell assay. *Nucleic Acids Res.* **34**, 1196–1204 (2006).
20. Rodriguez, R. *et al.* Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat. Chem. Biol.* **8**, 301–310 (2012).
21. Piazza, A. *et al.* Genetic instability triggered by G-quadruplex interacting Phen-DC compounds in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **38**, 4337–4348 (2010).
22. Riou, J.F. *et al.* Cell senescence and telomere shortening induced by a new series of specific G-quadruplex DNA ligands. *Proc. Natl. Acad. Sci. USA* **99**, 2672–2677 (2002).
23. Kale, S.P., Moore, L., Deininger, P.L. & Roy-Engel, A.M. Heavy metals stimulate human LINE-1 retrotransposition. *Int. J. Environ. Res. Public Health* **2**, 14–23 (2005).
24. Ostertag, E.M. & Kazazian, H.H. Jr. Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35**, 501–538 (2001).
25. Khan, H., Smit, A. & Boissinot, S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* **16**, 78–87 (2006).
26. Métifiot, M., Amrane, S., Litvak, S. & Andreola, M.L. G-quadruplexes in viruses: function and potential therapeutic applications. *Nucleic Acids Res.* **42**, 12352–12366 (2014).

ONLINE METHODS

General notes on computational genomics. We used the GRCh37 (hg19) unmasked reference sequence of the human genome, as accessed through the Ensembl database (http://www.ensembl.org). We identified repetitive elements by using the full RepeatMasker[27] annotation of the human genome through the supplied tables of the UCSC genome browser (http://www.genome.ucsc.edu). We plotted sequence logos[28] using the seqLogo library of R[29]. Estimated ages of L1 elements were recovered from Khan *et al.*[25]. We conducted multiple sequence alignments with ClustalW[30] using the default alignment matrix for nucleic acids and by iteratively refining individual alignment steps. All the other calculations and genomic analyses were done with in-house code written in the R programming language[29]. The performed calculations are discussed in brief in the sections below, with the complete details and reasoning presented in **Supplementary Note 1**.

Identification of L1-originated quadruplex sequences (LQS). Potential quadruplex sequences (PQSs) of the form $G_{3+}N_{1-12}G_{3+}N_{1-12}G_{3+}N_{1-12}G_{3+}$, where N is any base, including G, were recovered from the human genome. In the case of overlapping motifs, the longest sequences were considered. The frequency of each unique PQS was calculated. Six out of the 15 most frequent (from among 703,091) PQSs from the human genome were found to originate from L1 elements (**Supplementary Table 1**) and were used to define the LQS family as follows. The most frequent LQS representative, GGGGACTGTTGTGGGGTGGGGGGAGGGGGGAGGG, referred to as LQS^{ref} and later identified as having stemmed from the L1PA3 family, was used as a reference to reveal all related PQSs by hierarchical clustering analysis (**Supplementary Note 1**). We compared all PQSs with lengths similar to that of LQS^{ref} (34 nt) to LQS^{ref} directly, by calculating the Hamming distance[31], d_H, a unitless similarity measure between strings of equal length. For a given sequence in our analysis, d_H is the number of characters that differed, in a position-specific manner, from LQS^{ref}. The d_H distribution revealed a distinct family of PQSs with $d_H \leq 5$ (**Supplementary Fig. 1**) characterized by 10,269 occurrences of 3,238 unique sequences. Finally, we screened all PQSs, not restricted to 34-nt-long sequences, against the identified 34-nt core to find all occurrences of the related sequences. The resulting pool of sequences defined the LQS family.

Analysis of the LQS family. LQSs were assigned to the corresponding L1 remnants using the RepeatMasker annotation. The distribution and occurrence of PQSs in different L1 subfamilies are reported in **Supplementary Table 2**. For each retrotransposon subfamily, in order to focus on the relevant polymorphism, we tried to determine the original LQS sequences (**Supplementary Note 1**), with a marked prevalence in their copy numbers as compared to other PQSs found in the corresponding retrotransposon subfamily. To identify the original LQSs, we quantified the enrichment of a given PQS in each L1 family through the introduced G4 diversity index, DI^{G4}, which is the ratio of all unique PQSs versus the total number of PQSs found in a given element. Hominoid-specific elements, L1Hs and L1PA2–5, showed low DI^{G4}, indicating the preservation of the information content to identify the original LQS sequences. For subsequent analysis, we considered only the original LQSs, in order to correct for the noise introduced by time-accumulated random nucleotide substitutions (**Supplementary Table 3**, **Supplementary Note 1** and **Supplementary Fig. 1e**). We used multiple sequence alignment of the revealed original LQSs to construct a substitution-based tree, visualized via Dendroscope[32], rooted on the representative from the L1PA8A subfamily.

Structural stability of the original LQSs of young L1s. G4 motif stability was estimated as reflected in G4 scores inferred from the G4-seq experiments. G4-seq is a genome-wide G4-detection method that uses high-throughput sequencing[16] and takes advantage of polymerase stalling at G4 sites during sequencing, which causes the incorporation of identifiable mismatches. The maximum observed percentages of those mismatches (mm%) close to G4 motifs are reflective of G4 stability. Human G4-seq data (GEO GSE63874) were recovered, and the mm% values for all G4A1–5 LQSs (with 50-nt flanking regions) were extracted from the data set for K$^+$ versus Na$^+$ experiment. We then calculated the relative G4 scores (**Fig. 2c**) by dividing the obtained mm% values by 26.7 (the median mm% of the most stable G4A4 sequence; **Fig. 2b**). The significance of the differences among the mean G4 scores for the LQSs from different L1 subfamilies was then assessed via a one-tailed Mann–Whitney nonparametric test (**Supplementary Note 1**).

Recombinant DNA plasmids. Plasmids pL1$_{RP}$WT (EF06R) and pL1$_{RP}$NEG (EF05J) were obtained from Addgene. Among others, these plasmids contain the L1 element under control of an upstream CMV promoter and a retrotransposition indicator cassette. Plasmids lacking the conserved G4 motif (pL1$_{RP}$ΔG4) or harboring a mutated version of the sequence (pL1$_{RP}$mG4) were derived from pL1$_{RP}$WT via standard molecular cloning procedures. In brief, we constructed inserts with the corresponding sequences by oe-PCR using the primers L1fw and L1rv with the corresponding primers for each construct (**Supplementary Note 2**). The resulting inserts and the parent plasmid were digested with *SapI* (New England BioLabs) before ligation with T7 ligase (New England BioLabs). Ligation reactions were purified and transformed using XL-1 blue chemically competent cells. Ligation reactions were spread on LB agar plates containing ampicillin (150 µg/mL) and incubated at 37 °C overnight. Clones containing the desired plasmids were identified by Sanger sequencing (**Supplementary Fig. 3**), and plasmids used in the transfection experiments were purified with a Plasmid Midi Kit (Qiagen).

Cell culture and materials. HeLa cells were grown in an incubator under a humidified 5% CO_2 atmosphere at 37 °C in high-glucose (4.5 g/L) Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine calf serum (DMEM complete) and split at 70–80% confluence using trypsin-EDTA. The cell line was genotyped via short tandem repeat profiling and tested for mycoplasma. PDS[33], PhenDC3 (ref. 34) and 12459 (ref. 35) were synthesized as previously described. The synthesis and characterization of PDC12 and PDC20 are reported in **Supplementary Note 2**. Mercury (II) sulfide red (HgS) and propidium iodide (PI) were purchased from Aldrich.

Retrotransposition assay. HeLa cells were seeded in six-well dishes at a density of 1×10^5 cells per well and grown at 70% confluence in DMEM complete. Cells were transfected with the FuGENE 6 transfection reagent (Promega) according to the manufacturer's protocol. Each transfection well received 2 µg of plasmid DNA, 12 µL of FuGENE 6 reagent and 2 mL of DMEM complete. Puromycin selection (2 µg/mL) was started 48 h after transfection. Puromycin-resistant cells were selected by growth in DMEM complete containing 2 µg/mL puromycin. For small-molecule and chemical treatments, we added the appropriate volume of DMSO stock solutions to a puromycin-containing DMEM complete solution and changed the cell media every 2 d. PDS and PhenDC3 were added to the cell cultures 9 d after transfection. 12459, PDC12, PDC20 and HgS were added 2 d after transfection.

Fluorescence-activated cell sorting (FACS). After 14 d of culture, we prepared the cells for FACS analysis by washing them twice with PBS and then incubating them for 10 min with trypsin-EDTA. The suspended cells were collected by centrifugation, resuspended in PBS supplemented with PI (10 µg/mL) and kept on ice until FACS analysis. Cells were analyzed with a FACSCalibur system (BD Biosciences). Between 10,000 and 20,000 total events were monitored per sample. For live/dead gating, we excluded PI-stained cells. We analyzed living cells for fluorescence intensity using a blue argon laser (488 nm) and a filter set (533/30 band pass). Data were analyzed with the FlowJo Software.

Statistical tests. The significance of the observed G4-score reduction among the different G4A*i* sequences depicted in **Figure 2c** was tested via one-tailed Mann–Whitney nonparametric test. The null hypothesis was that the younger L1 G4 sequences had an average G4 score not lower than that of the older ones (**Supplementary Note 1**). The significance of the observed differences in retrotransposition assays was assessed via two-tailed unpaired *t* test. All the relevant information and the sample sizes are provided in the main text, figure captions and **Supplementary Note 1**.

Code availability. All the computer programs and genomic databases used in this study are openly available, as detailed in **Supplementary Note 1** and the citations therein. In-house-generated R scripts used for the analyses, data exploration and plotting are available from the authors upon request.

Data availability. Source data for **Figures 1b**, **2c**, **3d**, **4b** and **4c** are available online. All other data are available from the corresponding author upon reasonable request.

27. Smit, A.F.A., Hubley, R. & Green, P. *RepeatMasker Open-4.0* (Institute for Systems Biology, 2015).
28. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
29. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015).
30. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
31. Hamming, R.W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–160 (1950).
32. Huson, D.H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061–1067 (2012).
33. Rodriguez, R. *et al.* A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. *J. Am. Chem. Soc.* **130**, 15758–15759 (2008).
34. De Cian, A., Delemos, E., Mergny, J.-L., Teulade-Fichou, M.-P. & Monchaud, D. Highly efficient G-quadruplex recognition by bisquinolinium compounds. *J. Am. Chem. Soc.* **129**, 1856–1857 (2007).
35. Mailliet, P., Riou, J.-F., Mergny, J.-L., Laoui, A., Lavelle, F. & Petitgenet, O. Dérivés arylamines et leur application comme agent antitélomerase [Arylamine derivatives and their use as anti-telomerase agents]. Patent application WO 2001040218 (2001).