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

Aleksandr B. Sahakyan,[1,2,§] Pierre Murat,[1,2,§] Clemens Mayer,[1,2] and Shankar Balasubramanian[1,2,3,*]

[1] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK.
[2] Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK.
[3] School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SP, UK.

[§] These authors contributed equally to this work.
[*] Correspondence to: sb10031@cam.ac.uk (S.B.)

Long Interspersed Nuclear Elements (LINEs) are ubiquitous transposable elements in higher eukaryotes that play a significant role in shaping genomes due to their abundance. Here, we report that guanine-rich sequences in the 3'-UTR 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 stabilising the G4 motif of a human-specific LINE-1 element with 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 is comprised of TEs, and, owing to their association with genomic instability through *de novo* insertion and recombination events, TEs are responsible for a number of genetic disorders and cancers.[2,3] Long Interspersed Nuclear Elements-1 (LINE-1 or L1) are a type of 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 a RNA polymerase II promoter, two open reading frames (ORF1 and ORF2 encoding for a RNA-binding protein and a reverse transcriptase with endonuclease activity respectively), and a 3'-UTR that harbours a guanine-rich sequence together with a polyadenylation signal and terminates with an adenine-rich tail (**Fig. 1a**).[4] Unlike LINE elements in plants[5] and some metazoans,[6,7] in which a conserved, stem-loop secondary structures at the 3'-tail control *de novo* insertions by recruiting the ORF2-encoded protein,[8] mammalian 3'-UTR 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 the human and primate genomes.[9] Although the 3'-UTR of primate L1 elements are devoid of canonical secondary structures, they display conserved guanine-rich sequences (**Fig. 1a**) with the ability to fold *in vitro* into G-quadruplex (G4) structures.[10,11] G4s are non-canonical secondary structures formed by guanine-rich nucleic acids and stabilised by the stacking of guanine tetrads held by Hoogsteen base pairing.[12] Recent works highlighted that putative G-quadruplex sequences (PQSs) 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] Based on these observations, we hypothesised a functional role for G4-forming sequences within L1 elements and have carried out a study to evaluate the contribution of the L1-3'-UTR on retrotransposition activity.

RESULTS

Guanine-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 (LQS family, see on-line **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, and are characteristic of **p**otential **q**uadruplex **s**equences (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 family with shared origin (on-line **Methods**, **Supplementary Table 1**, **Supplementary Fig. 1a-c**). In total, the analysis identified 3,228 unique LQSs that are located in 15,724 genomic locations, with all but 5 present in the 3'-UTR of L1 elements (**Supplementary Fig. 1d**). This highly preferential co-localisation 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 their corresponding L1 subfamilies (**Supplementary Table 2**), we then sought to identify original G4 sequences present in their respective viable states. To reveal the L1 subfamilies for which time-accumulated random substitutions left enough sequence preference to infer the original LQS, we introduced a G4 diversity index (DI^{G4}), defined as the ratio of unique PQSs over the total number of PQSs found in a given L1 subfamily. A DI^{G4} value of 1 reflects that all PQSs found in a given L1 subfamily are different, each affected by different substitutions. Smaller values indicate lower diversity, where the original PQSs can still be differentiated by their abundance. DI^{G4} values increase with the age of L1 subfamilies, demonstrating the erosive effect of time-accumulated, random substitutions (**Fig. 1b**). Conversely, for the youngest lineage of L1 elements, specific to hominoid apes and humans, DI^{G4} values are smaller and allowed for the robust identification of the original G4 sequences for the L1PA5-2 and L1Hs subfamilies (**Supplementary Table 3**, **Supplementary Fig. 1e**). A 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'-UTR of L1 elements during their viable states. While the remnants of younger elements allowed the identification of the original sequences, time-accumulated nucleotide substitutions have deteriorated PQSs found in older L1 remnants (**Fig. 2a**).

Mutations in 3'-UTR G4 motifs are coupled with L1 speciation

Using a 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 does the resulting tree correctly recapitulate the age hierarchy of the hominoid lineage of L1s, but it also identifies a cascade of single-nucleotide substitutions in the original G4

sequence of different L1 subfamilies. This pattern of nucleotide substitutions demonstrates that mutations in the G4 motif are coupled with L1 sub-speciation (**Fig. 2b**). Using a recently developed sequencing technique (G4-seq) for the genome-wide assignment of G4-stabilitites,[16] we further identified a gradual decrease of G4 formation propensities (on-line **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'-UTR) are subject to negative selection, which is supported by the reported fitness cost of L1 activity in humans.[17]

Overall, our computational analyses demonstrates that G4 sequences in the 3'-UTR of L1 elements are conserved structural features across different L1 sub-families, in which single-nucleotide substitutions caused a gradual decrease in G4-stability that is coupled with L1-sub-speciation and the number of insertions. These observations are consistent with a functional role of G-rich sequences at 3'-UTRs of L1 elements, specifically through the formation of G4 structures. Notably, L1 subfamilies harbouring more stable G4s display higher genomic copy numbers (**Supplementary Table 3**), consistent with a link between G4 stability and L1 retrotransposition activity (*vide infra*).

G4 motif alteration modulates L1Hs mobility in cultured cells

To test this hypothesis, we assessed the impact of G4 formation and stability on the activity of an L1 element. We employed 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 *retinitis pigmentosa* patient. The $L1_{RP}$ element harbours 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 is inserted into a pCEP4 vector and 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. HeLa cells were transfected with $pL1_{RP}$WT – a construct expressing the unmodified TE – and cultured for 14 days under antibiotic selection. At this time, EGFP production could be detected by fluorescence microscopy in cells (**Fig. 3b,c**). A quantitative assessment of $L1_{RP}$ retrotransposition efficiency was obtained by counting the resulting number of EGFP-positive cells *via* a FACS analysis. Following $pL1_{RP}$WT transfection, 4.96±0.32 % (mean values ± s.d.; n = 9

independent transfection experiments) of the cells were found to be EGFP positive (**Fig. 3d**). Conversely, a construct expressing a functionally inactive $L1_{RP}$ (harbouring two missense mutations in ORF1) did not yield any detectable retrotransposition event under comparable conditions (0.07±0.02 % EGFP positive cells (mean values ± 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 a ~30 % decrease of retrotransposition activity (3.46±0.32 % EGFP positive cells (mean values ± s.d.; n = 9 independent transfection experiments)) compared to $L1_{RP}$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, 9 G-to-A mutations to disrupt G4 formation while preserving the purine content of the sequence) resulted in a similar decrease in retrotransposition activity (3.48±0.24 % EGFP positive cells (mean values ± s.d.; n = 9 independent transfection experiments), **Fig. 3c,d**), suggesting that the G4 structural motif, rather than the sequence, modulates $L1_{RP}$ retrotransposition. Taken together, these results suggest that the conserved G4 motif found in the 3'-UTR of hominoid L1 elements contributes to their activities. Because retrotransposition of TEs is a multi-step mechanism, we postulate that the conserved G4 motif is one of the *cis*-regulatory elements within L1 retrotransposons, modulating their activity and influencing the selection and evolution of hominoid TEs. Based on the observation that the presence of G4 structures increases retrotransposition activity, factors that promote the formation of and (or) stabilise the L1 G4 folded structures would be expected to stimulate retrotransposition.

Stabilisation of the 3'-UTR G4 motif stimulates L1Hs mobility

To assess the effect of stabilising the L1Hs G4 structure on retrotransposition, we investigated the effect of small-molecule G4-ligands on $L1_{RP}$ activity. We employed four different small molecules that stabilise G4 structures (**Fig. 4a**). Pyridostatin (PDS)[20], PhenDC3[21] and 12459[22] have each been reported to affect cellular pathways in a manner consistent with G4 stabilisation, while PDC12 was recently identified in our laboratory through a cellular screening assay. Biophysical characterisation confirmed that all four small molecules stabilize the L1Hs G4 structure in both DNA and RNA (**Supplementary Fig. 4**) and we measured their cell-growth inhibition properties to establish a sub-cytotoxic dosing range (**Supplementary Fig. 5a**). To evaluate the potential of these ligands to modulate retrotransposition, HeLa cells were first transfected with either $pL1_{RP}$WT or $pL1_{RP}\Delta G4$ and

subsequently grown in presence of a sub-GI_{50} dose of each compound. As before, retrotransposition efficiencies of both TEs were assessed by FACS analysis. All G4-stabilising molecules were found to stimulate the retrotransposition frequency of $L1_{RP}$WT but not $L1_{RP}$ΔG4 (**Fig 4b**, **Supplementary Fig. 5a-d**). It is noteworthy that the control molecule PDC20, that is structurally related to PDC12 but unable to stabilise L1Hs quadruplexes *in vitro* (**Fig. 4b**, **Supplementary Fig. 5a-d**), did not have an effect on the activity of either construct. Conversely, a heavy metal salt (HgS), known to stimulate human L1 mobility,[23] was found to increase the frequency of retrotransposition of both the $L1_{RP}$WT and $L1_{RP}$ΔG4 elements (**Fig. 4b**, **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 activities of $L1_{RP}$WT or $L1_{RP}$ΔG4 transfected cells when incubated with an increasing concentration of PDC12 (up to 40 μM). We observed a concentration-dependent increase of $L1_{RP}$WT activity (~2-fold at higher concentrations), while no significant stimulation (p-value > 0.05) was observed for $L1_{RP}$ΔG4 at any ligand concentration tested (**Fig. 4c**, **Supplementary Fig. 5e-g**). Taken together, these data support that the formation and stabilisation of the G-quadruplex 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 harbouring more stable G4s display higher genomic copy numbers, are in line with this hypothesis.

DISCUSSION

Herein we have presented computational and experimental evidence to support that stable G4 formation in the 3'-UTR of L1 elements contributes to hominoid L1 selection and mobility. Although the contribution of G4 motifs on L1 mobility is moderate, it is worth noting that retrotransposition events are generally rare (~10^{-1} events per generation[24]) and that L1 elements have co-evolved 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, constitute a significant part of the forces that shape the human genome.

Interestingly, the non-autonomous TEs that rely on the L1 machinery for retrotransposition, such as Alu and SVA elements,[11] and retroviral LTRs, such as in HIV genome,[26] display conserved G4-forming sequences. This observation suggests that G4 motifs may be drivers of gene copy variation and horizontal gene transfer. Hence G4s ought to be considered part of forces that drive genome evolution. The molecular mechanism by which G4 motifs modulate L1 retrotransposition could

involve L1 retrotransposition at DNA (replication, recombination or transcription regulation) and (or) RNA (reverse-transcription, stabilisation and transport of L1 mRNA) levels and it will be insightful to explore such possibilities in detail in future work.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary Information is available in the online version of the paper.

ACKNOWLEDGEMENTS

S.B. is a Wellcome Trust Senior Investigator (grant no. 099232/z/12/z). The Balasubramanian group is supported by a European Research Council Advanced Grant (no. 339778) and receives core funding from Cancer Research UK.

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.

Correspondence and requests for materials should be addressed to S.B.

REFERENCES

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. Genet.* **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. *Nucl. 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,** (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. Elem.* **4,** e28084 (2014).
14. Kejnovsky, E., Tokan, V. & Lexa, M. Transposable elements and G-quadruplexes. *Chromosome Res.* **23,** 615–623 (2015).
15. Huppert, J. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucl. 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. Biotech.* **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., DeBerardinis, R. J., Moran, J. V. & Kazazian, H. H. Determination of L1 retrotransposition kinetics in cultured cells. *Nucl. Acids Res.* **28,** 1418–1423 (2000).
19. Farkash, E. A. Gamma radiation increases endonuclease-dependent L1 retrotransposition in a cultured cell assay. *Nucl. 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. *Nucl. 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. Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35,** 501–538 (2001).
25. Khan, H. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* **16,** 78–87 (2006).
26. Metifiot, M., Amrane, S., Litvak, S. & Andreola, M. L. G-quadruplexes in viruses: function and potential therapeutic applications. *Nucl. Acids Res.* **42,** 12352–12366 (2014).

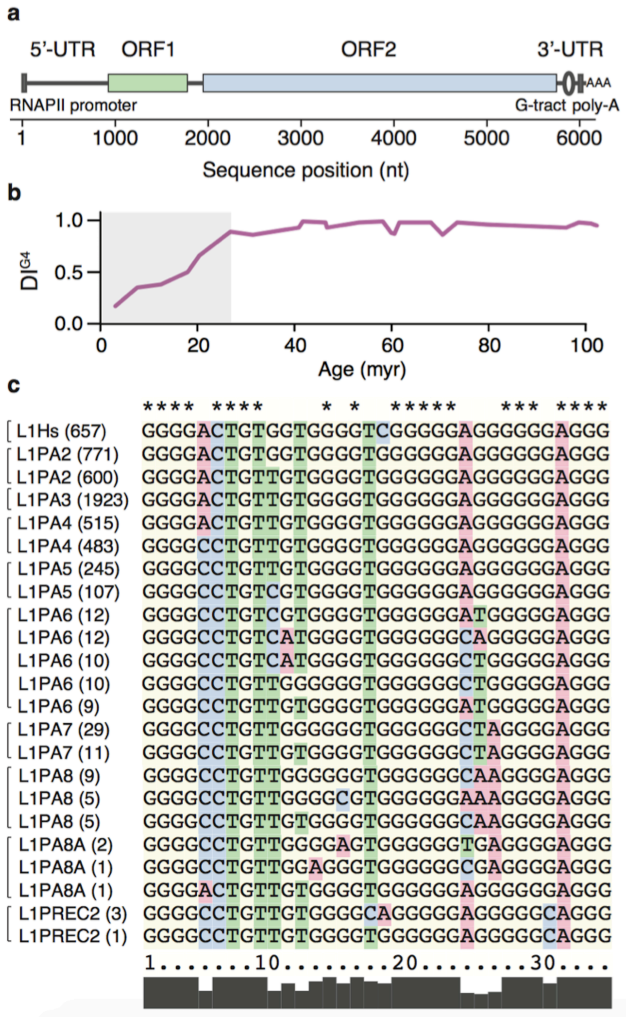


Figure 1 | G-rich sequences are a hallmark of young active L1 retrotransposons. (**a**) Sequence organisation of the L1 element, showing the position of the G-rich sequence. (**b**) Change in the DI^{G4} index, as a function of divergence age of the source L1 subfamilies. The shaded rectangle depicts the region where the DI^{G4} index is low, hence the time-accumulated substitutions have not interfered too much with the sequences of the corresponding age, and the original sequence can be robustly identified. (**c**) Most frequently occurring G4 sequences found within the 3'-UTR of primate L1 elements and their number of occurrence in the human genome, in brackets, together with a multiple sequence alignment plot highlighting the core nucleotide substitutions.

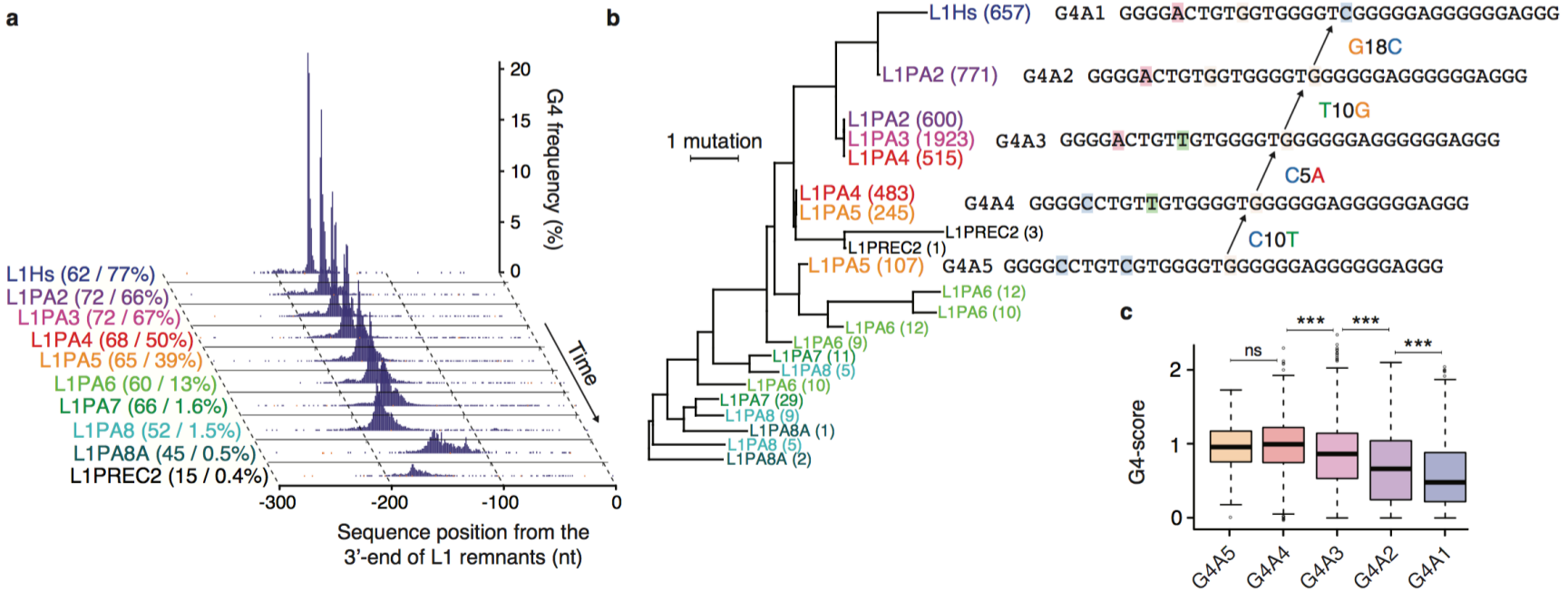


Figure 2 | Sequence variation of L1-3'-UTR-encoded G4 is coupled with L1 sub-speciation. (**a**) Distribution of the 5'-start positions of all the PQSs found in the remnants of different L1 subfamilies. The positions are referenced in relation to the variable 3'-end of L1 remnants. The numbers in the 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. The tree was built accounting for the L1 subfamilies with at least 5 LQSs found in their remnants, using the neighbour-joining method based on the multiple sequence alignment (**Fig. 1c**). The subfamily of the retrotransposon, the number of instances in the given branch, the sequences of the most recent hominoid lineage and the substitution cascade along the L1PA5 to L1Hs speciation events are reported. (**c**) Gradual decrease in G4-motif stability in the young L1 elements, as assessed by G4-scores inferred from the G4-seq experiments (see **on-line Methods**). Each individual box in the boxplot **c** shows the distribution of G4-scores from all instances of the corresponding G4Ai (i = 5 to 1) intact sequence (with varying flanks) in the human genome covered by the G4-seq experiment, and is constructed via the median (central line), first and third quartiles of the distribution (box limits). The whiskers depict the range of values within the 1.5 times IQR (interquartile range), with points indicating the outliers. From left to right, the sample sizes in **c** are 125, 807, 2945, 765 and 505 and the reported comparison p-values are 7.73×10^{-1}, 7.68×10^{-16}, 2.96×10^{-15} and 2.49×10^{-4} (one-tailed Mann-Whitney non-parametric test, see **Supplementary Note 1**). The boxplot annotations mean ns: not statistically significant (p-value > 0.05), ***: p-value < 0.001.

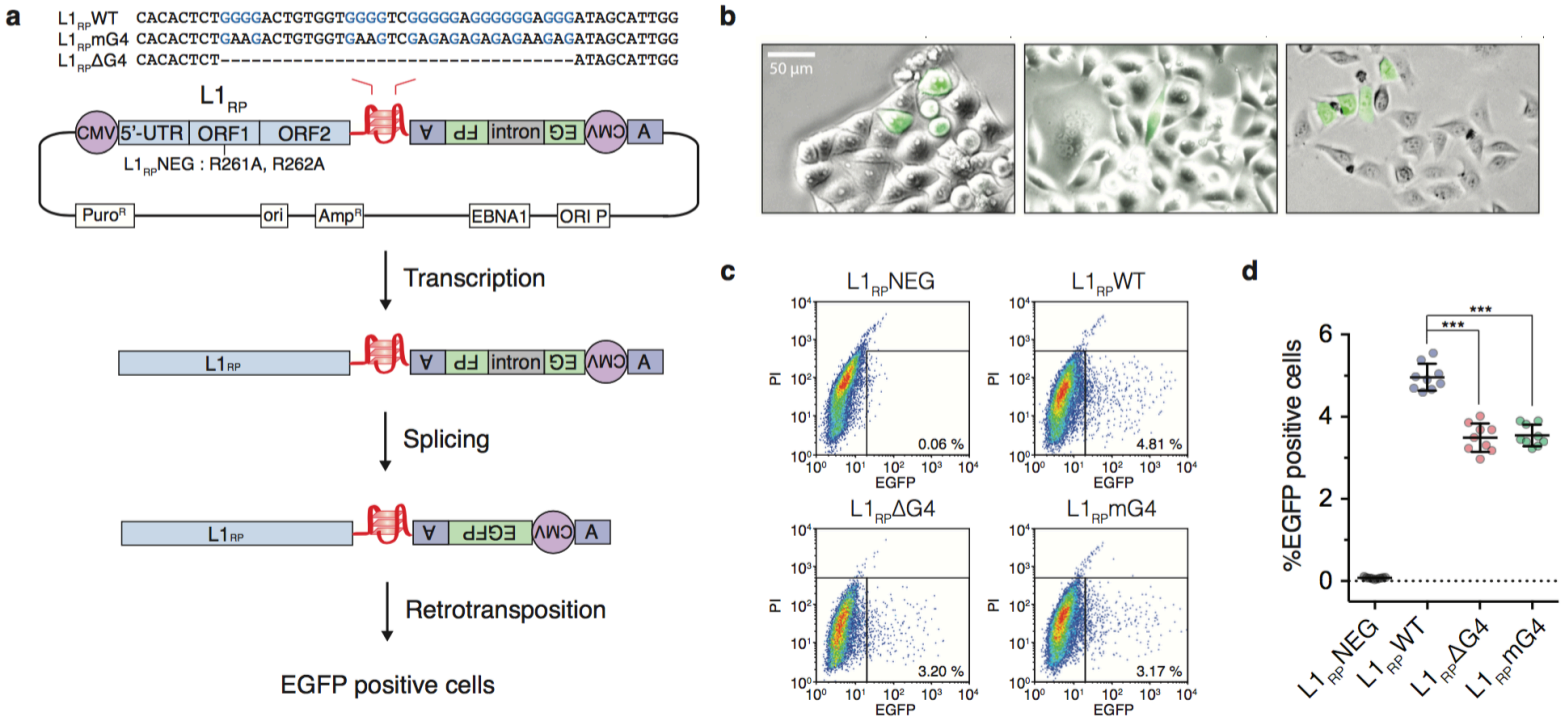


Figure 3 | The G4 motif, encoded within the 3'-UTR of the human $L1_{RP}$ element, contributes to its retrotransposition. (**a**) Schematic representation of the L1 retrotransposition assay used in this study. The 3'-end of the human $L1_{RP}$ element in the pCEP4 episomal expression vector encodes an 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 $pL1_{RP}$ΔG4 display respectively either a mutated or deleted quadruplex motif. $pL1_{RP}$NEG expresses an inactivated $L1_{RP}$ element. Following transfection of an EGFP-tagged $L1_{RP}$ element and antibiotic selection, retrotransposition is detected by EGFP fluorescence under UV light. (**b**) Example snapshots from a fluorescence microscopy (400× total magnification, the scale is shown in µm) of HeLa cells 14 days post transfection with the $pL1_{RP}$WT construct. (**c**) Representative FACS profiles of HeLa cells transfected with $pL1_{RP}$NEG, $pL1_{RP}$WT, $pL1_{RP}$ΔG4 or $pL1_{RP}$mG4. Deletion or mutation of the G4 motif within the 3'-UTR of the $L1_{RP}$ element resulted in a ~30% decrease of retrotransposition activity compared to the WT element. ORF1 mutation ($L1_{RP}$NEG) resulted in no detectable retrotransposition events. (**d**) Relative transposition efficiency of the four elements, as assessed by quantifying the percentage of EGFP positive cells by FACS analysis 14 days after transfection (data represent mean values ± s.d.; n = 9 independent transfection experiments). The p-values in **d** (***: p-value < 0.001) were calculated via two-tailed unpaired t-test.

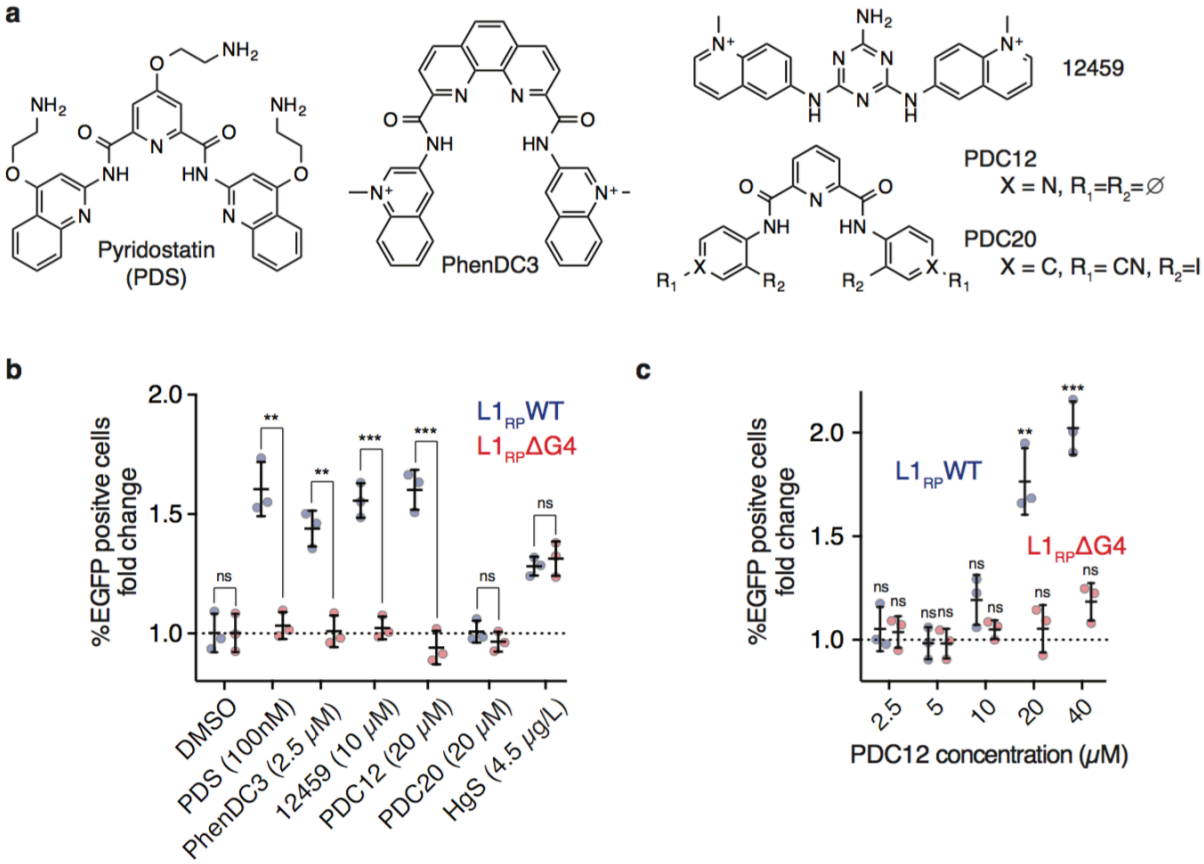


Figure 4 | Stabilisation of the G4 motif in the 3'-UTR stimulates the human $L1_{RP}$ mobility. (**a**) Structure of the quadruplex ligands PDS, PhenDC3, 12459 and PDC12 together with the negative control PDC20. (**b**) Effect of small molecule treatments on the relative transposition efficiency of the $L1_{RP}$WT and $L1_{RP}$ΔG4 elements, as assessed by FACS analysis (data represent mean values ± s.d.; n = 3 independent transfection experiments). (**c**) Concentration dependent stimulation of $L1_{RP}$WT retrotransposition but not $L1_{RP}$ΔG4 by the small molecule PDC12 (data represent mean values ± s.d.; n = 3 independent transfection experiments). The p-values in **b** and **c** (ns: not statistically significant, p-value > 0.05, **: p-value < 0.01, ***: p-value < 0.001) were calculated via two-tailed unpaired t-test.

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 (www.ensembl.org). Repetitive elements were identified using the full RepeatMasker[27] annotation of the human genome through the supplied tables of the UCSC genome browser (www.genome.ucsc.edu). Sequence-logos[28] were plotted using the *seqLogo* library of *R*[29]. Estimated ages of L1 elements were recovered from Khan et al. [25] Multiple sequence alignments were performed 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 using the in-house code written in *R* programming language[29] and available from the authors upon request. The sections that refer to the performed calculations are briefed below, with the complete details and reasoning presented in the **Supplementary Information**.

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. 6 out of the 15 most frequent 703,091 PQSs from the human genome were found to originate from L1 elements (**Supplementary Table 1**) and were used to define the L1-originated quadruplex sequence (LQS) family as follows. The most frequent LQS representative, GGGGACTGTTGTGGGGTGGGGGGAGGGGGGAGGG, referred to as LQS^{ref} and later identified to stem from the L1PA3 family, was used as a reference to reveal all related PQSs by hierarchical clustering analysis (**Supplementary Note 1**). All PQSs with lengths similar to LQS^{ref} (34 nucleotides) were directly compared to LQS^{ref}, by calculating the Hamming distance,[31] d^H, a unitless similarity measure between the strings of equal length. For a given sequence in our analysis, d^H is the number of characters that differ, in a position specific manner, from LQS^{ref}. The d^H distribution revealed a distinct family of PSQ with $d^H \leq 5$ (**Supplementary Fig. 1**) characterised by 10,269 occurrences of 3,238 unique sequences. Finally all PQSs, not restricted to 34-nt-long sequences, were screened against the identified 34-nt core to find all occurrences of the related sequences. This 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 reveal 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. For identifying 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 over 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, only the original LQSs were considered, in order to correct for the noise introduced by time-accumulated random nucleotide substitutions (**Supplementary Table 3**, **Supplementary Note 1**, **Supplementary Fig. 1e**). Multiple sequence alignment of the revealed original LQSs was used to construct a substitution-based tree, visualised *via* Dendroscope,[32] while rooted on the representative from 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 using high-throughput sequencing[16] that takes advantage of polymerase stalling at G4 sites during sequencing, causing the incorporation of identifiable mismatches. The maximum observed percentages of those mismatches (*mm%*) close to G4 motifs are reflective of the G4 stability. Human G4-seq data (GEO accession number: GSE63874) were recovered and the *mm%* values for all G4A1-5 LQSs (with 50-nt flank regions) were extracted from the K^+ vs. Na^+ dataset. The relative G4-score (**Fig. 2c**) was then calculated by dividing the obtained *mm%* values by 26.7 (the median *mm%* of the most stable G4A4 sequence, **Fig. 2b**). The significance of 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 ($pL1_{RP}\Delta$G4) or harbouring a mutated version of the sequence ($pL1_{RP}$mG4) were obtained from $pL1_{RP}$WT by standard molecular cloning procedures. In brief, inserts with the corresponding sequences

were constructed by oe-PCR using the primers L1fw and L1rv with the corresponding primers for each construct (**Supplementary Information**). The resulting inserts and the parent plasmid were digested with *SapI* (New England BioLabs) prior to 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 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 purified with a Plasmid Midi Kit (Quiagen).

Cell culture and materials. HeLa cells were grown in a humidified, 5 % CO_2 incubator at 37°C in high glucose (4.5 g·L^{-1}) Dulbecco's modified Eagle's medium supplemented with 10 % fetal bovine calf serum (DMEM complete) and split at 70-80 % confluency using trypsin EDTA. The cell line was genotyped via STR profiling and mycoplasma tested. PDS[33], PhenDC3[34] and 12459 (patent WO 0140218) were synthesised as previously described. The synthesis and characterisation of PDC12 and PDC20 is 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 6-well dishes at a density of 1×10^5 cells/well and grown at 70 % confluency in DMEM complete. Cells were transfected with the FuGENE 6 transfection reagent (Promega) following the manufacturer's protocol. Each transfection well received 2 μg plasmid DNA, 12 μL FuGENE 6 reagent and 2 mL DMEM complete. Puromycin selection (2 μg·mL^{-1}) was started 48 h after transfection. Puromycin-resistant cells were selected by growth in DMEM complete containing 2 μg·mL^{-1} puromycin. Small molecule and chemical treatments were achieved by adding the appropriate volume of DMSO stock solutions to a puromycin-containing DMEM complete solution and changing the cell media every two days. PDS and PhenDC3 were added to the cell cultures 9 days after transfection. 12459, PDC12, PDC20 and HgS were added 2 days after transfection.

Fluorescence-activated cell scanning (FACS). After 14 days of culture, the cells were prepared for FACS analysis by washing twice with PBS and then incubating for 10 min with trypsin EDTA. The suspended cells were collected by centrifugation, re-suspended in PBS supplemented with PI (10 μg·mL^{-1}) and kept on ice until FACS analysis. Cells were analysed with a FACSCalibur system (BD

Biosciences). Between 10,000 and 20,000 total events were monitored per samples. Live-dead gating was performed by excluding PI-stained cells. Living cells were analysed for fluorescence intensity using a blue argon laser (488 nm) and a filter sets (533/30 band pass). Data were analysed with the FlowJo Software.

Statistical tests. The significance of the observed G4-score reduction among different G4Ai sequences depicted in **Fig. 2c** was tested using one-tailed Mann-Whitney non-parametric test. There, the null hypothesis was that the younger L1 G4 sequences had an average G4-score not lower than the older ones (**Supplementary Note 1**). The significance of the observed differences in retrotransposition assays was assessed using two-tailed unpaired t-test. All the relevant information and the sample sizes are provided within the main text, figure captions and the supplementary information files.

Code availability. All the used computer programs and genomic databases are openly available, as detailed in the Supplementary Information and the citations within. Different, in-house *R* scripts used for the analyses, data exploration and plotting are available from the authors upon request.

Data availability. The data supporting the findings of this study are available within the paper and the supplementary information files. Source data for **Figures 1b**, **2c**, **3d**, **4b** and **4c** are provided with the paper on-line.

References (for online Methods)

27. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0* (2015).
28. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18,** 6097–6100 (1990).
29. R Core Team. R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* (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).

SUPPLEMENTARY NOTES AND TABLES

SUPPLEMENTARY NOTE 1

Further details on the performed computations. Throughout the computational part of the work, we followed a path where, to objectively define the G-quadruplex (G4) sequences that belong to the 3'-UTRs of L1 elements (L1 quadruplex sequences or LQS family), we first analysed a wide range of potential quadruplex sequences (PQSs) from the human genome. Operating on a set of PQS data, the general motif of which engulfs the G-rich sequence found in the human specific active L1 elements,[1] we could define the LQS family in an unbiased manner, leaving out the ubiquitous similarities that were to be expected between any two random G-rich sequences due to the presence of the G-tracks.

Potential quadruplex sequences from the human genome. The potential quadruplex sequences (PQSs) were localised using a regular expression search in each human nuclear chromosome. The well-characterised sequences that are capable of forming G-quadruplex (G4) structures conform the $G_{g1}N_{L1}G_{g2}N_{L2}G_{g3}N_{L3}G_{g4}$ motif,[2] where g1, g2, g3 and g4 are integer numbers equal to or greater than 3, L1, L2 and L3 are integers in between 1 and L^{max} (inclusive), N is any nitrogenous base, including guanine. In this work, the regular expression was defined with the maximum loop size 12 ($L^{max} = 12$)[3] that captures many of the observed quadruplex forming sequences found in the human genome.[4] Whenever nested PQSs were encountered, the longest sequence encompassing the constituent PQSs were retrieved. For capturing the canonical G4 sequences, the searched motif was thus $\{G_{3+}N_{1-12}\}_3G_{3+}$. This resulted in the identification of 703,091 PQSs, close to the number reported before.[3]

Similarity measure among the sequences of equal length. As a unitless similarity measure among sequences (character strings), we used the Hamming distance d^H metric.[5] d^H between two equal-sized character strings is the number of characters that differ, in a position specific manner, from each other. The $d^H = 0$ means that the compared sequences are identical. The maximum possible d^H distance is equal to the size of the examined strings (number of characters within), and shows that all the positions are different in the compared sequences.

The most abundant LQS representative. While analysing the PQS sequences originating from L1 elements (LQSs) within the context of all human PQSs defined above, we observed that the LQSs are the most abundant quadruplex sequences in the human genome. The frequency of each unique

PQS in the pool of 703,091 sequences was counted. As can be seen from **Supplementary Table 1**, where only the 15 most frequent unique PQSs are shown. The most frequent PQS is rather simple and occurs 3,484 times in the human genome. However, the 2nd, 4th, 6th, 7th, 8th and 13th most frequent unique PQSs, in the list of the most abundant 15, originate from the L1 elements, as cross checked against the RepeatMasker annotations,[6] hence belong to the LQS family. Furthermore, the summed genomic copy number of all the LQSs in the list of top 15 is 4,770. This collectively makes the LQS family be the most abundant among PQSs in the human genome, a feature that, unlike prior work,[7] became evident due to relaxing the 7-nt constraint on the allowed loop size in the PQS definition. We can also see (**Supplementary Table 1,** blue sequences) that the most frequent LQS representatives contain the following 34-nt-long core:

GGGGACTGTTGT**GGGGTGGGGGGA**GGGGGGA**GGG**

which will be denoted as LQS^{ref} hereafter. Please note, as it turns out later in the work, LQS^{ref} is the original G4A3 quadruplex sequence of the L1PA3 subfamily of L1 retrotransposons.

Identification of sequences belonging to the LQS family. Owing to the presence of G-tracks in the sequences, all PQSs have a high degree of similarity that does not stem from their shared origin. Therefore, to identify all the L1 quadruplex sequences in an unbiased manner, we looked for sequences where we could surely state the relatedness to the most abundant L1 quadruplex sequence - LQS^{ref}. We first analysed all the PQSs, with the same size (34-nt) as LQS^{ref}. For each such PQS, we calculated the d^H distance[5] from the LQS^{ref} reference with the goal that examining the d^H distribution would give us an idea about the d^H values expected from any unrelated, random, PQSs, and reveal whether we have outliers, i.e. related representatives. The d^H distribution indeed revealed a pattern with three peaks (**Supplementary Fig. 1**). The largest peak depicts d^Hs between LQS^{ref} and any unrelated 34-nt PQSs, which can still have matches with the reference sequence owing to the presence of G-tracks by definition. The middle peak is still ambiguous, since is, in part, formed by the summation of the shoulders of the first and the last peaks. Hence, to define the core of the LQS family, we considered all the sequences that comprised the first peak and were apart by not more than $d^H = 5$ from the LQS^{ref} reference sequence (red shaded area in **Supplementary Fig. 1a**). This technique identified 10,269 cases comprised of 3,238 unique sequences from the original PQS pool. The sequence-logo[8] (**Supplementary Fig. 1a**, top) reveals the relative variation of nucleobases at each position of the 34-nt-long core, demonstrating the positions where the most number of variations occur. Next, the pool of 703,091 PQSs, not restricted to only the 34-nt sequences, was screened to find all the instances where any of the 3,238 unique cores ($d^H \leq 5$) were found. This further increased the size of the related sequences (L1 quadruplex

sequences, LQS) to 15,724, forming our final unbiased set of quadruplex sequences that belong to the LQS family.

Identification of the original LQS sequences of different L1 subfamilies. Having the outcomes of the analyses described above, we revealed the PQSs and LQSs co-localised with the remnants of L1 retrotransposons by using the full RepeatMasker[6] annotation. The numbers and distribution of the L1 remnants and quadruplex sequences found in different L1 subfamilies are presented in **Supplementary Table 2** and **Fig. 2a**.

With an increase in age,[9-11] the remnants of different L1 subfamilies have accumulated an increasing amount of nucleobase substitutions. This accumulation of random substitutions makes it difficult to confidently infer the original LQS sequences at the 3'-UTR for each L1 subfamily via simple sequence alignment and subfamily-wise examination of consensus sequences. The repetitive and short nature of LQSs with an overwhelming abundance of a single base, G, deems such comparisons even more ambiguous and problematic. **Fig. 2a** also demonstrates the erosive effect of the time-accumulated substitutions on the information content of the L1 remnants, as we observe a decrease in the identifiable PQS or LQSs for the older L1 elements, despite such elements originally possessing a unique LQS.

To identify the L1 subfamilies that are not old enough for the time accumulated aberrations in their sequences to wipe out the original LQS moieties in most of their remnants, we introduced a measure that we call G4 diversity index - DI^{G4}. DI^{G4} was defined as the ratio of all the unique PQSs over the total number of PQSs found in a given population (a given L1 subfamily in our particular case). $DI^{G4} = 1$ means that all the PQSs in the population are unique, without any two sequence repeating each other, since the unique number of PQSs is equal to the number of all PQSs in the ratio that forms DI^{G4}. Smaller values of DI^{G4} depict an emergence of a preferred sequence(s), the occurrence of which is repeated multiple times. As can be seen from **Supplementary Table 2**, nearly all L1-family retrotransposons that are old and altered by time-accumulated substitution and recombination events (small n^{PQS}/n^{LINE}) have high DI^{G4}, hence we cannot robustly state anything about any specific PQS being the original one. In contrast, the hominoid primate- and human-specific L1 retrotransposons (L1PA2-5 and L1Hs subfamilies) with high PQS occurrence have rather small DI^{G4}, indicating a strong preservation of the information content pointing to the original LQS sequence in those families (**Fig. 1b**). Not surprisingly, all L1 subfamilies that possess a small DI^{G4} are the ones where we have a substantial identifiable LQS presence. **Supplementary Fig. 1e** and **Supplementary Table 3** present the sequence logo and PQS vs. LQS occurrence data for all the L1 subfamilies that have at least five LQS sequences found in their genomic population. We analysed the frequency of the individual PQSs in each subfamily, and **Supplementary Table 3**

shows the most frequent/original PQS sequences for each L1 subfamily. As can be seen from the table, wherever an LQS member is identified in a given type of mobile element, it reveals itself as the most preferred sequence contributing into the lowered DI^{G4} value, hence all the consensus sequences in **Supplementary Table 3** belong to the LQS family. The detailed sequence-logo plots, presented in **Supplementary Fig. 1e**, show the frequency of all the bases at each position of the 34-nt core of the LQS family. These frequencies also suggest that the identified L1 subfamilies, where DI^{G4} signals the preservation of the information content (**Fig. 1b**), feature one or two prevalent sequences (**Fig. 1c** and **Supplementary Table 3**), with all the other variations occurring substantially less-frequently. There are also distinct base differences between different retrotransposon types, suggesting that we should consider the reconstruction of the relatedness tree, based on just the 34-nt-long LQS core sequence.

Identifying a cascade of single-nucleotide substitutions in LQS. Since the L1 retrotransposons are largely inactive and stayed as such for a long time in the evolution of genomes,[12] they have freely accumulated different substitutions that have nothing to do with the speciation of the retrotransposons.[13] Hence, for each retrotransposon subfamily, in order to clean the data from such random mutations and focus on the relevant polymorphism, we have considered only the original LQS sequences (see above), with a marked prevalence in their copy numbers as compared to other PQSs found in the corresponding retrotransposon subfamily. The selected sequences and the number of cases are shown in **Fig. 1c**, along with the outcome of their multiple sequence alignment. The multiple sequence alignment of the selected sequences, which belong to the LQS family and were deemed prevalent in each subfamily of retrotransposons, was done using the ClustalW program.[14] The default alignment matrix for nucleic acids was used, and the individual alignment steps were iteratively refined. The tree was then constructed through a neighbour joining algorithm and visualised via Dendroscope,[15] rooted on the most distant L1PA8A subfamily. The corresponding tree is shown in **Fig. 2b**. The figure reconstructs the sub-tree of the primate-specific L1-family members, depicting the relatedness and evolutionary precedence of the subfamilies in the hominoid lineage.[9-11,13]

Importantly, we can trace the evolution and speciation of the L1PA5 up to the L1PA2 and the most recent human-specific L1Hs subfamilies as a cascade of sequential single-nucleotide substitutions. This most pronounced lineage is known to have been amplified and evolved in the hominoid primates throughout the past 25 myr.[11] It is interesting to note that along the branch revealed in our analysis several subfamilies were, most probably, occurring during the transition periods for the LQS shift (**Fig. 2b**), where they either had LQS segments of both types (before and after the substitution), or frequently transposed into the same loci, leading to the mix of 3'-end

sequences in the contemporary human genome. For example, L1PA4 (age ~18 myr[10]) has nearly equal number of instances with the LQS being in the state before (483 cases) and after (515 cases) the C5A substitution. Similarly, L1PA2 (~7.6 myr[10]) transits the T10G substitution, with 600 and 771 cases of LQS before and after the event. In contrast, the L1PA5 (~20.4 myr[10]), L1PA3 (~12.5 myr[10]) and the most recent human-specific L1Hs (~3.1 myr[10]) members have inherited a distinct state of LQS family sequence (**Fig. 2b**).

The revealed strong coupling of sub-speciation of hominoid L1 elements with the state of their respective LQS segments (**Fig. 2b)** suggests that the LQS could have an important role in the activity of L1 elements.

Determination of the G4 stabilities as reflected in G4-seq experiments. The recent development of a sequencing technique - G4-seq - that experimentally scans for G4 structures in DNA, and the availability of such experimental outcome for the entire human genome,[4] provides us with an opportunity to explore the stability change of the identified LQSs that emerge along the cascade of single-nucleotide substitutions during L1 sub-speciation. We took the data for the human genome Na^+ vs. K^+ gradient experiment (GEO accession number: GSE63874). The data contained the mismatch percentage of the base incorporation (averaged to a 15-nt-resolution bins), after the polymerase stalling occurs around quadruplex structures, for all the mapped positions in the human genome. The mismatch percentage, reflected in the G4-score (see **on-line Methods**), was shown to be a good marker of G4 structural stability (the higher is mismatch level, the more substantial is the polymerase stalls at a G4 structure), though with a large variation of values that reflect the influence of flanking sequences at around G4s.[4] We next screened all the loci in the human genome that had the exact LQSs (the sequences G4A5 to G4A1 in **Fig. 2b**) identified along the cascade of substitutions. Wherever a mismatch percentage data was found for a given LQS locus (with 50-nt flank regions), the maximum value was assigned to LQS, as an overall stability measure. We found 125, 807, 2945, 765 and 505 such instances for the G4A5, G4A4, G4A3, G4A2 and G4A1 sequences correspondingly. The distribution of the mismatch levels for each LQS along the cascade is shown in **Fig. 2c**. The results demonstrate a gradual decrease of the G4 stability coupled with the emergence of younger L1 subfamilies. To take into account the difference in the number of overlaps found for each LQS families, we quantified the significance of the noted shifts in G4-scores using one-tailed Mann-Whitney nonparametric test (**Fig. 2c**). The latter test has a greater efficiency than the standard t-test for the distributions that deviate from normal, and has efficiency similar to the t-test for normal distributions. The null hypothesis in the test, while comparing two adjacent distributions (such as the ones for G4A2 vs. G4A1, **Fig. 2c**) was that the average G4-score from the distribution in younger L1 subfamily (G4A1) is not smaller than the same value in older

L1 subfamily (G4A2). Such quantifications of the significance produced 7.73×10^{-1}, 7.68×10^{-16}, 2.96×10^{-15} and 2.49×10^{-4} p-values while comparing the G4-score distributions in {G4A5, G4A4}, {G4A4, G4A3}, {G4A3, G4A2} and {G4A2, G4A1} pairs, where the comparison of only the first pair gave insignificant outcome due to the small number of G4A5 overlaps found in the G4-seq data (only 125).

Overall, here we found that the identified cascade of single-nucleotide substitutions in LQSs gradually decreased the G-quadruplex structural stability in the human DNA. This observation, along with the revealed coupling of the substitution cascade with young L1 sub-speciation, presents a significant computational evidence for the presence of a role that the LQS plays at 3'-UTRs of L1 elements through the formation of G-quadruplex structures exclusively. As can be seen from the experiments below, we found that the stabilised quadruplex structures stimulate the L1 mobility. This sheds further light on the decrease of the native G4 stability during the host DNA-L1 co-evolution. The evolution proceeds to improve the fitness of the host organism, hence the L1 elements that pass through the host generations give rise to subfamilies that are relatively less mobile, hence more tolerable for the host organism. The latter can also be judged from the simple number of genomic copy numbers (n^{LINE}) of the young L1 subfamilies (**Supplementary Table 3**), where the older hominoid-specific L1 elements exerted much more violent amplification (higher n^{LINE}) than the younger ones.

G-rich quadruplex forming sequence of the active L1Hs retrotransposons (G4A1). The inferred 34-nt long LQS sequence (G4A1, **Fig. 2b**) at the 3'-UTR of the human-specific L1 retrotransposon is spanning the positions 5859-5892 of the 6064-nt-long sequence of complete L1Hs element:

GGGGACTGTGGT**GGGG**TC**GGGGG**A**GGGGGG**A**GGG**

The high degree of conservation and completeness of these youngest representatives of L1 retrotransposons in the human genome allows the inference of its original G-rich sequence via a mere examination of the consensus sequence of all the L1Hs remnants. Looking up all the reportedly active human retrotransposons,[1] we verified the presence of G4A1 in them.

The G4A1 sequence features five G-tracks. Since four G-tracks are needed for G4 structure formation, G4A1 harbours multiple possibilities for quadruplex formation. However, the prior characterisation of the quadruplex forming propensity omitted either the last[16] or the first[17] G-track, and no prior study demonstrated the importance of the G-rich trail in the retrotransposition efficiency (see below).

SUPPLEMENTARY NOTE 2

Synthesis of PDC12 and PDC20. Chemicals were purchased from Sigma-Aldrich. NMR spectra were acquired on a Bruker DRX-400 instrument at ambient probe temperature (300 K). Notation for the ^{1}H NMR spectral splitting patterns includes: singlet (s), doublet (d), triplet (t), broad (br) and multiplet/overlapping peaks (m). Signals are quoted as d values in ppm. Mass spectra were recorded on a Micromass Q-Tof (ESI) spectrometer.

PDC12: 2,6-Pyridinedicarbonyl dichloride (102 mg, 0.5 mmol) was combined with 4-aminopyridine (94 mg, 1 mmol). Dichloromethane (5 ml) followed by triethylamine (140 µl, 1 mmol) were added and the mixture was stirred at room temperature overnight. The resulting precipitate was collected by filtration and washed with dichloromethane (20 ml) before drying under vacuum. Yield: (0.029 g, 18.2 %); Calculated mass: 320.1148 g·mol^{-1}, Experimental mass: 320.1136 g·mol^{-1}; ^{1}H NMR (400 MHz, DMSO-d6): d 11.33 (s, 2H), 8.63 (m, 2H), 8.45 (m, 2H), 8.36 (m, H), 8.12 (m, 6H); ^{13}C NMR (100 MHz, DMSO-d6): d 162.8, 159.8, 150.5, 148.5, 145.2, 140.3, 126.2, 114.6, 108.8.

PDC20: 2,6-Pyridinedicarbonyl dichloride (102 mg, 0.5 mmol) was combined with 4-amino-3-iodobenzonitrile (244 mg, 1 mmol). Dichloromethane (5 ml) followed by triethylamine (140 µl, 1 mmol) were added and the mixture was stirred for 2 minutes before leaving to react in a CEM Discover microwave at 150 °C, 300W for 20 minutes. The reaction mixture was cooled and the resulting precipitate collected by filtration. The crude product was washed with dichloromethane (20 ml) and purified by column chromatography using methanol 10 %: chloroform 90 %: TEA 0.5 % as the eluent. Yield: (0.045 g, 14.5 %); Calculated mass: 619.9036 g·mol^{-1}, Experimental mass: 619.8421 g·mol^{-1}; ^{1}H NMR (400 MHz, DMSO-d6): d 10.09 (s, 2H), 8.42 (m, 5H), 7.96 (m, 2H), 7.88 (m, 2H); ^{13}C NMR (100 MHz, DMSO-d6): d 161.7, 148.1, 142.5, 140.7, 132.9, 126.9, 126.0, 118.4, 110.3, 97.4.

UV spectroscopy. UV melting curves (followed at 295 nm) were collected with a Varian Cary 400 Scan UV-visible spectrophotometer. Oligonucleotide solutions were prepared at final concentrations of 4 μM in 10 mM lithium cacodylate, EDTA 1mM, 1-10 mM KCl, pH 7.2. The samples were annealed by heating to 95 °C for 10 min and were then slowly cooled to room temperature. Each sample was transferred to a quartz cuvette with a 1 cm path length, covered with

a layer of mineral oil, placed in the spectrophotometer and equilibrated at 5 °C for 10 min. Samples were then heated to 95 °C at a rate of 1 °C·min^{-1}, with data collection every 1 °C. Melting temperature ($T_{1/2}$) values are the temperature at which 50 % of the quadruplex structures are unfolded.

Circular dichroism (CD) spectroscopy. CD spectroscopy experiments were conducted on a Chirascan Plus spectropolarimeter with a quartz cuvette with an optical path length of 1 mm. Oligonucleotide solutions were prepared at a final concentration of 1.5 to 10 μM in 10 mM lithium cacodylate, EDTA 1 mM, 1-100 mM KCl, pH 7.2. The samples were annealed by heating at 95 °C for 10 min and were slowly cooled to room temperature. Scans were performed over the range of 200-320 nm. Each trace is the result of the average of three scans taken with a step size of 1 nm, a time per point of 1 s and a bandwidth of 1 nm. A blank sample containing only buffer was treated in the same manner and subtracted from the collected data. The data were finally baseline corrected at 320 nm. Temperature denaturation studies followed by CD spectroscopy were conducted similarly in the presence of small molecules directly added to the samples from DMSO stock solutions.

NMR spectroscopy. ^{1}H NMR spectra were recorded at 298 K using a 500-MHz Bruker Avance 500 TCI spectrometer equipped with a cryogenic TCI ATM probe. Water suppression was achieved using excitation sculpting. The oligonucleotides were annealed in 10 mM PBS (pH 7.0) supplemented with 100 mM KCl and 10 % D_2O at a final concentration of 0.1 mM. The samples were annealed by heating at 95 °C for 10 min and slowly cooled to room temperature.

Fluorescence spectroscopy (FRET melting). Fluorescence experiments were conducted on a Varian Cary Eclipse spectropolarimeter using a quartz cuvette with an optical path length of 1 cm. A dual-labeled L1HS-DNA oligonucleotide, referred to as Fl-L1HS-DNA-G4, was used. The donor fluorophore was 6-carboxyfluorescein (FAM), and the acceptor fluorophore was 6-carboxytetramethylrhodamine (TAMRA). Fl-L1HS-DNA-G4 was annealed at a final concentration of 100 nM in 10 mM lithium cacodylate, EDTA 1mM, 5 mM KCl, pH 7.2. Samples were excited at 494 nm, and the emission at 580 nm. Small molecules were added to the samples from DMSO stock solutions.

Luminescent cell viability assay. GI_{50} values of growth inhibition were determined using the cell viability assay CellTiter-GloTM (Promega). Cells were plated in 96 well plates at a density of 1500 cells per well in 100 μL of media and incubated for 24 h. Compounds were added in serial dilutions (in a range between 0 – 320 μM) at a volume of 100 μl per well at the respective concentrations.

Cell viability was measured after 96 h using the manufacturer's protocol. All measurements were made in triplicate.

SEQUENCE OF OLIGONUCLEOTIDES USED IN THIS STUDY

Cloning of recombinant DNA plasmids:

Name	Sequence (5' to 3')
L1_fw	CAG GAA ATA CAG AGA ACG CC
L1_rv	CCG CTA TCA GGA CAT AGC G
L1_dG4_fw	GGA CAC AGG AAG GGG AAT ATC ACA CTC TAT AGC ATT GGG AGA TAT ACC TAA TGC
L1_dG4_rv	AGA GTG TGA TAT TCC CCT TCC TGT GTC C
L1_mG4_fw	GAA GAC TGT GGT GAA GTC GAG AGA GAG AGA AGA GAT AGC ATT GGG AGA TAT ACC TAA TGC
L1_mG4_rv	CTC TTC TCT CTC TCT CGA CTT CAC CAC AGT CTT CAG AGT GTG ATA TAT CCC CTT CCT GTG TCC
L1_seq	GAG TTC ATA TCC TTT GTA GGG

Biophysical studies:

Name	Sequence (5' to 3')
L1HS-DNA-G4	d(GGGGACTGTGGTGGGGTCGGGGGAGGGGGGAGGG)
L1HS-RNA-G4	r(GGGGACUGUGGUGGGGUCGGGGGAGGGGGGAGGG)
Fl-L1HS-DNA-G4	FAM-d(GGGGACTGTGGTGGGGTCGGGGGAGGGGGGAGGG)-TAMRA
dsDNA	d(GCATAGTGCGTGGCGTTTAGC)

SUPPLEMENTARY REFERENCES

1. Brouha, B. *et al*. Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. USA* **100,** 5280–5285 (2003).
2. Huppert, J. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucl. Acids Res*. **33,** 2908–2916 (2005).
3. Maizels, N. & Gray, L. T. The G4 genome. *PLoS Genet*. **9,** e1003468 (2013).
4. Chambers, V. S. *et al*. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotech*. **33,** 877–881 (2015).
5. Hamming, R. W. Error detecting and error correcting codes. *Bell Syst. Tech. J*. **29,** 147–160 (1950).
6. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0,* 2013-2015 at http://www.repeatmasker.org
7. Todd, A. K., Johnston, M. & Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucl. Acids Res*. **33,** 2901–2907 (2005).
8. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res*. **18,** 6097–6100 (1990).
9. Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol*. **17,** 915–928 (2000).

10. Khan, H. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res*. **16,** 78–87 (2006).
11. Lee, J. *et al*. Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* **390,** 18–27 (2007).
12. Kazazian, H. H. Mobile elements: drivers of genome evolution. *Science* **303,** 1626–1632 (2004).
13. Giordano, J. *et al*. Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLoS Comput*. *Biol*. **3,** e137 (2007).
14. Larkin, M. A. *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
15. Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst*. *Biol*. **61,** 1061–1067 (2012).
16. 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).
17. Lexa, M. *et al*. Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics* **15,** (2014).

SUPPLEMENTARY TABLES

Rank	PQS sequence	n^{PQS}
1	**GGG** A **GGG** AGGT **GGG** G **GGG**	3484
2	**GGGG** ACTGTTGT **GGGG** TGG **GGGG** AGG **GGGG** AGGG	**1560**
3	**GGG** A **GGG** A **GGG** A **GGG**	1475
4	**GGGG** ACTGTTGT **GGGG** TGG **GGGG** AGG **GGGG** AGGGATAGCATTGGG	**943**
5	**GGG** A **GGG** AGGT **GGG** GG **GGG**	843
6	**GGGG** ACTGTGGT **GGGG** TCG **GGGG** AGG **GGGG** AGGGATAGCATTGGG	**701**
7	**GGGG** CCTGTTGT **GGGG** TGG **GGGG** AGG **GGGG** AGGG	**698**
8	**GGGG** ACTGTGGT **GGGG** TGG **GGGG** AGG **GGGG** AGGGATAGCATTGGG	**613**
9	**GGG** A **GGG** A **GGG** A GGG A**GGG**	596
10	**GGG** T **GGG** CGT **GGG** CTTGGC **GGG**	327
11	**GGGG** TGG **GGGG** AGG **GGGG** A **GGG**N	313
12	**GGG** GGATTTGGCA **GGG** TCAT **GGG** ACAATAGTGGA **GGG**	293
13	**GGG** CCTGTTGT **GGGG** TGG **GGGG** AGG **GGGG** AGGG	**255**
14	**GGG** A **GGG** A **GGG** A **GGG** AGGGAGGG	236
15	**GGG** A **GGG** AGGT **GGG** GGG **GGG**	233

Supplementary Table 1 | The 15 most frequent PQSs in the human genome. n^{PQS} is their genomic copy numbers. The shaded rows outline the PQSs that originate from L1 retrotransposons (LQS). The blue sequences are the ones with the LQS^{ref} core.

L1 types	n^{LINE}	n^{PQS}	n^{PQS}/n^{LINE}	DI^{G4}	n^{LQS}
HAL1	26967	1312	0.05	0.98	-
HAL1-2a_MD	1483	48	0.03	0.97	-
HAL1-3A_ME	1901	1101	**0.58**	0.94	-
HAL1b	4807	77	0.02	0.96	-
HAL1N1_MD	2	0	0.00	-	-
L1HS	1528	**954**	**0.62**	**0.17**	**733**
L1M	974	1	0.00	1.00	-
L1M1	9362	429	0.05	0.95	-
L1M2	8981	339	0.04	0.99	-
L1M2a	534	56	0.10	1.00	-
L1M2a1	124	32	0.26	1.00	-
L1M2b	240	40	0.17	0.89	-
L1M2c	431	90	0.21	1.00	-
L1M3	6751	25	0.00	1.00	-
L1M3a	625	39	0.06	0.97	-
L1M3b	553	49	0.09	0.98	-
L1M3c	1098	60	0.05	0.98	-
L1M3d	537	7	0.01	1.00	-
L1M3de	462	9	0.02	0.78	-
L1M3e	779	154	0.20	0.99	-
L1M3f	696	34	0.05	1.00	-
L1M4	17916	165	0.01	0.92	-
L1M4b	6355	174	0.03	0.97	-
L1M4c	6096	114	0.02	0.98	-
L1M5	63753	282	0.00	0.94	-
L1M6	6882	49	0.01	1.00	-
L1M7	4341	45	0.01	0.96	-
L1MA1	4220	1014	0.24	0.98	1
L1MA10	5013	241	0.05	0.99	-
L1MA2	7478	1684	0.23	0.98	1
L1MA3	8904	1614	0.18	0.98	1
L1MA4	9955	801	0.08	0.95	1
L1MA4A	6165	768	0.12	0.98	1
L1MA5	4432	420	0.09	0.99	-
L1MA5A	3406	429	0.13	1.00	-
L1MA6	5410	448	0.08	0.99	1

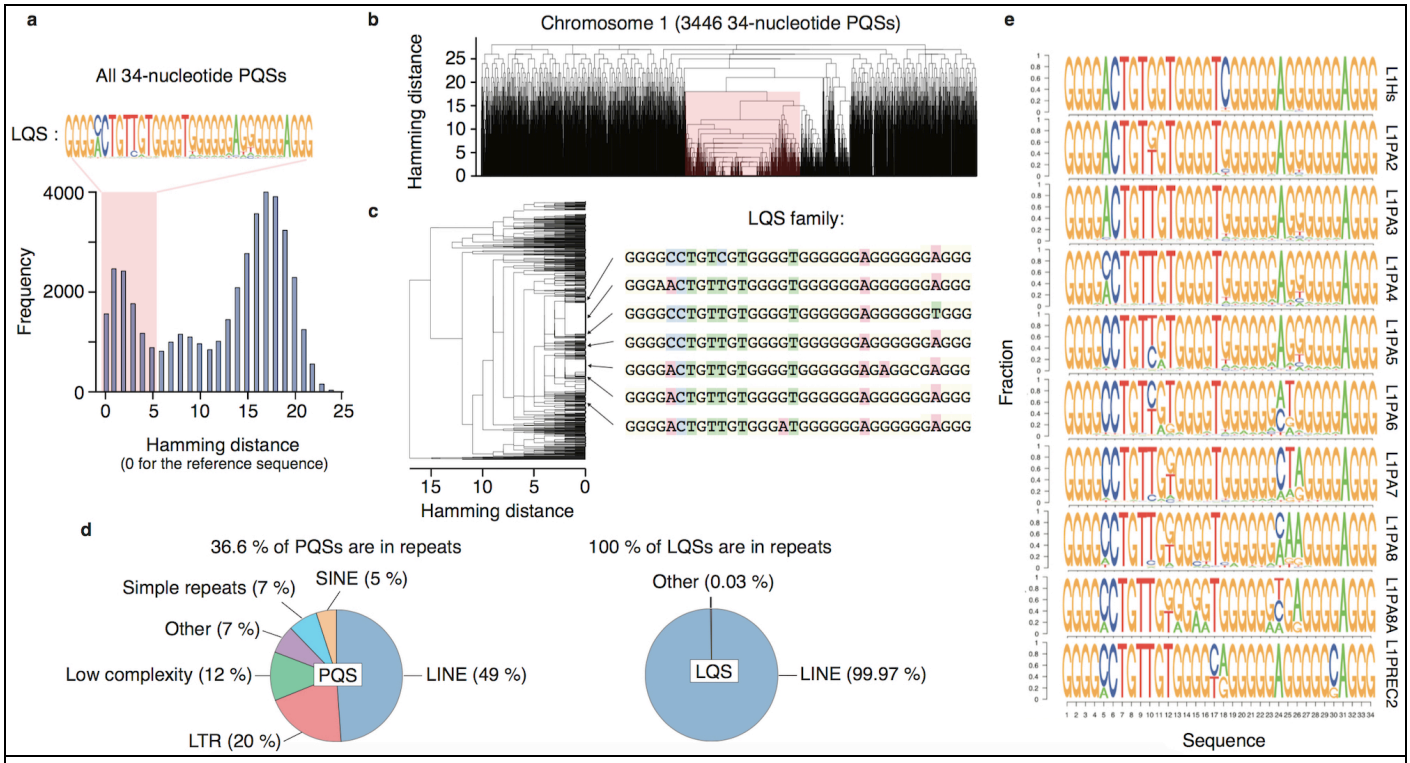
L1MA7	8439	677	0.08	0.98	-
L1MA8	10931	800	0.07	0.99	1
L1MA9	16569	1181	0.07	0.97	-
L1MB1	6141	348	0.06	0.97	-
L1MB2	8901	753	0.08	0.97	2
L1MB3	17156	1967	0.11	0.97	-
L1MB4	9161	529	0.06	0.97	3
L1MB5	9758	975	0.10	0.97	-
L1MB7	22924	1801	0.08	0.98	-
L1MB8	16461	1449	0.09	0.97	-
L1MC	11821	34	0.00	1.00	-
L1MC1	13054	1112	0.09	0.97	1
L1MC2	6341	727	0.11	0.98	-
L1MC3	13238	477	0.04	0.97	-
L1MC4	29055	471	0.02	0.97	-
L1MC4a	27328	492	0.02	0.98	-
L1MC5	20356	411	0.02	0.99	-
L1MCa	7482	168	0.02	0.97	-
L1MCb	2139	64	0.03	0.97	-
L1MCc	3390	90	0.03	0.99	-
L1MD	8612	35	0.00	1.00	-
L1MD1	6800	494	0.07	0.98	-
L1MD2	10973	833	0.08	0.98	1
L1MD3	5276	160	0.03	0.98	-
L1MDa	7024	218	0.03	0.94	1
L1MDb	1070	61	0.06	0.96	-
L1ME1	31502	1785	0.06	0.98	1
L1ME2	12862	451	0.04	0.99	-
L1ME2z	7604	286	0.04	0.99	-
L1ME3	9252	435	0.05	0.99	-
L1ME3A	15948	700	0.04	0.99	-
L1ME3B	8253	330	0.04	0.99	-
L1ME3C	12438	395	0.03	0.96	-
L1ME3D	4507	144	0.03	0.96	-
L1ME3E	4922	203	0.04	0.95	-
L1ME3F	4876	149	0.03	1.00	-
L1ME4a	42873	183	0.00	0.98	-
L1ME5	3384	199	0.06	0.99	-
L1MEa	369	7	0.02	1.00	-
L1MEb	1811	36	0.02	1.00	-
L1MEc	18607	217	0.01	0.96	1
L1MEd	10565	125	0.01	0.97	-
L1MEe	11783	66	0.01	0.98	-
L1MEf	13932	162	0.01	0.99	-
L1MEg	17926	152	0.01	0.93	-
L1MEg1	911	12	0.01	1.00	-
L1MEg2	761	5	0.01	1.00	-
L1P	159	0	0.00	-	-
L1P1	3085	31	0.01	0.84	-
L1P2	1524	102	0.07	0.94	-
L1P3	3327	247	0.07	0.76	-
L1P3b	82	7	0.09	1.00	-
L1P4	3655	13	0.00	1.00	-
L1P4a	560	93	0.17	0.94	-
L1P4b	149	47	0.32	0.79	-
L1P4c	45	8	0.18	1.00	-
L1P4d	160	36	0.23	0.93	-
L1P4e	192	17	0.09	1.00	-
L1P5	682	3	0.00	1.00	-
L1PA10	7044	2312	0.33	0.98	3
L1PA11	4107	1205	0.29	0.98	2
L1PA12	1764	375	0.21	0.91	1
L1PA13	8901	2406	0.27	0.88	-

L1PA14	3040	752	0.25	0.87	-
L1PA15	8251	1541	0.19	0.86	-
L1PA15-16	1368	131	0.10	0.96	-
L1PA16	13927	3030	0.22	0.96	1
L1PA17	4816	974	0.20	0.97	-
L1PA2	4867	3525	**0.72**	**0.35**	**2338**
L1PA3	10565	7577	**0.72**	**0.38**	**5114**
L1PA4	11763	7993	**0.68**	**0.5**	**4018**
L1PA5	11171	7244	**0.65**	**0.66**	**2831**
L1PA6	5849	3483	**0.60**	0.89	452
L1PA7	12897	8464	**0.66**	0.86	137
L1PA8	7998	4165	**0.52**	0.93	61
L1PA8A	2466	1098	**0.45**	0.99	5
L1PB	1752	22	0.01	0.90	-
L1PB1	13229	5055	0.38	0.93	-
L1PB2	2828	943	0.33	0.99	-
L1PB3	3578	561	0.16	0.98	1
L1PB4	7428	1189	0.16	0.93	-
L1PBa	2184	215	0.10	0.96	-
L1PBa1	404	85	0.21	1.00	-
L1PBb	264	33	0.13	1.00	-
L1PREC2	7756	1157	0.15	0.88	5

Supplementary Table 2 | Occurrence of quadruplex sequences in different L1 subfamilies. n^{LINE} is the number of different L1 mobile elements in the human genome identified via the RepeatMasker database, n^{PQS} the number of the potential quadruplex sequences found in those elements, the n^{PQS}/n^{LINE} ratio represents the fraction of the remnants where PQS are preserved, DI^{G4} the G4 diversity index and n^{LQS} the number of LQS sequences identified to be part of the same L1-originated family of quadruplexes. DI^{G4} is defined as the ratio of unique PQS and all the PQSs found in a given L1 subfamily. A DI^{G4} value of 1 depicts the maximum diversity, where all the identified PQSs in a given L1 subfamily are unique. Values less than 1 indicate the presence of favoured PQSs. The L1 subfamilies are ordered in an alphabetical order. The shaded rows outline the entries with high PQS and LQS contents.

L1 type	n^{LINE}	n^{PQS}	n^{PQS}/n^{LINE}	DI^{G4}	n^{LQS}	n^{LQS}/n^{PQS} %	Consensus sequence
L1HS	**1528**	**954**	**0.62**	**0.17**	**733**	**76.83**	GGGG A CTGT G G T GGGG T C GGGGG A G G GGG G AGGG
L1PA2	**4867**	**3525**	**0.72**	**0.35**	**2338**	**66.33**	GGGG A CTGT[G|T]G T GGGG T G GGGGG A G G GGG G AGGG
L1PA3	**10565**	**7577**	**0.72**	**0.38**	**5114**	**67.49**	GGGG A CTGT T G T GGGG T G GGGGG A G G GGG G AGGG
L1PA4	**11763**	**7993**	**0.68**	**0.50**	**4018**	**50.27**	GGGG[C|A]CTGT T G T GGGG T G GGGGG A G G GGG G AGGG
L1PA5	**11171**	**7244**	**0.65**	**0.66**	**2831**	**39.08**	GGGG C CTGT[T|C]G T GGGG T G GGGGG A G G GGG G AGGG
L1PA6	*5849*	*3483*	*0.60*	*0.89*	*452*	*12.98*	GGGG C CTGT[C|T]G T GGGG T G GGGGG[A|C]T G GGG G AGGG
L1PA7	*12897*	*8464*	*0.66*	*0.86*	*137*	*1.62*	GGGG C CTGT T G[G|T]GGGG T G GGGGG C T A GGG G AGGG
L1PA8	*7998*	*4165*	*0.52*	*0.93*	*61*	*1.46*	GGGG C CTGT T G[G|T]GGGG T G GGGGG[C|A]A A GGG G AGGG
L1PA8A	*2466*	*1098*	*0.45*	*0.99*	*5*	*0.46*	GGGG C CTGT T G[G|T]GGGG T G GGGGG[TCA]G A GGG G AGGG
L1PREC 2	*7756*	*1157*	*0.15*	*0.88*	*5*	*0.43*	GGGG C CTGT T G T GGGG[C|T][A|G]GGGGG A G G GGG C AGGG

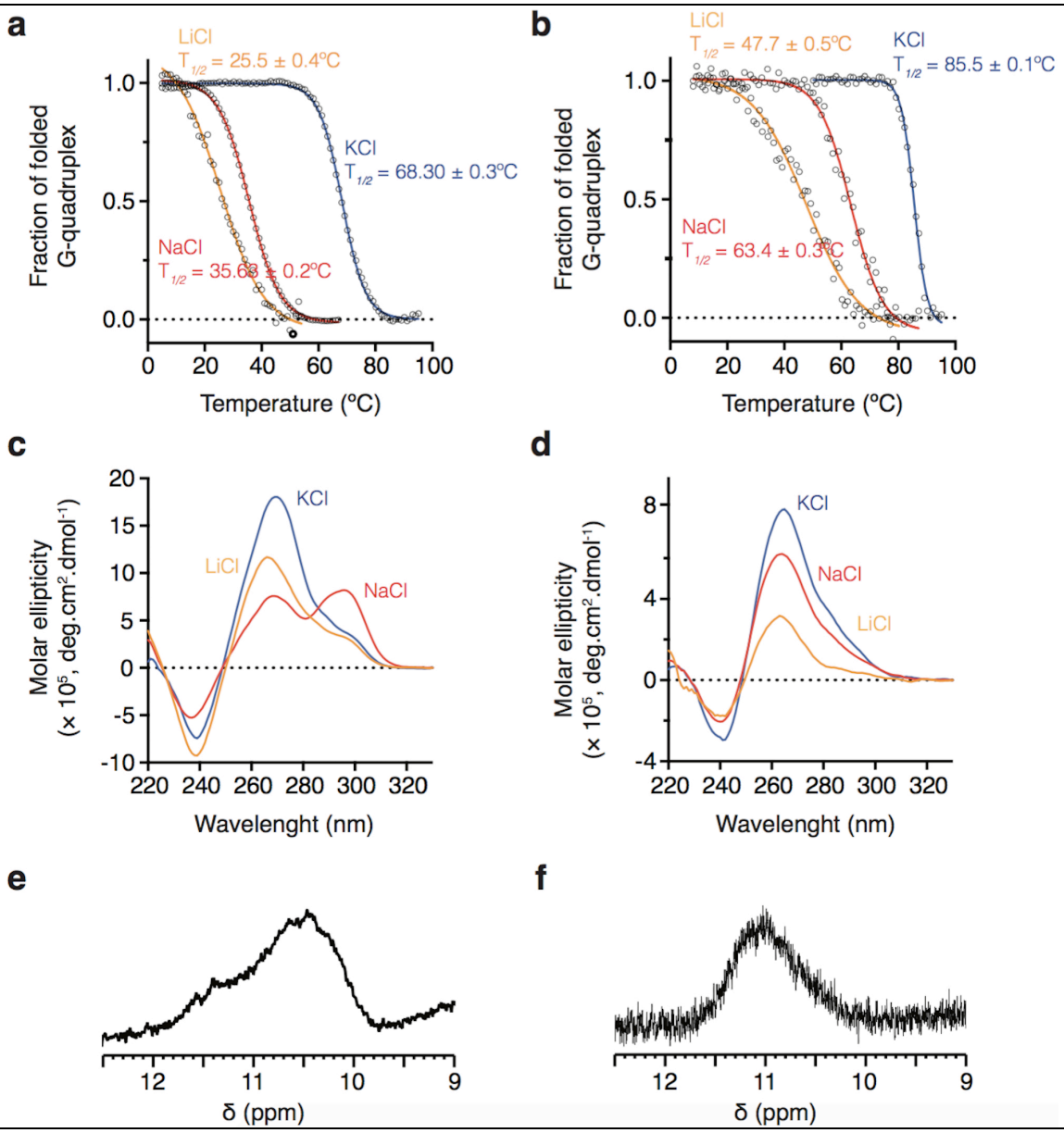
Supplementary Table 3 | Selected set of L1-family retrotransposons in the human genome with at least five LQS identified in their remnants. n^{LINE} is the genomic copy number, n^{PQS} the number of sequences that contained a potential quadruplex sequence, DI^{G4} the G4 diversity index for the PQS that occur in the corresponding retrotransposon, n^{LQS} the number of elements containing a LQS sequence, together with the consensus sequences of the most frequent PQSs. The consensus sequences are coloured at the positions where polymorphism occurs. If a position has two equally frequent variants, two bases are indicated with [$base^a$|$base^b$] notation, where $base^a$ and $base^b$ are the two most frequent variants with $base^a$ having a slight prevalence in frequency. The shaded rows outline the entries with high PQS and LQS contents, as inferred from the n^{PQS}/n^{LINE} and n^{LQS}/n^{PQS} ratios.



Supplementary Figure 1

34-nucleotide-long potential quadruplex sequences (PQSs) used for defining the L1-originated quadruplex sequence (LQS) family.

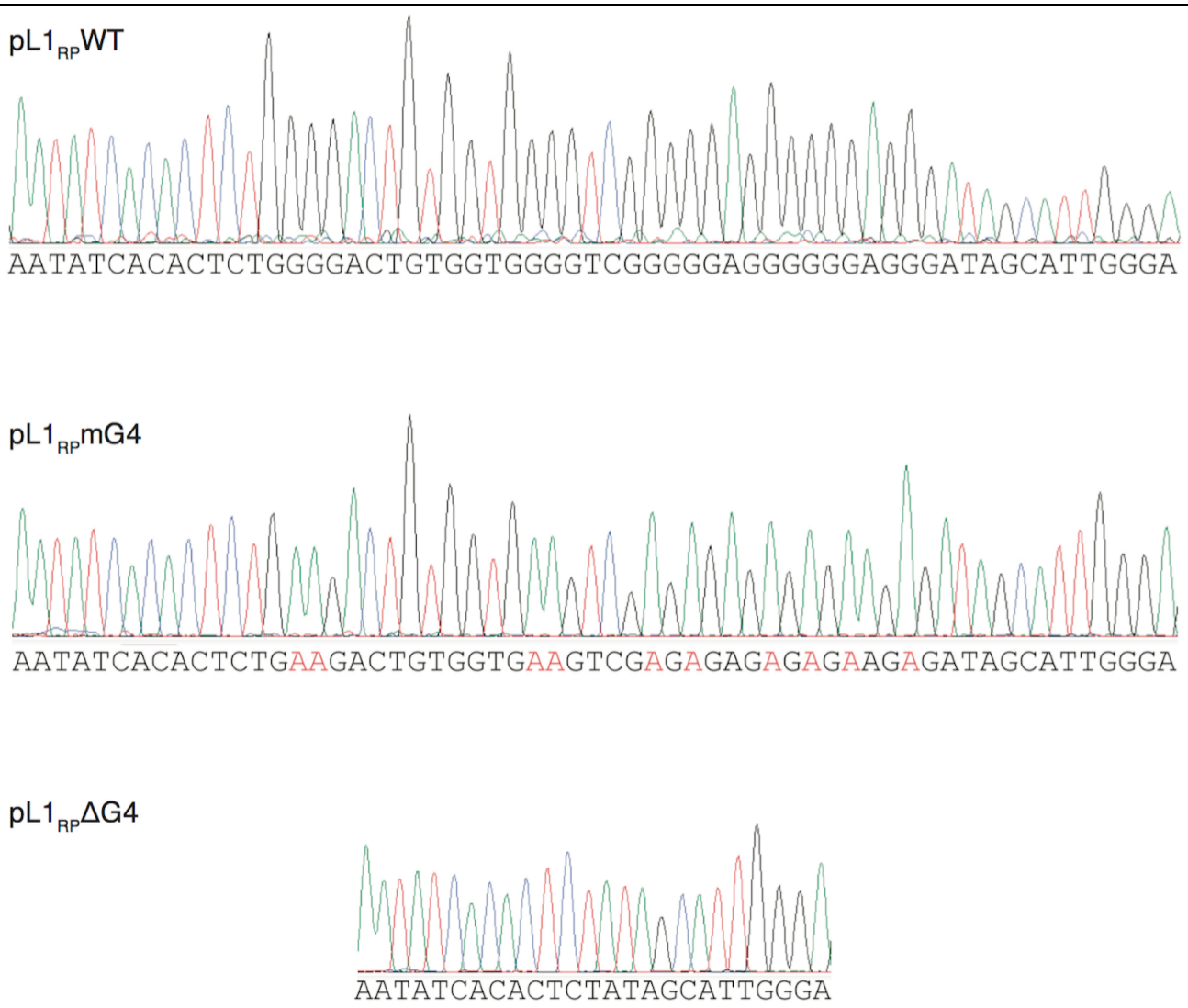
(**a**) Distribution of the Hamming distance of all 34-nt-long PQSs referenced against the most frequent PQS from L1 retrotransposons - LQS^{ref}. The peak with [0,5] borders, highlighted in red, defines the LQS family. The consensus sequence of the LQS family is illustrated via a sequence-logo plot. (**b**) Hierarchical clustering of all the 34-nt-long PQSs in the chromosome 1 with the Hamming distance applied as a similarity metrics. (**c**) Sub-trees depicting the region encompassing the LQS family (red box on panel **b**) together with some selected examples. (**d**) Genomic localisation of repeat-associated PQSs and LQSs. It is noteworthy that almost all LQS sequences are found within the 3'-UTR of L1 elements. (**e**) Sequence composition of the LQS family of quadruplexes in different L1 subfamilies. Sequence-logo plots in **e** show the base frequencies at each LQS position in the retrotransposon remnants of a given subfamily. The analysed subfamilies contain at least 5 conserved quadruplex sequences belonging to the LQS family.



Supplementary Figure 2

The G-rich sequence found within the 3'-UTR of L1Hs fold into a G-quadruplex motif *in vitro* at both the DNA and RNA levels.

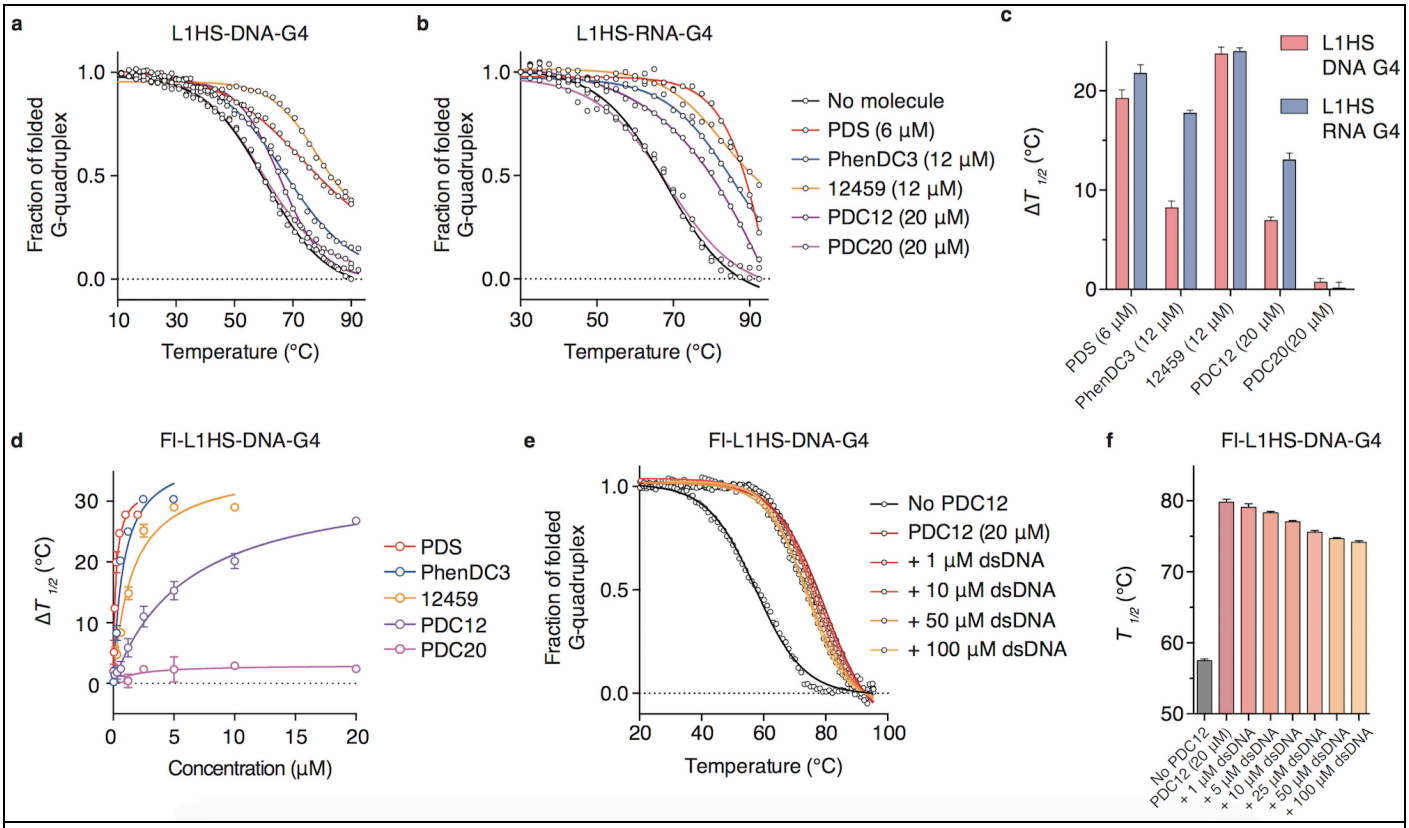
UV-melting profile followed at 295 nm of the L1Hs-DNA-G4 (**a**) and L1Hs-RNA-G4 (**b**) in the presence of 10 mM LiCl (orange line), NaCl (red line) or KCl (blue line). A cation-dependent hypochromic transition is characteristic of G4 formation. Circular dichroism (CD) spectra of the L1Hs-DNA-G4 (**c**) and L1Hs-RNA-G4 (**d**) in the presence of 100 mM LiCl (orange line), NaCl (red line) or KCl (blue line). All CD spectra are characterised by a minimum at 240 nm and a maximum at 263 nm characteristic of G4 formation. Expansion of the ^{1}H NMR spectra of L1Hs-DNA-G4 (**e**) and L1Hs-RNA-G4 (**f**) pre-annealed in a 100 mM KCl containing buffer. Both NMR spectra exhibit characteristic imino proton signals shifted downfield (between 10 to 12 ppm) characteristic of Hoogsteen hydrogen bonding and G4 formation.



Supplementary Figure 3

Sanger sequencing traces for $L1_{RP}W$, $L1_{RP}mG4$ and $L1_{RP}\Delta G4$ sequences.

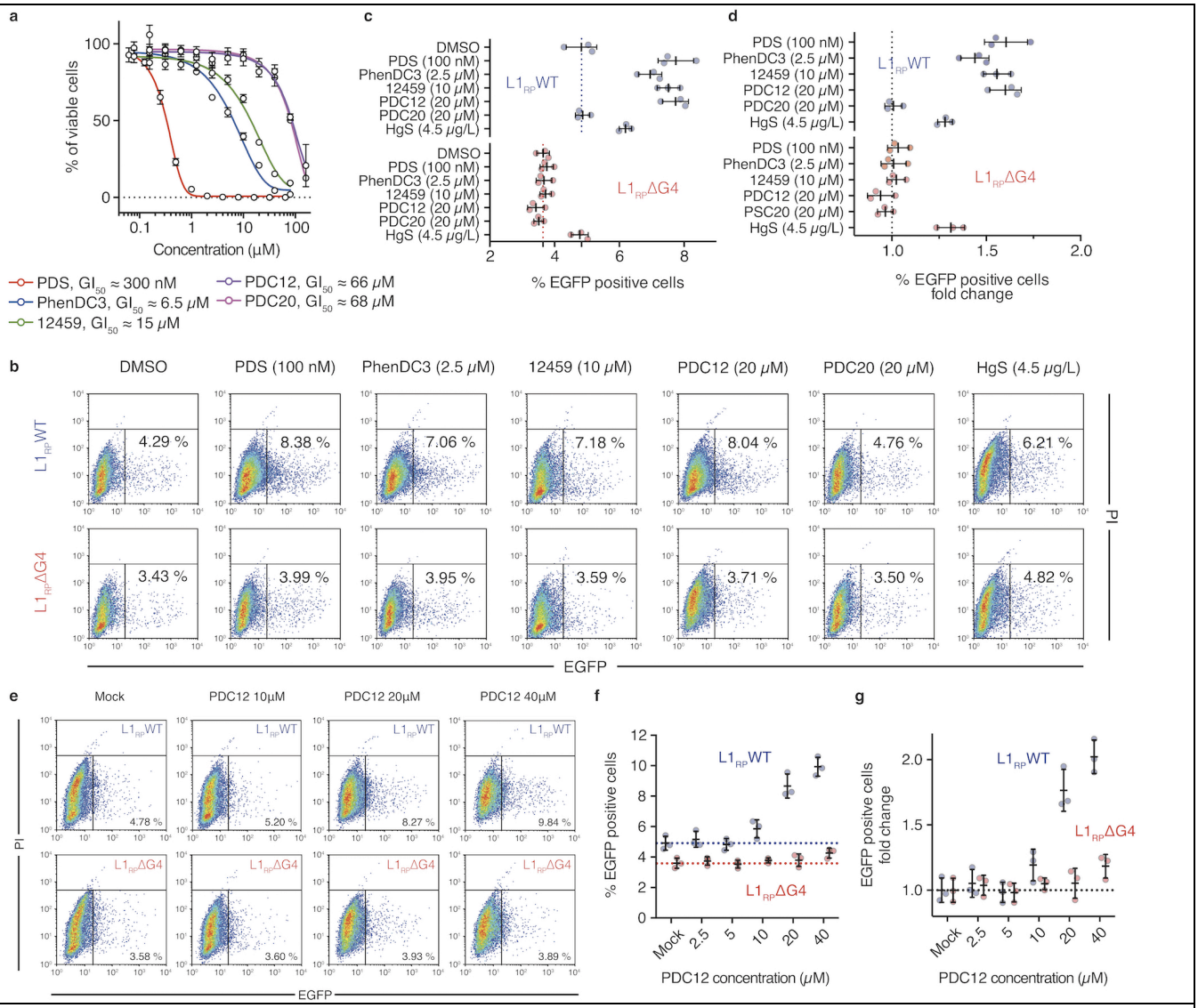
The quadruplex motif found in the plasmid $pL1_{RP}WT$ (Addgene EF06R) was either mutated ($pL1_{RP}mG4$) or deleted ($pL1_{RP}\Delta G4$) by oe-PCR. Clones containing the desired plasmids were identified by Sanger sequencing.



Supplementary Figure 4

Biophysical characterisation of the interaction of small-molecule G4-ligands with the DNA and RNA L1Hs quadruplexes.

Thermal denaturation profiles of L1HS-DNA-G4 (**a**) and L1HS-RNA-G4 (**b**) in the presence or absence of the small molecules PDS, PhenDC3, 12459, PDC12 and PDC20 followed by CD spectroscopy at 263 nm. The nucleic acids were annealed at 1.5 µM in a 5 mM and 1mM KCl containing buffer for the DNA or RNA quadruplex respectively. (**c**) Extracted melting temperatures ($T_{1/2}$) from the CD melting experiments (data represent mean values ± s.d.; n = 3 technical replicates, i.e. independent melting experiments). (**d**) Concentration-dependent stabilisation effect of the small molecules on Fl-L1HS-DNA-G4 (a dual fluorescently labelled L1HS DNA quadruplex) followed by FRET melting experiments (data represent mean values ± s.d.; n = 3 technical replicates). (**e**-**f**) Thermal denaturation profiles and extracted melting temperatures ($T_{1/2}$) of Fl-L1Hs-DNA-G4 in the presence of PDC12 and an increasing concentration of a double-stranded DNA competitor (dsDNA). Fl-L1Hs-DNA-G4 was annealed at 100 nM in a 5 mM KCl containing buffer. All molecules, except PDC20, were found to stabilise both the DNA and RNA L1Hs quadruplexes. The newly reported compound PDC12 was found to be selective for the L1Hs-DNA quadruplex over double-stranded DNA (data in **f** represent mean values ± s.d.; n = 3 technical replicates).



Supplementary Figure 5

Small-molecule G4-ligands stimulate retrotransposition of the human $L1_{RP}$ element in HeLa cells.

(**a**) Growth inhibition properties of the small molecules. GI_{50} values of growth inhibition were determined using the cell viability assay CellTiter-GloTM after 4 days of incubation (data represent mean values ± s.d.; n = 3 independent cell cultures). (**b**) Representative FACS profiles of HeLa cells transfected with either pL1$_{RP}$WT or pL1$_{RP}$ΔG4 and treated with the different quadruplex ligands PDS, PhenDC3, 12459 and PDC12. PDC20 and HgS were used as negative and positive controls respectively. (**c**) Percentage of EGFP positive cells for the different treatments. The blue and red lines report the mean values of the percentages of EGFP positive cells transfected with pL1$_{RP}$WT or pL1$_{RP}$ΔG4 mock treated with DMSO. (**d**) Percentage of EGFP positive cells fold changes compared to mock treatment. (**e**-**g**) The small molecule PDC12 stimulates the retrotransposition of the human $L1_{RP}$ element in HeLa cells in a concentration dependent manner. (**e**) Representative FACS profiles of HeLa cells transfected with either L1$_{RP}$WT or L1$_{RP}$ΔG4 and treated with an increasing concentration of PDC12. (**f**) Percentage of EGFP positive cells for different doses of PDC12. The blue and red lines report the mean values of the percentages of EGFP positive cells transfected with L1$_{RP}$WT or L1$_{RP}$ΔG4 in the absence of PDC12 respectively. (**g**) EGFP positive cells fold changes compared to mock treatment (DMSO) for different doses of PDC12. The individual frames in **c**, **d**, **f** and **g** represent the mean values ± s.d. for n = 3 independent experiments.