LETTER

doi:10.1038/nature13760

An evolutionary arms race between KRAB zinc-finger genes *ZNF91/93* and SVA/L1 retrotransposons

Frank M. J. Jacobs[1]*†, David Greenberg[1,2]*†, Ngan Nguyen[1,3], Maximilian Haeussler[1], Adam D. Ewing[1]†, Sol Katzman[1], Benedict Paten[1], Sofie R. Salama[1,4] & David Haussler[1,4]

Throughout evolution primate genomes have been modified by waves of retrotransposon insertions[1–3]. For each wave, the host eventually finds a way to repress retrotransposon transcription and prevent further insertions. In mouse embryonic stem cells, transcriptional silencing of retrotransposons requires KAP1 (also known as TRIM28) and its repressive complex, which can be recruited to target sites by KRAB zinc-finger (KZNF) proteins such as murine-specific ZFP809 which binds to integrated murine leukaemia virus DNA elements and recruits KAP1 to repress them[4,5]. KZNF genes are one of the fastest growing gene families in primates and this expansion is hypothesized to enable primates to respond to newly emerged retrotransposons[6,7]. However, the identity of KZNF genes battling retrotransposons currently active in the human genome, such as SINE-VNTR-Alu (SVA)[8] and long interspersed nuclear element 1 (L1)[9], is unknown. Here we show that two primate-specific KZNF genes rapidly evolved to repress these two distinct retrotransposon families shortly after they began to spread in our ancestral genome. ZNF91 underwent a series of structural changes 8–12 million years ago that enabled it to repress SVA elements. ZNF93 evolved earlier to repress the primate L1 lineage until ~12.5 million years ago when the L1PA3-subfamily of retrotransposons escaped ZNF93's restriction through the removal of the ZNF93-binding site. Our data support a model where KZNF gene expansion limits the activity of newly emerged retrotransposon classes, and this is followed by mutations in these retrotransposons to evade repression, a cycle of events that could explain the rapid expansion of lineage-specific KZNF genes.

KAP1 mediates transcriptional silencing of retrotransposons and protects genome integrity through repression of retrotransposition activity[10,11]. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) analysis revealed that in human embryonic stem cells (hESCs), KAP1 predominantly associates with active primate-specific classes of retrotransposons such as SVA and L1PA (Extended Data Fig. 1)[11,12]. Similarly, in mouse ESCs (mESCs) KAP1 primarily associates with mouse-lineage-specific retrotransposon classes (Extended Data Fig. 2)[12]. These data support the hypothesis that species-specific KZNFs recruit KAP1 to species-specific retrotransposon classes that recently invaded the host's genome[7,13]. To test this, we determined the fate of primate-specific retrotransposons in a non-primate background using trans-chromosomic mESCs that contain a copy of human chromosome 11 (E14(hChr11) cells[14], hereafter termed trans-chromosomic 11 (TC11)-mESCs). In the TC11-mESC cellular environment, primate-specific retrotransposons, including SVA and L1PA elements, are derepressed and gain activating histone H3 Lys 4 (H3K4me3) marks (Fig. 1a, b and Extended Data Fig. 1e). As a result of this de-repression, a majority of SVA (51%), human-specific L1 (L1Hs) (93%) and some other L1PA elements, such as L1PA4 (16%), become aberrantly transcribed. These findings suggest primate-specific retrotransposons have a transcriptional potential[15,16] that is repressed by primate-specific factors.

Promising candidates for these factors are the approximately 170 KZNF genes that emerged during primate evolution[7] (Extended Data Fig. 3a). We reasoned that a KZNF gene responsible for protecting genome integrity, most critical in the germ line, must be highly expressed in hESCs. So we focused on 14 highly expressed, primate-specific KZNF genes (Extended Data Fig. 3b) and tested each candidate for a role in repressing SVA retrotransposons, which first appeared in great apes 18–25 million years (Myr) ago[8], and are still active[17]. We set up a luciferase assay based screen in mESCs in which an SVA element cloned upstream of a minimal SV40 promoter strongly enhances luciferase activity (Extended Data Fig. 4a). Each candidate KZNF was co-expressed with the SVA–luciferase construct to determine its effect on reporter activity. Of all KZNFs tested, ZNF91 most dramatically decreased SVA-driven luciferase activity, reducing activity to 16 ± 4% relative to an empty-vector-transfected control (Fig. 2a). Some other KZNFs had modest effects on this reporter, but were not further analysed, as those with the strongest effect also inhibited the OCT4 (also known as POU5F1) enhancer, which is not KAP1-bound in ESCs, and therefore suggests a nonspecific effect (Extended Data Fig. 7a). Structure–function analysis of SVA revealed that the variable number tandem repeat (VNTR) domain is necessary

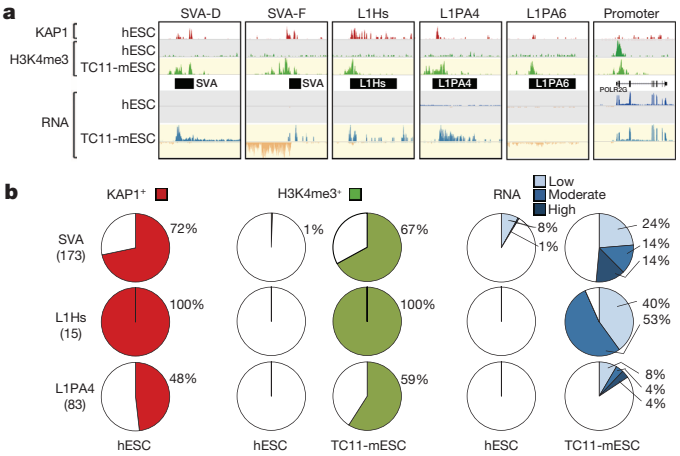


Figure 1 | SVAs and L1PAs are derepressed in a non-primate cellular environment. a, KAP1, H3K4me3 ChIP-seq and RNA sequencing (RNA-seq) coverage tracks for a selection of KAP1-bound primate-specific retrotransposons derepressed in TC11-mESCs (yellow) relative to hESCs (grey). H3K4me3 signal on promoters is similar in hESCs and TC11-mESCs. **b**, Percentages of SVA, L1Hs and L1PA elements on human chromosome 11 positive for KAP1, H3K4me3 and relative levels of transcription (see Methods) in hESC and TC11-mESCs. Total elements of each type on human chromosome 11 in parentheses.

[1]Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA. [2]Molecular, Cell and Developmental Biology, University of California Santa Cruz, Santa Cruz, California 95064, USA. [3]Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA. [4]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California 95064, USA. †Present addresses: Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam 1098 XH, The Netherlands (F.M.J.J.); Gladstone Institute of Virology and Immunology, San Francisco, California 94158, USA (D.G.); Mater Research Institute, University of Queensland, Queensland 4101, Australia (A.D.E.).
*These authors contributed equally to this work.

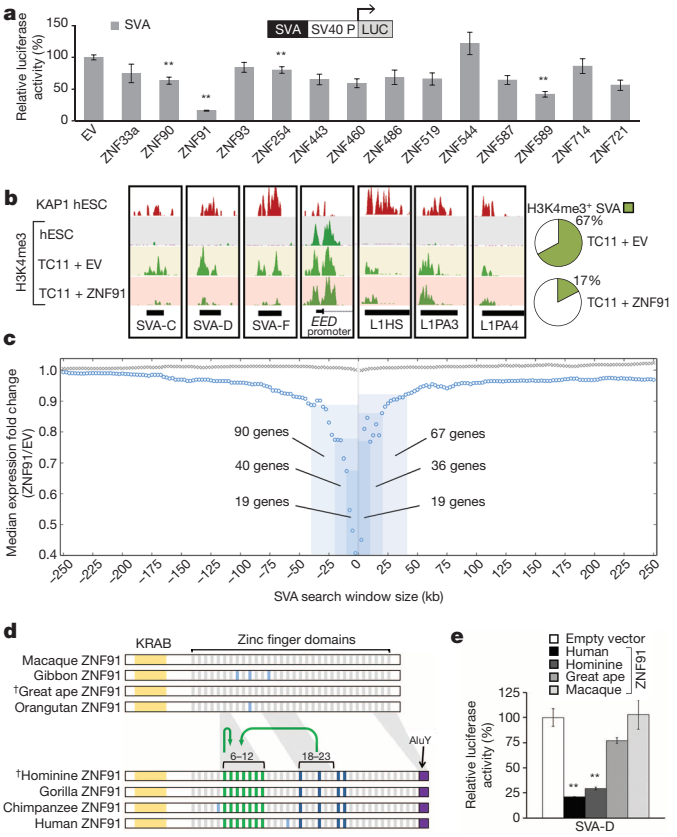


Figure 2 | SVA elements are repressed by primate-specific ZNF91. **a**, Relative luciferase activity of a SVA-D–SV40–luciferase-reporter after co-transfection of KZNFs in mESCs. EV, empty vector. **b**, KAP1 and H3K4me3 ChIP-seq coverage tracks for a selection of loci in hESCs and TC11-mESCs transfected with an empty vector (TC11 + EV) or ZNF91 (TC11 + ZNF91). Pie charts show percentages of H3K4me3-positive SVAs on human chromosome 11. **c**, Median fold expression change (ZNF91 relative to empty vector), for genes with (blue circles) or without (grey crosses) an SVA within the indicated genomic distance among the 994 expressed human chromosome 11 genes; kb, kilobases. **d**, ZNF91 structural evolution. Green stripes, duplicated zinc-fingers; blue stripes, zinc-fingers that changed contact residues in the lineage to humans (dark blue) or in other lineages (light blue). Green arrows indicate segmental duplications. Dagger symbols indicate reconstructed ancestral proteins. **e**, Relative SVA_D–SV40–luciferase activity in the presence of various ZNF91 proteins. **a**, **e**, $^{**}P < 0.01$; error bars are s.e.m.

and sufficient for ZNF91-mediated repression of luciferase activity (Extended Data Fig. 4b, c). Furthermore, transfection of TC11-mESCs with human ZNF91 restored the repression of deregulated SVAs on human chromosome 11, causing a strong decrease of the aberrant H3K4me3 ChIP-seq signal at SVAs, while leaving other derepressed elements such as L1Hs or L1PAs unaffected (Fig. 2b and Extended Data Fig. 5a). Transfection of ZNF91 also significantly repressed aberrant transcription of SVA repeats, indicating that ZNF91 is sufficient to restore transcriptional silencing of SVAs. (Extended Data Fig. 5b). No such effects were observed for other primate KZNFs (*ZNF90*, *ZNF93*, *ZNF486*, *ZNF826*, *ZNF443*, *ZNF544* or *ZNF519*) transfected in TC11-mESCs, validating the specificity of the ZNF91–SVA interaction (Extended Data Fig. 5c). Cellular genes near SVAs on human chromosome 11 in TC11-mESCs were also repressed by ZFN91, with the distance of a gene to an SVA as the major factor governing the amount of bystander repression (Fig. 2c), supporting the hypothesis that the host response to retrotransposon insertion has significantly impacted human gene expression patterns[11,15,16].

ZNF91 emerged in the last common ancestor (LCA) of humans and Old-World monkeys and has undergone dramatic structural changes, including the addition of seven zinc-fingers in the LCA of humans and gorillas[18] (Fig. 2d). We reconstructed ancestral versions of ZNF91 by parsimony analysis (Extended Data Fig. 6a, b) and found that ZNF91 as it probably existed in the LCA of humans and gorillas (ZNF91$^{\text{hominine}}$) was able to repress the SVA-luciferase reporter in a similar fashion to human ZNF91 (Fig. 2e). However, ZNF91 as it existed in the LCA of humans and orangutans (ZNF91$^{\text{great ape}}$) only reduced luciferase activity to around 80% of baseline and macaque ZNF91 completely lacked the ability to repress SVA-driven luciferase activity. The importance of the seven recently added hominine zinc-fingers was further supported by deletion analysis of ZNF91 (Extended Data Fig. 6c). These findings suggest that the changes in ZNF91 between 8–12 Myr ago have markedly improved the protein's ability to bind and repress SVA.

In our KAP1 ChIP experiments, KAP1 also showed a strong association with the 5′ untranslated region (UTR) of L1PA elements. None of the 14 KZNFs had a significant effect on the 5′ UTR of the current active L1Hs[9,19] cloned upstream of the luciferase reporter when tested in mESCs. However, ZNF93 significantly reduced luciferase activity of a reporter with the 5′ UTR of a KAP1-positive L1PA4 element (62 ± 10%, Extended Data Fig. 7a). To verify the recruitment of ZNF93 to L1PA4 elements on the human genome, we performed ChIP-seq analysis on hESCs using antibody ab104878, which recognizes ZNF93 and co-immunoprecipitates KAP1 (Extended Data Fig. 7b, c). We found that ZNF93 binds to the 5′ end of L1PA4, the ancestral subtypes L1PA6 and L1PA5, and the descendant subtype L1PA3 (Fig. 3a and Extended Data Fig. 7d). To validate that the ab104878 ChIP-seq signal on L1PAs is derived from ZNF93, we performed ab104878-ChIP analysis followed by quantitative PCR on TC11-mESC transfected with ZNF93 or an empty vector and found significant enrichment of the L1PA4 5′ UTR compared to a LTR12C control element (Extended Data Fig. 7e). No consistent ZNF93 binding was detected at L1PA7 or older subtypes nor at the most recently evolved L1PA2 and L1Hs (Fig. 3a). Comparative sequence analysis revealed that the absence of ZNF93 binding in L1Hs and L1PA2 can be explained by a 129-base-pair (bp) deletion in the 5′ UTR that spans the ChIP-determined ZNF93- and KAP1-binding sites (Fig. 3b). The deletion is also present in ~50% of L1PA3 elements, resulting in distinct subgroups of shorter (L1PA3-6030) and longer (L1PA3-6160) L1PA3 elements, but is not present in L1PA4–6 families.

To investigate the interaction of ZNF93 with the 129-bp L1PA element, we tested a series of L1PA4 segments cloned upstream of an OCT4-enhancer fused to an SV40-promoter and luciferase-reporter in mESCs (Fig. 3c). Both the 129-bp element and a 51-bp sub-fragment were sufficient to confer ZNF93-mediated repression of the luciferase reporter, and this repression was abolished by elimination of the 51-bp portion in the 129-bp fragment (129Δ51$^{\text{L1PA4}}$). The 51-bp element encompasses a computationally predicted DNA binding motif for the 17 fingers of ZNF93[20] and the central 18 bp of this region displays strong similarity to the predicted recognition motif of zinc-fingers 8–13 of human ZNF93 (Fig. 3d). A ZNF93 variant that has all contact residues in zinc-fingers 8–13 replaced by serine residues (ZNF93serF), a modification that abolishes DNA binding selectivity[21], was unable to repress luciferase activity of the L1PA4 elements (Fig. 3e), suggesting that fingers 8–13 of ZNF93 are important for recognition of the 129-bp element in L1PA3-6 retrotransposons.

ZNF93 emerged in the LCA of apes and Old-World monkeys and reconstruction of the evolutionary history of the ZNF93 protein by parsimony suggests that dramatic changes took place in the LCA of orangutans and humans between 12–18 Myr ago (ZNF93$^{\text{great ape}}$; Extended Data Fig. 8a). Indeed, macaque ZNF93 does not have the ability to repress the 129-bp or 51-bp element of L1PA4 in the luciferase assay, but ZNF93$^{\text{great ape}}$ represses at levels similar to ZNF93$^{\text{human}}$ (Extended Data Fig. 8b), suggesting changes in the ape lineage probably enabled ZNF93 to regulate L1 activity.

To explore the function of the lost 129-bp element, we created a version of L1Hs with this sequence restored in its 5′ UTR (L1Hs+129$^{\text{L1PA4}}$), or a scrambled version of this 129-bp sequence (L1Hs+129scramble$^{\text{L1PA4}}$) as a control, and compared retrotransposition efficiencies to wild-type L1Hs in HEK293FT cells in an *in vitro* retrotransposition assay[22,23]. In

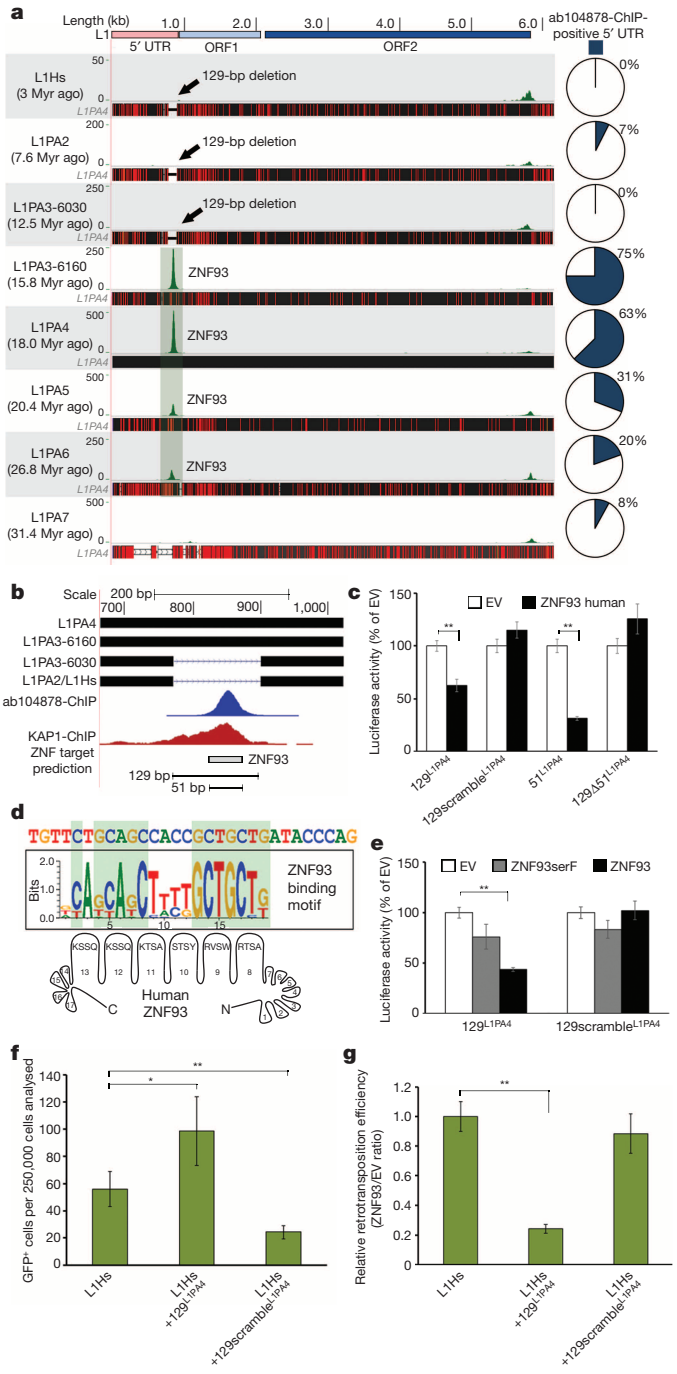


Figure 3 | L1PA elements are repressed by primate-specific ZNF93. **a**, Green peaks represent genome-wide ab104878-ChIP-seq peak-summits mapped to L1PA consensus sequences. Black horizontal bars, alignment to L1PA4; red lines, divergent positions. **b**, The 129-bp deletion and predicted 51-bp ZNF93 binding motif (grey bar) relative to L1PA4. **c**, Relative activity of OCT4-enhancer–luciferase-reporters after co-transfection of an empty vector (EV) or ZNF93. 129^{L1PA4}, 129-bp fragment of L1PA4; $129\Delta51^{LIPA4}$, 129-bp fragment without the 51-bp part; $129scramble^{L1PA4}$, scrambled 129-bp fragment; 51^{L1PA4}, 51-bp fragment. **d**, Consensus central sequence of ab104878-ChIP-seq summits for L1PA4, aligned with the predicted recognition motif of ZNF93 zinc-fingers 8–13. **e**, Relative activity for OCT4-enhancer–luciferase-reporters after co-transfection of EV, ZNF93serF or ZNF93. **f**, Number of GFP-positive cells derived from retrotransposition events of L1Hs, L1Hs+129 and L1Hs+129scrambled constructs in HEK cells ($n = 7$). **g**, Same as **f** but showing the ratio of retrotransposition events after co-transfection with ZNF93 compared to an empty vector. **c**, **e**, **f**, **g**, $*P < 0.05$; $**P < 0.01$; error bars are s.e.m.

this assay, a retrotransposition event results in green fluorescent protein (GFP) expression (Extended Data Fig. 9). L1Hs+129^{L1PA4} shows a 1.76-fold (± 0.45 s.e.m.) higher retrotransposition activity compared to wild-type L1Hs, an effect not seen with L1Hs+$129scramble^{L1PA4}$ (Fig. 3f), suggesting that this 129-bp sequence promotes retrotransposition. Importantly, co-expression of ZNF93 significantly reduced retrotransposition of L1Hs+129^{L1PA4} to just 24% (± 3% s.e.m.) relative to L1Hs, but had no significant effect on L1Hs+$129scramble^{L1PA4}$ (Fig. 3g).

These data suggest the 129-bp sequence, as it once existed in the 5′ UTR of L1PA subfamilies, may have been beneficial to L1 mobilization, but since ZNF93 evolved to bind this element, losing it allowed the L1 lineage to escape ZNF93-mediated repression, providing net selective advantage. Indeed, phylogenetic analysis of L1PA3 elements and calculation of the average distance of L1PA3-6030 and L1PA3-6160 elements from the respective consensus sequences, suggests that L1PA3-6030 elements lacking the 129-bp element have expanded more recently in our genome than L1PA3-6160 elements, showing an estimated age of 12.5 and 15.8 million years, respectively (Extended Data Fig. 10a). This strongly suggests that loss of the ZNF93-binding site—and thereby the evasion of the host repression—propagated a new wave of L1 insertions in great ape genomes.

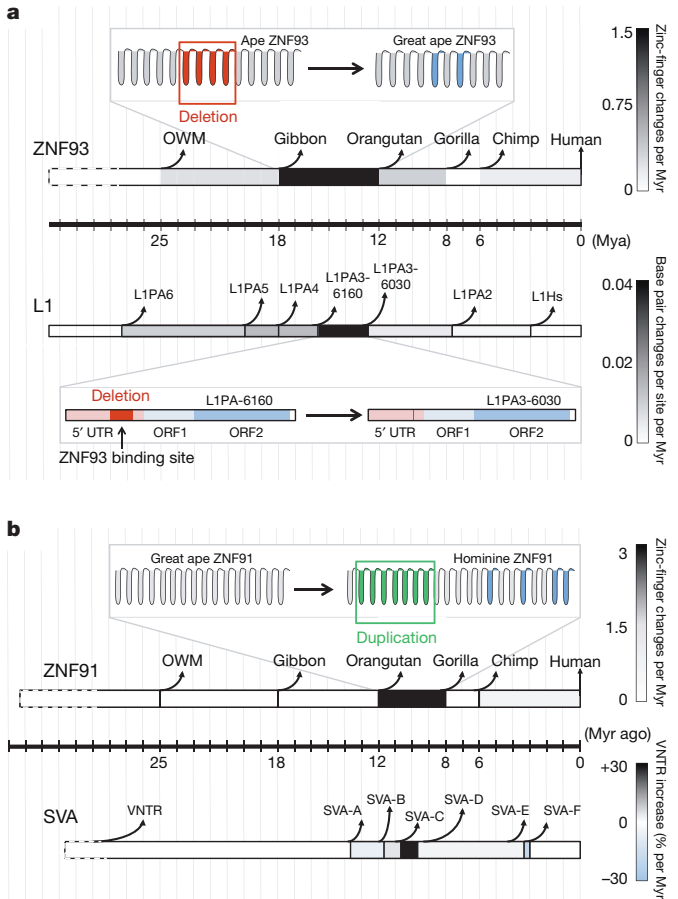


Figure 4 | Dynamic patterns of co-evolution between ZNFs and target retrotransposons. **a**, **b**, Schematic showing the evolution of L1PA[9] and SVA[8] retrotransposons parallel to the structural evolution of ZNF93 and ZNF91 along an evolutionary timescale. Colouring of ZNF91 and ZNF93 horizontal bars represent zinc-finger changes per million years during the time interval indicated. Red zinc-fingers, deletion; blue zinc-fingers, change in contact residues; green zinc-fingers, duplication. Colouring of retrotransposon horizontal bars represents base-pair substitutions, deletions or insertions per site per million years (L1PA), or percentage increase in VNTR size per million years (SVA). Myr, million years; OWM, Old-World monkey.

Repeated turnover of the 5′ UTR occurred in early L1PA evolution[9] and was previously thought to be associated with competition for host factors[24]. Our results suggest turnover was instead driven by avoidance of host factors. The precise removal of the ZNF93-binding site probably took place soon after ZNF93 underwent a series of structural changes, suggesting the deletion may have been driven by improved host repression of L1PA activity (Fig. 4a). In a similar fashion, the structural changes in ZNF91 allowing it to repress SVA elements may have driven the further evolution of new and different SVA-subtypes in gorillas, chimpanzees and humans, a pattern that is not observed in orangutans, which diverged before ZNF91 had undergone these structural changes (Extended Data Fig. 10b). Notably, the size of the VNTR region of SVA, the prime interaction site of ZNF91, has increased during the timeframe of structural changes to ZNF91 (Fig. 4b and Extended Data Fig. 10c).

Our data support a model in which modifications to lineage-specific KZNF genes are used by the host to repress new families of retrotransposons as they emerge, which in turn drives the evolution of newer families of retrotransposons, in a continuing arms race. Because repression affects nearby genes, KZNFs have probably been co-opted for other functions that persisted long after the original transposon expansion they first evolved to repress had subsided[25], fuelling the evolution of more complex gene-regulatory networks. Unlike an arms race with an external pathogen, retrotransposons are host DNA, suggesting that a mammalian genome is itself in an internal arms race with its own DNA, and thereby inexorably driven towards greater complexity.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 December 2013; accepted 7 August 2014.
Published online 28 September 2014.

1. Kazazian, H. H. Mobile elements: drivers of genome evolution. *Science* **303,** 1626–1632 (2004).
2. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nature Rev. Genet.* **10,** 691–703 (2009).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
4. Wolf, D. & Goff, S. P. TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. *Cell* **131,** 46–57 (2007).
5. Wolf, D. & Goff, S. P. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature* **458,** 1201–1204 (2009).
6. Birtle, Z. & Ponting, C. P. Meisetz and the birth of the KRAB motif. *Bioinformatics* **22,** 2841–2845 (2006).
7. Thomas, J. H. & Schneider, S. Coevolution of retroelements and tandem zinc finger genes. *Genome Res.* **21,** 1800–1812 (2011).
8. Wang, H. *et al.* SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.* **354,** 994–1007 (2005).
9. 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).
10. Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463,** 237–240 (2010).
11. Turelli, P. *et al.* Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. *Genome Res.* **24,** 1260–1270 (2014).
12. Castro-Diaz, N. *et al.* Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev.* **28,** 1397–1409 (2014).
13. Huntley, S. *et al.* A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res.* **16,** 669–677 (2006).
14. Kai, Y. *et al.* Enhanced apoptosis during early neuronal differentiation in mouse ES cells with autosomal imbalance. *Cell Res.* **19,** 247–258 (2009).
15. Gifford, W. D., Pfaff, S. L. & Macfarlan, T. S. Transposable elements as genetic regulatory substrates in early development. *Trends Cell Biol.* **23,** 218–226 (2013).
16. Ward, M. C. *et al.* Latent regulatory potential of human-specific repetitive elements. *Mol. Cell* **49,** 262–272 (2013).
17. Hancks, D. C. & Kazazian, H. H. Active human retrotransposons: variation and disease. *Curr. Opin. Genet. Dev.* **22,** 191–203 (2012).
18. Bellefroid, E. J. *et al.* Emergence of the ZNF91 Krüppel-associated box-containing zinc finger gene family in the last common ancestor of anthropoidea. *Proc. Natl Acad. Sci. USA* **92,** 10757–10761 (1995).
19. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nature Rev. Genet.* **12,** 615–627 (2011).
20. Persikov, A. V., Osada, R. & Singh, M. Predicting DNA recognition by Cys_2His_2 zinc finger proteins. *Bioinformatics* **25,** 22–29 (2009).
21. Moore, M., Choo, Y. & Klug, A. Design of polyzinc finger peptides with structured linkers. *Proc. Natl Acad. Sci. USA* **98,** 1432–1436 (2001).
22. Ostertag, E. M., Prak, E. T., DeBerardinis, R. J., Moran, J. V. & Kazazian, H. H. Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic Acids Res.* **28,** 1418–1423 (2000).
23. Kimberland, M. L. *et al.* Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Hum. Mol. Genet.* **8,** 1557–1560 (1999).
24. Swergold, G. D. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.* **10,** 6718–6729 (1990).
25. Lowe, C. B., Bejerano, G. & Haussler, D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl Acad. Sci. USA* **104,** 8005–8010 (2007).

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

Acknowledgements This work was supported by California Institute of Regenerative Medicine (CIRM) facility awards (FA1-00617, CL1-00506-1.2) and scholar awards (TG2-01157) to F.M.J.J. and D.G. and F.M.J.J. also received a Human Frontier Science Program Postdoctoral fellowship (LT000689). D.H. is an Investigator of the Howard Hughes Medical Institute. S.K. is supported by the California Institute for Quantitative Biosciences, A.D.E. was supported by TCGA U24 24010-443720, M.H. by EMBO ALTF 292-2011, and B.P. and N.N. by ENCODE U41HG004568. We thank F. Wianny and C. Dehay (Lyon University) for the LYON-ES1 macaque embryonic stem cells; M. Oshimura and T. Inoue (Tottori University) for the E14(hChr11) trans-chromosomic embryonic stem cells, N. Pourmand and the UCSC genome sequencing center; B. Nazario (UCSC Institute for the Biology of Stem Cells) for flow cytometry assistance; M. Batzer (LSU) and K. Han (Dankook University) for L1CER sequences; L. Carbone (OHSU) for gibbon genomic DNA; A. Smit (ISB, Seattle) for discussions on L1PA evolution; D. Segal (UC Davis) for advice on ZNF mutations; H. Kazazian, D. Hancks and J. Goodier (JHMI) for retrotransposition plasmids and advice; K. Tygi, C. Vizenor, J. Rosenkrantz, W. Novey, S. Kyane and B. Mylenek for technical assistance and the entire Haussler laboratory for discussions and support.

Author Contributions F.M.J.J., D.G., D.H. and S.R.S. designed and analysed the experiments. F.M.J.J. performed RNA-seq, ChIP-seq and reintroduction of primate ZNFs in trans-chromosomic mESCs; D.G. performed ZNF cloning, luciferase reporter and retrotransposition assays; N.N., D.G., A.D.E. and B.P. performed resequencing and analysis to complete the *ZNF91* and *ZNF93* loci in various primates; N.N. and B.P. reconstructed the evolutionary history of ZNF91 and ZNF93 ZNF domains; M.H. generated a Repeatmasker UCSC-Browser and hub, ZNF-binding site predictions and VNTR length analysis; S.K. processed and analysed RNA-seq and ChIP-seq data; A.D.E. analysed SVA numbers in great apes and SVA–gene-expression correlations. F.M.J.J., D.G., S.R.S. and D.H. wrote the manuscript.

Author Information The data discussed in this publication have been deposited in the NCBI Gene Expression Omnibus and are accessible through GEO Series accession number GSE60211. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.H. (haussler@soe.ucsc.edu).

METHODS

Embryonic stem cell culture and ZNF overexpression analysis. Human (H9) ESC colonies were maintained as described (http://www.wicell.org). Colonies were manually passaged at a 1:3 ratio onto plates containing mitomycin-C-treated mouse embryonic fibroblasts that were seeded at a density of 35,000 cells cm^{-2} on 0.25% gelatin-coated plates (porcine; Sigma) the day before. Mouse transchromosomic E14(hChr11) (TC11) ESCs were cultured on mouse embryonic fibroblast feeder layers as described[14]. For transfections, cells were cultured on gelatin for two passages and transfected with 24 μg of ZNF and 1 μg of GFP expression vectors per 10 cm plate of cells, using lipofectamine 2000 (Invitrogen). Cells were cultured for an additional 40 h, harvested with trypleE reagent (Life technologies) and washed three times and collected in fluorescence-activated cell sorting (FACS) buffer (1× PBS, 2% fetal bovine serum (FBS), 5 mM EDTA). GFP-positive cells were sorted using a FACSAria III (BD Biosciences) and samples were used for RNA isolation and ChIP analysis.

RNA-seq library preparation. RNA was treated with RQ1 DNaseI (Promega) for 1 h at 37 °C and total RNA was cleaned up using the RNAeasy Mini kit (Qiagen). For each sample, the non-ribosomal fraction of 5 μg of total RNA was isolated using a Ribo-Zero rRNA removal Kit (Epicentre) following the manufacturer's protocol (Lit. 309-6/2011). For the non-ribosomal fraction of RNA, double stranded (ds) complementary DNA was synthesized as described previously[26] using dUTP in the second strand synthesis and USER digest before amplification to retain strand specificity. Clean-up steps were performed using RNA Clean & Concentrator or DNA Clean & Concentrator kits (Zymo Research). Double stranded cDNA was used for library preparation following the Low Throughput Guidelines of the TruSeq DNA Sample Preparation kit (Illumina), with the following additions. Size selections were performed before and after cDNA amplification on an E-gel Safe Imager (Invitrogen) using 2% E-gel SizeSelect gels (Invitrogen). The cDNA fraction of 300–400 bp in size (including adapters) was isolated and purified. For adaptor ligations, 1 μl instead of 2.5 μl of DNA Adaptor Index was used. Indexed libraries were pooled and sequenced on the Illumina HiSEQ platform. Two biological replicate samples were analysed for empty-vector-transfected cells and ZNF91-transfected cells, three biological replicate samples were analysed for human ESCs and two for rhesus macaque LYON-ES1 ESCs. Data can be viewed on the UCSC browser: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hubUrl=http://hgwdev.soe.ucsc.edu/~max/jacobs2014/hub.txt&position=chr11:60180780-60680779.

Mapping and analysis of RNA-seq data. All samples were mapped using Tophat2 (ref. 27) with Bowtie2 (ref. 28) as the underlying alignment tool. The input Illumina fastq files consisted of paired-end reads with each end containing 100 bp. The target genome assembly for the human samples was GRCh37/UCSC-hg19 for hESCs, or a hybrid target genome of mm9-hChr11 for TC11-mESCs, and Tophat was additionally supplied with a gene model (using its '-GTF' parameter) with data from the hg19 UCSC KnownGenes track[29]. For multiply-mapped fragments, only the highest scoring mapping determined by Bowtie2 was kept. Only mappings with both read ends aligned were kept. Potential PCR duplicates (mappings of more than one fragment with identical positions for both read ends) were removed with the samtools 'rmdup'[30] function, keeping only one of any potential duplicates. The final set of mapped paired-end reads for a sample were converted to position-by-position coverage of the relevant genome assembly using the bedtools 'genomeCoverageBed'[31] function. To determine the count of fragments mapping to a gene, the position-by-position coverage was summed over the exonic positions of the gene. This gene total coverage was divided by a factor of 200, to account for the 200 bp of coverage induced by each mapped paired-end fragment (100 bp from each end), and rounded to an integer. For the human samples, this was calculated for each gene in the UCSC Known Gene set. For input to DESeq[32] all genes with non-zero counts in any sample were considered. Two replicates of each sample were combined per the DESeq methodology.

For Fig. 2c, the median fold change in expression (ZNF91/EV, vertical axis) for genes with an SVA element within some distance (blue circles) and genes without an SVA element within the same distance (grey crosses) were plotted against the up- or downstream distance from each gene. A total of 994 expressed genes were considered. Points were computed every 2.5 kb, For every window size starting at 2.5 kb and progressing cumulatively up to 250 kb in 2.5 kb intervals upstream and downstream of genes on chromosome 11, we identified the set of genes with and without at least one SVA element within the window. For the two sets (genes with SVA and genes without SVA), at every window size we calculated the median fold change in gene expression (ZNF91/EV) using the DESeq results from TC11-mESCs transfected with either ZNF91 or an empty vector. The python script to generate the figure and the associated data are available at http://hgwdev.sdsc.edu/~ewingad/Tc11SVAFig2e.tar.gz.

Chromatin immunoprecipitation (ChIP), ChIP-qPCR and ChIP-seq library preparation. Human (H9) and mouse ESCs (46C and transchromosomic TC11) were crosslinked in 1% formaldehyde for 10 min on ice by adding 1/10 volume of freshly prepared 11× crosslinking solution (50 mM Hepes (pH 8.0); 0.1 M NaCl; 1 mM EDTA; 0.5 mM EGTA; 11% formaldehyde). The crosslinking reaction was quenched by adding glycine to a final concentration of 0.125 M and incubating for 5 min on ice. For KAP1-ChIP and ChIP with the KZNF antibody ab104878, cells were washed three times in PBS + 0.1% BSA and dissolved in ten packed cell volumes 0.3% SDS-lysis buffer (10 mM Tris (pH 8.0); 1 mM EDTA (pH 8.0); 0.3% (w/v) SDS + Complete Proteinase Inhibitor Cocktail (Roche)). Cells were incubated on ice for 20 min and cells were lysed in a pre-chilled Dounce homogenizer by ten strokes with pestle B. Cell lysate was transferred to a 15 ml conical (hESC) or 1.5 ml tube (mESC) and chromatin was sheared to an average size of ~500 bp in a Bioruptor Sonicator (Diagenode) (settings: HIGH; 30 s on; 60 s off; 10–12 cycles). Sonicated lysate was transferred to 2 ml tubes and three lysate volumes of immunoprecipitation buffer (50 mM Tris-HCl (pH 8.0); 150 mM NaCl; 5 mM MgCl; 0.5 mM EDTA; 0.2% NP-40; 5% glycerol; 0.5 mM dithiothreitol); Complete Protease Inhibitor Cocktail was added. Debris was pelleted by centrifugation for 15 min at 12,000*g* at 4 °C and supernatant was transferred to a new 2 ml vial. Supernatant was pre-cleared with 50 μl of Sheep-anti-Rabbit (M-280) Dynabeads (Invitrogen) for 4 h at 4 °C. Dynabeads (Invitrogen) were blocked with BSA according to the Dynabeads manual. Pre-cleared lysate was incubated with 10 μl of dynabeads suspension pre-bound for 4 h with an excess of anti-KAP1 antibody (ab10484), or anti-KRAB ZNF-antibody (ab104878). Immunoprecipitation was performed overnight at 4 °C on a rotator. Immunocomplexes were washed six times in freshly prepared RIPA buffer (50 mM Hepes (pH 8.0); 1 mM EDTA (pH 8.0); 1% (v/v) NP-40; 0.7% (w/v) deoxycholate; 0.5 M LiCl; Complete Proteinase Inhibitor Cocktail) and once in TE buffer (10 mM Tris-HCl (pH 8.0); 1 mM EDTA (pH 8.0)). H3K4me3-ChIP (H3K4me3 antibody: Milipore; catalogue no. 07-473; lot no. JBC1888194) was performed following the Roadmap Epigenome Project Protocol (April 19, 2010 version) available at http://www.roadmapepigenomics.org/protocols/type/experimental/. Immunocomplexes were eluted from the beads by incubation at 67 °C for 20 min in ChIP elution buffer (TE + 1% SDS) and vortexing every 2 min; cross-linking was reversed by incubation at 67 °C overnight. ChIP DNA was treated with RNase A/T for 2 h at 37 °C and Proteinase K for 2 h at 55 °C. NaCl was added to a final concentration of 200 mM and ChIP DNA was extracted twice with phenol/chloroform/iso-amyl-alcohol (25:24:1) and twice with chloroform/iso-amyl-alcohol (24:1). ChIP DNA was ethanol precipitated and dissolved in nuclease-free water. ChIP DNA was cleaned up one extra time using Zymo PCR purification columns.

To determine the genome-wide binding of ZNF93, we performed chromatin immunoprecipitation (ChIP) analysis, using a KRAB ZNF antibody (ab104878) which was originally raised against a peptide in ZNF486 that displays 88% identity to ZNF93 and we show is capable of recognizing ZNF93 (Extended Data Fig. 7b, c). Notably, the size of the protein immunoprecipitated by ChIP from hESC lysates corresponds to the size of ZNF93 and not ZNF486, suggesting that this antibody predominantly immunoprecipitates the highly expressed ZNF93. To establish that ZNF93 can direct ab104878 to the L1PA4 5′ UTR, ChIP-quantitative-PCR was performed on ab104878-ChIP-DNA derived from three biological replicates of TC11-mESCs transfected with either *pCAG-EV*, where EV represents an empty vector, or *pCAG-ZNF93*. Quantitative PCR was performed on a Roche LightCycler 480 II, using primers to amplify an amplicon in the 5′ UTR of L1PA4 (forward: CATTTGCGGTTCACCAATATC; reverse: GCTAGAGGTCCACTCCAGAC) and LTR12C (forward: GCACTTGAGGAGCCCTTCAG; reverse: ACACCTCCCTG CAAGCTGAG).

For ChIP-seq analysis, ChIP-DNA was used for library preparation following the Low Throughput Guidelines of the TruSeq DNA Sample Preparation kit (Illumina), with the following minor additions. Size selections were performed before and after amplification on an E-gel Safe Imager (Invitrogen) using 2% E-gel SizeSelect gels (Invitrogen). The ChIP-DNA fraction of 300–400 bp in size (including adapters) was isolated and purified. For adaptor ligations, 1 μl instead of 2.5 μl of DNA Adaptor Index was used. Indexed libraries were pooled and sequenced on the Illumina HiSEQ platform. For ChIP-seq analysis in hESCs, three biological replicates of KAP-ChIP, two biological replicates of H3K4me3-ChIP and two biological replicates of ab104878-ChIP were analysed, and for H3K4me3 ChIP-seq analysis in TC11-mESCs, two biological replicate samples were analysed for empty-vector-transfected cells and ZNF91-transfected cells, and one sample was analysed for other KZNF genes reported in Extended Data Fig. 5c. Data can be viewed on the UCSC browser: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hubUrl=http://hgwdev.soe.uc.

MACS ChiP-seq peak calling. All samples were mapped using Bowtie[28] using input Illumina fastq files consisting of paired-end reads. The human samples were mapped to the GRCh37/UCSC hg19 genome assembly. Only fully paired-end, uniquely mapping reads were kept. Potential PCR duplicates (mappings of more than one fragment with identical positions for both read ends) were removed with the samtools 'rmdup'[30] function, keeping only one of any potential duplicates. Based on the paired-end mappings, the median length of the fragments was determined for each sample. For input to MACS 1.4 (ref. 33) only the read1 mappings were used and the median

fragment length was used to determine the '-shiftsize' parameter. For each ChIP sample mappings, the corresponding input DNA sample mappings were used as a control. The UCSC table browser[34] was used to select MACS peaks that were called in both biological replicates. The overlap between KAP1 ChIP-seq replicates is ~30%, which is lower than expected and can probably be best explained by numerous retrotransposon and promoter regions on the genome displaying a low level of (possibly transient) KAP1 binding that may be below threshold in one replicate, and above threshold in the other.

Quantification of ChIP-seq and RNA-seq data for Figs 1b and 2b. For specific retrotransposon classes, the percentage of elements on human chromosome 11 (a total of 173 SVA elements; 15 full-length L1Hs elements; 84 full-length L1PA4 elements) that overlapped with KAP1 ChIP-seq peaks and H3K4me3 ChIP-seq peaks in hESCs and TC11-mESCs was determined using the UCSC table browser. Only L1PAs >5700 bp were considered to select (near) full-length L1 elements for the analysis. Transcription derived from individual SVA, full-length L1Hs and full-length L1PA4 human chromosome 11 elements in hESCs and TC11-mESCs was scored manually based on the RNA-seq coverage track uploaded in the UCSC browser, using a fixed scale that was normalized for relative sequencing depth. Level of transcription was divided in four categories: no (~0–10 reads); low (~10–30 reads); moderate (~30–50 reads) and high transcription (>50 reads). Isolated reads were not counted as transcription, nor were elements scored as transcribed when the transcription covering the retrotransposon was clearly part of exonic or intronic expression of genes. For Fig. 2b, only H3K4me3 ChIP-seq peaks that had a minimal 'score' of 100 for both empty-vector-transfected and ZNF91-transfected TC11-mESCs were considered. The 'score' is a value defined by MACS analysis representing the 'height' of each ChIP-seq signal, and the score of 100 is an arbitrary cut-off that we chose. This provides a quantitative measure of the percentage of SVAs on chromosome 11 that display a reduction of the H3K4me3 signal. For the pie charts in Fig. 3a, we used the UCSC table browser to determine the percentage of full-length L1PA elements on chromosome 11 that overlapped with an ab104878-ChIP-seq peak in the 5′ UTR (5′-most 1000 bp of each individual L1PA element). This analysis was based on 15 L1Hs, 54 L1PA2, 29 L1PA3-6030, 36 L1PA3-6160, 83 L1PA4, 39 L1PA5, 41 L1PA6, 50 L1PA7 and 14 L1PA8 full-length elements. The following should be noted about the discrepancy between the pie charts showing a small fraction of L1PA2 (7%) and L1PA7 (8%) that overlap with ab104878-ChIP-seq peaks in the 5′ UTR, and the repeat browser tracks on the left where no ab104878 ChIP-summit is observed for these elements. The annotation of L1PAs on the RepeatMasker track is based on ~500 bp in the 3′ UTR only, whereas the L1PA reference sequences in the repeat browser we used to generate the ChIP-seq summit tracks in Fig. 3a are based on the consensus of full-length L1PA sequences. In the RepeatMasker track that was used to make the pie-charts, we noticed incidental mis-annotations for these highly similar L1PA subfamilies. In particular, some L1PAs appear to be one subtype on the 3′ end (based on which they were categorized) yet are annotated as a different subfamily on the 5′ end. In fact, manual analysis of the 7% of repeat-masker-annotated L1PA2 fragments positive for KZNF-ChIP, revealed that all are mis-annotations and based on the consensus of the full length L1PA sequence should have been categorized as L1PA4 or L1PA3.

Immunoblotting. Human ESC (H9) and ZNF-transfected TC11-mESCs and HEK cells were lysed in 50 mM Tris-HCl (pH 8.0); 150 mM NaCl; 5 mM MgCl; 0.5 mM EDTA; 0.2% NP-40; 5% glycerol; 0.5 mM dithiothreitol and complete protease inhibitor cocktail (Roche) and centrifuged at max speed for 10 min at 4 °C to remove debris. Cleared lysates were subjected to SDS–PAGE on Nupage (Invitrogen) 4–12% protein gels for SDS–PAGE and transferred to nitrocellulose as described in the Nupage manual. Blots were incubated overnight in 5% non-fat dried milk in PBS-T and incubated with 1:1000 anti-KAP1 antibody (ab10484), 1:1000 anti-KZNF antibody (ab104878) or 1:1000 anti-haemagglutinin (HA; ab9110) antibody in PBS for 3 h and goat-anti-rabbit-HRP secondary antibody for 30 min at room temperature. Blots were incubated with SuperSignal West Dura Extended Duration Substrate (Thermo Scientific) and visualized on a Biorad Chemidoc MP system.

Plasmids. KZNF cDNAs were amplified from hESC cDNA, isolated from IMAGE clones or synthesized (Genscript) and cloned into pCAG EN (Addgene 11160) for transient transfections. For generation of the luciferase constructs, SVA_D (Hg19: chr11: 65, 529, 663-65, 531, 199) was synthesized (Genscript); the OCT4-enhancer region (OCT4Enh; Hg19: chr6: 31, 139, 549-31, 141, 393) was amplified by PCR from hESC gDNA, and L1PA4-5′ UTR (chr11: 74, 005, 653-74, 006, 113) was synthesized (IDT, gBlock) and were cloned upstream of a pGL4CP–SV40[34] luciferase-reporter construct. Retrotransposition assay constructs were modified from pCPE4–$L1_{RP}$–GFP[22]. Detailed plasmid descriptions and sequences of inserts can be found in Supplementary Information File 1.

Luciferase assay. Luciferase assay was carried out according to Promega dual-luciferase kit instructions and as previously published[34]. 46C[35] mESCs were plated in the afternoon on gelatin-coated 24-well plates at 35,000 cells per cm^2. The next morning, media was changed and 200 ng of pCAG–ZNF was co-transfected with 20 ng of SV40–luciferase reporter and 2 ng of pRL–TK–renilla (a 10:1 firefly to renilla ratio) per 24 wells using Lipofectamine2000 in duplicate wells. Twenty-four hours after transfection, wells were washed once with PBS, harvested with 100 µl of Passive Lysis Buffer for 15 min on a room-temperature rocker. Each well is then read in duplicate as 40 µl of lysate was transferred twice to a 96-well white opti-plate and combined with 50 µl of LARII substrate and read on a Perkin-Elmer luminometer and Wallace Victor Light software counting 1 s per well. Next, lysate and substrate was combined with 50 µl of Stop & Glo reagent to quench and measure renilla activity to control for transfection efficiency. Data were normalized in Microsoft Excel by dividing firefly by renilla and the average of four technical replicate measurements was taken as a raw value of activity. This activity was further normalized against an SV40–luciferase control for each KZNF pCAG construct. Final values are displayed, where for each biological replicate pCAG empty vector is set to 100%. Statistical testing was performed with a two-tailed Student's *t*-test and statistical differences of $P < 0.01$ are indicated in the figures. The following number of biological replicates were used: Fig. 2a: empty vector, $n = 42$; ZNF90, $n = 6$; ZNF91, $n = 17$; ZNF93, $n = 9$; ZNF254, $n = 10$; ZNF443/ZNF460/ZNF486/ZNF519/ZNF 544/ZNF 587/ZNF589/ZNF714/ZNF721/ZNF33a, $n = 3$. Fig. 2e: empty vector, $n = 6$; human ZNF91, $n = 3$; hominine ZNF91, $n = 3$; great ape ZNF91, $n = 3$; macaque ZNF91, $n = 3$. Fig. 3c: empty vector , $n = 6$; ZNF93, $n = 3$. Fig. 3e: empty vector, $n = 6$; ZNF93, $n = 4$; ZNF93serF, $n = 6$. Extended Data Fig. 4a, $n = 6$. Extended Data Fig. 4b: no VNTR, $n = 9$; partial VNTR, $n = 3$; no hex/Alu, $n = 2$; no hex, $n = 2$; full-length SVA, $n = 15$; SINE-R, $n = 3$. Extended Data Fig. 4c, $n = 3$. Extended Data Fig. 6c: empty vector , $n = 42$; ZNF91 (1–11), $n = 4$; ZNF91 (1–24), $n = 7$; ZNF91 (1–30), $n = 4$); ZNF91 (1, 2, 23–36), $n = 3$. Extended Data Fig. 7a, $n = 3$. Extended Data Fig. 8b, $n = 4$.

Retrotransposition assay. The full length L1Hs retrotransposition reporter construct[22], was modified to have the 129-bp element of L1PA4 (L1Hs+129^{L1PA4}) or a scrambled 129-bp sequence (L1Hs+129scramble ^{L1PA4}) inserted at the corresponding position where the 129-bp element is present in L1PA4 and lost in L1PA3-6030. See Supplementary Information File 1 for more details on the cloning of these constructs. Retrotranspositon assay of L1Hs and related 129^{L1PA4}-containing constructs was carried out based on established protocols[22,36]. HEK293FT cells were plated at 35,000 cells per cm^2 on 6-well plates and incubated overnight in DMEM+FBS (without penicillin or streptomycin). The next day, cells were transfected with 300 ng of L1Hs reporter and 1 µg of pCAG–empty-vector or pCAG–ZNF93 using lipofectamine 2000/Optimem (Invitrogen); media was changed after 6 h per manufacturer recommendations. Cells were maintained and on day 4 cells were harvested with TrypLE, washed twice with PBS, placed on ice and incubated with propidium iodide. For each transfection 250,000 cells were analysed for GFP-positive and dead cells on a BD LSR II. Data were gated and analysed in FlowJo software to determine the number of live, GFP-positive cells. Statistical testing was carried out using a two-tailed Student's *t*-test; $n = 7$ biological replicates.

Repeat Browser. We constructed a consensus sequence of SVA_D and L1PA elements. To remove extremely short and long copies, we first eliminated the longest 2% of the copies in the genome, then took the 50 longest sequences annotated by RepeatMasker (http://www.repeatmasker.org) in the UCSC genome[37], aligned them with MUSCLE and constructed a consensus sequence from the multiple alignment. We created a version of the UCSC genome browser using this consensus as a reference sequence. MACS summits of KZNF(ab104878)-ChIP-seq and KAP1-ChIP-seq were mapped to the repeat browser for Fig. 3a, b (repeat browser: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hub_27057_repeats2&position=L1PA3long%3A1-6157&hgsid=389007373_caeGCkR66TMstaDYHuKAyt6txDQD).

Multi-species *ZNF91* and *ZNF93* nucleotide sequence identification. We focused on finding homologues in other species for the fourth exon of human *ZNF91* and *ZNF93*, which contains all the important functional domains of the genes, including the KRAB domains and all the zinc-finger domains. Using BLAT from the UCSC genome browser toolset to align the human *ZNF91* (ENST00000300619) genomic nucleotide sequence (UCSC Hg19 chr19: 23, 539, 498-23, 579, 269, from 1 kb upstream to 1 kb downstream), we identified the best reciprocal hit *ZNF91* sequences in the chimpanzee (panTro4), gorilla (gorGor3), orangutan (ponAbe2), gibbon (nomLeu3), rhesus macaque (rheMac2) and baboon (papAnu2) genomes. Of note, for rhesus macaques, we used the rheMac2 assembly because we have identified a potential assembly error in the *ZNF91* fourth-exon region of the latest assembly, rheMac3, which resulted in an early stop codon. The *ZNF91* sequence obtained from rheMac2 was validated by RNA-seq data.

For *ZNF93*, the human fourth exon is located at: UCSC Hg19, chr19: 20, 043, 993-20, 045, 627. We extracted the homologous regions in other species using the UCSC 100 vertebrate species multiple sequence alignment (UCSC browser (http://genome.ucsc.edu/), Multiz Alignments of 100 Vertebrates track). To refine the alignments, we independently aligned the human *ZNF93* fourth-exon nucleotide sequence to these homologous regions together with their immediate upstream and downstream regions (using BLAT) and manually analysed and ensured the

quality of the alignments. We obtained homologues for chimpanzee (panTro4 chr19: 20, 255, 111-20, 256, 670), gorilla (partial homologue due to missing information, gorGor3 chr19: 20, 328, 848-20, 330, 482), orangutan (partial homologue due to missing information, ponAbe2 chr19_random: 3, 818, 660-3, 820, 506), green monkey (chlSab1 chr6: 18, 428, 342-18, 430, 231), rhesus macaque (rheMac3 chr3: 73, 136, 331-73, 137, 882), crab-eating macaque (macFas5 chr19: 20, 589, 892-20, 591, 781) and baboon (papHam1 scaffold15384: 40, 473-42, 362). We aligned these sequences back to the human genome and validated that *ZNF93* was their best match. We used RAxML to construct a phylogenetic tree for these sequences and sequences of human *ZNF93* and its close relatives *ZNF90*, *ZNF737* and *ZNF626*. The results confirmed that these sequences were closest to human *ZNF93*. To check for reciprocal best matches, we aligned the human *ZNF93* fourth-exon sequence to the species genome assemblies. Due to high repetitiveness of the zinc-finger domains and high diversity of the sequences across species, the alignments resulted in a large number of matches, many of which spanned large regions (that is, false positive matches with large 'introns'). We manually analysed these alignments and confirmed that the regions listed above were the best matches.

The *ZNF93* match in gibbon (nomLeu3 chr10: 54, 583, 066-54, 586, 723) contains long insertions, indicating that there are potential errors in the gibbon reference assembly (and/or that the exon is broken into multiple exons in gibbons, and/or that the gibbon exon contains extra bases). In the next section, we explain how we used PCR to correct assembly errors in the gibbon reference to obtain a valid gibbon homologue.

Genome assembly correction at primate *ZNF91* and *ZNF93* loci. Alignments of both translated amino acid and nucleotide sequences revealed that the identified orangutan and gorilla sequences had scaffold gaps within the fourth exon of the gene *ZNF91*, which includes crucial zinc-fingers. To fill in the gaps and correct assemblies we used genomic DNA from orangutan and gorilla fibroblasts (Coriell, San Diego Zoo), and performed PCR using a selection of primers that are provided in Supplementary Information File 2. Cloned PCR products were Sanger sequenced and sequences were aligned to the corresponding assemblies as well as to the human genome using BLAT. Only clones that mapped uniquely with at least 90% coverage to the corresponding regions were kept. Similarly, orangutan and gorilla sequences had scaffold gaps within the fourth exon of the gene *ZNF93*. We used genomic DNA from Sumatran orangutan and gorilla fibroblasts (San Diego Zoo) to fill in these gaps.

We identified potential assembly errors in the gibbon reference assembly (nomLeu3). To obtain a confident homologue of the fourth exon of *ZNF93* in gibbons, we used gDNA of gibbon species *Hylobates pileatus*, *Hyloblates gabriellae* and *Nomascus leucogenys*, which were a gift from L. Carbone (Oregon Health Sciences University Primate Center) and purchased from Coriell Cell Repositories. Purified PCR products were ligated into PCR4-TOPO (Invitrogen) and sequenced. The resulting sequences were aligned to the gibbon reference assembly (nomLeu3) and were manually analysed and assembled into the consensus gibbon *ZNF93* fourth-exon sequence. The reference gibbon assembly nomLeu3 contains one tandem duplication (of the corresponding human domains 6–12) and one long insertion (~1 kb), both were refuted by sequence evidence obtained from this experiment.

Reconstructing the evolutionary history of ZNF91. Multiple sequence alignments revealed a 588-bp subsequence containing seven extra zinc-fingers in the human, chimpanzee and gorilla genomes that are not present in the orangutan, gibbon, rhesus macaque and baboon genomes. This additional sequence corresponds to zinc-fingers 6–12 of the human protein. Using BLAT to align the human copy of this sequence to the human genome, human zinc-fingers 7–12 (2–7 of the subsequence) have the best reciprocal homology to zinc-fingers 18–23 of human *ZNF91*, indicating that the subsequence was initially created by a local segmental duplication. Further analysis revealed human zinc-finger 6 (the first zinc-finger of the additional subsequence) is a near exact, best-reciprocal match of human zinc-finger 7 (the second zinc-finger of the additional sequence), indicating that after the initial intra-gene segmental duplication there was a secondary tandem duplication of the first zinc-finger. BLAT analysis revealed the additional subsequence is not present anywhere in the orangutan and other outgroup genomes. To reconstruct a parsimonious nucleotide level evolutionary history of *ZNF91*, we constructed a global multiple sequence alignment using PRANK[38] (http://www.ebi.ac.uk/goldman-srv/prank/), which simultaneously aligns the sequences and infers the ancestral sequences using a realistic model of insertion, deletion and substitution evolution. To include the two inferred duplication events in this history we created edited versions of the human, chimpanzee and gorilla sequences with the additional duplicated sequence removed and included, for each species, as two extra input nucleotide sequences, one of the first additional zinc-finger (zinc-finger 6 in the human protein), and the second of the subsequent 6 additional zinc-fingers (zinc-fingers 7–12 in the human protein). As PRANK requires a phylogenetic tree, we supplied a tree that reflects the accepted species phylogeny, but which included the additional duplications branching off after the speciation from orangutans (Extended Data Fig. 6a). There were four amino acid changes in DNA-contacting residues in the relatively short critical time 12–8 Myr after orangutans branched off and before the human–chimpanzee–gorilla split. This together with the duplications mentioned above gives an indication of positive selection. The full multiple species alignment is shown in Supplementary Information File 3.

Reconstructing the evolutionary history of ZNF93. Multiple sequence alignment and sequence analyses (Extended Data Fig. 8a) revealed a deletion of four zinc-finger domains (located between human domains 5 and 6) in the common ancestral great ape lineage after the split with gibbons (deleted in orangutans, gorillas, chimpanzees and humans, but present in gibbons and Old-World monkeys (crab-eating macaques, rhesus macaques, baboons and green monkeys). Domains 5 and 6 (with respect to humans) are identical to each other in the great ape species. Domain 13 (with respect to humans) is missing in Old-World monkeys and is identical to domain 12 in all apes, suggesting that this domain is likely the result of a tandem duplication event that occurred in the ape last common ancestor, after the split with non-ape Old-World monkeys. Domain 17 (with respect to humans) is present in humans, crab-eating macaques and baboons (its presence or absence in rhesus macaques is unknown due to missing data), and missing in green monkeys, gibbons, orangutans, gorillas and chimps. Analysing the nucleotide sequences shows that one nucleotide insertion in the ape common ancestor (with respect to Old-World monkeys) results in an early stop codon and the loss of this domain, and a compensatory deletion of four nucleotides in humans (with respect to apes) nullifies the effect of the previous ape mutation and results in restoration of domain 17 in humans. So human ZNF93 is not like the protein of other apes. The multiple sequence alignments were obtained and validated using MUSCLE[39], MAFFT[40] and PRANK[38] and the ancestral reconstruction was constructed using PRANK. The full multiple species alignment is shown in Supplementary Information File 4.

Phylogenetic analysis and calculation of evolutionary divergence of L1PA3-6030 and L1PA3-6160 subclasses. Fifty sequences of L1PA3-6030, 50 sequences of L1PA3-6160, 3 sequences of L1PA2 and 3 sequences of L1PA4 were aligned by ClustalW in MEGA6 software package[41]. Only full-length L1PAs were selected. For phylogenetic analysis, the sequence downstream of the 129-bp element (L1PA4 and L1PA3-6160), or the corresponding position (L1PA2 and L1PA3-6030) was used to generate phylogenetic trees. Multiple methods were used (Maximum Parsimony, Minimum Likelihood and Minimum Evolution) to generate trees with comparable outcome. The phylogenetic tree generated by the Minimum Evolution method[42] was used to calculate the divergence times for all branching points with the RelTime method[43].

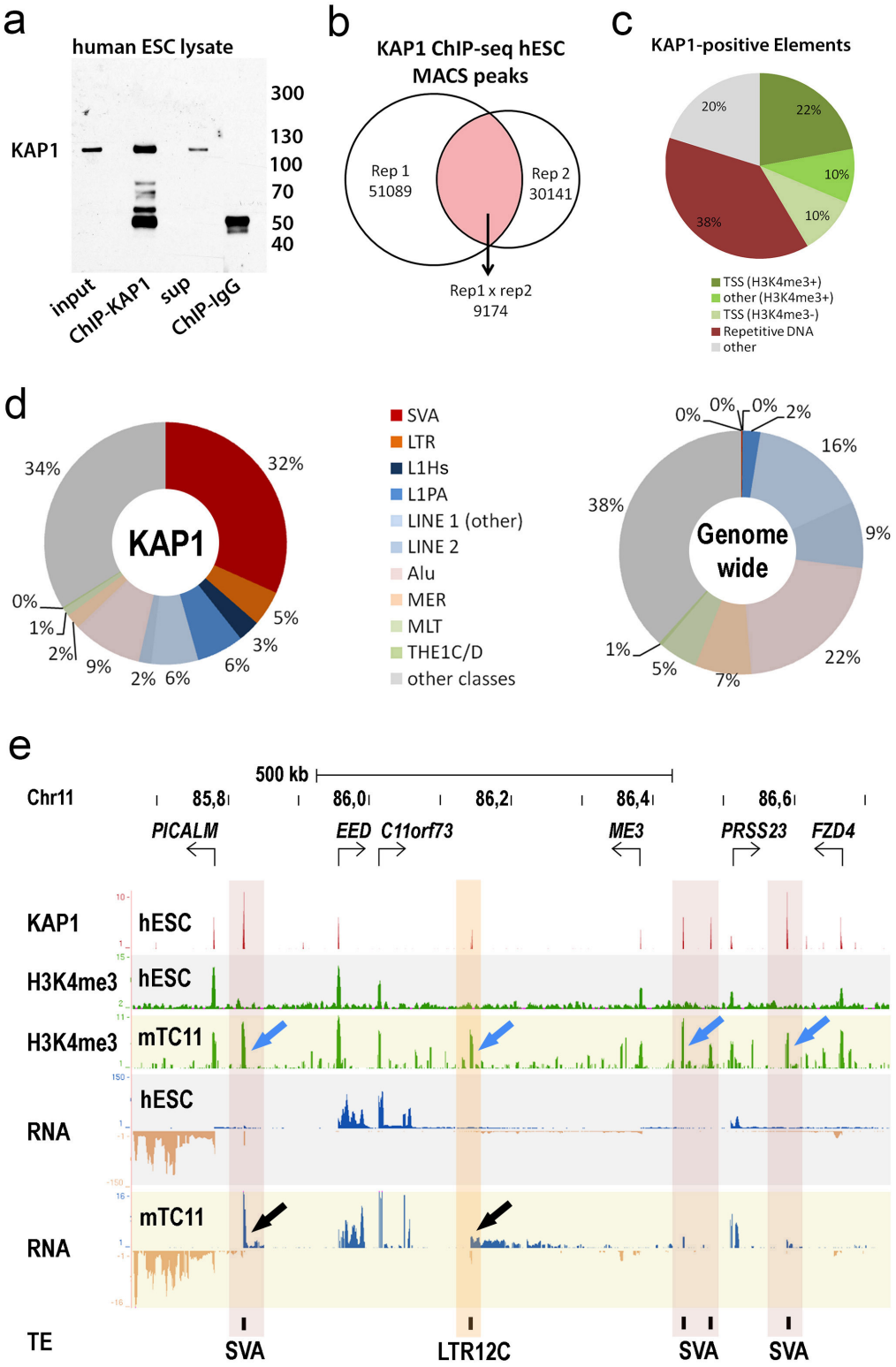
To calculate the average divergence from consensus, first consensus sequences were calculated for L1PA3-6030 and L1PA3-6160 from 150 full-length elements of each subclass using EMBOSS software (http://www.emboss.sourceforge.net/). Each consensus sequence was aligned in MEGA6 with the respective 150 full-length element by ClustalW. In order to be able to compare values for L1PA3-6030 and L1PA3-6160 to divergence values for other L1PA subfamilies, determined previously[9], we used the 500 bp of the 3′ end of the L1PA3 subclasses, and excluded the poly(A)-stretch at the 3′ end of L1PAs. The pairwise distances for each of the 151 (500 bp) sequences (150 individual L1PAs and 1 consensus) were calculated in MEGA6 and plotted in a distance matrix. The average distance (divergence) from consensus was determined by calculating the mean distance (± s.e.m.) from the consensus sequence to each individual L1PA3 element. The age of each L1PA3 subclass was estimated using a base-pair substitution rate of 0.17% per million years (Myr)[9].

VNTR size analysis for SVA-subfamilies. We extracted RepeatMasker SVA elements in the human genome as annotated in the UCSC Genome Browser RepeatMasker track (Hg19.rmsk). Each element was annotated with Tandem Repeat Finder[44] to identify all base pairs covered by a tandem repeat. While VNTR and HEX domains are both tandem repeats, we assumed that the length of the HEX region is a lot shorter and relatively fixed compared to the VNTR, so in the following we use the length of all base pairs masked by Tandem Repeat Finder as a proxy for the length of the VNTR. SVAs annotated by RepeatMasker as multiple adjacent SVA fragments can correspond to a single full-length SVA element. Therefore, to restrict our analysis to unbroken full-length elements, we concentrated on elements that displayed an intact SVA structure, with at least 800 bp of sequence outside of the VNTR region, a size that corresponds to the sizes of Alu and SINE-R combined. For this enriched set of SVAs the histogram of VNTR lengths is plotted in Extended Data Fig. 10c.

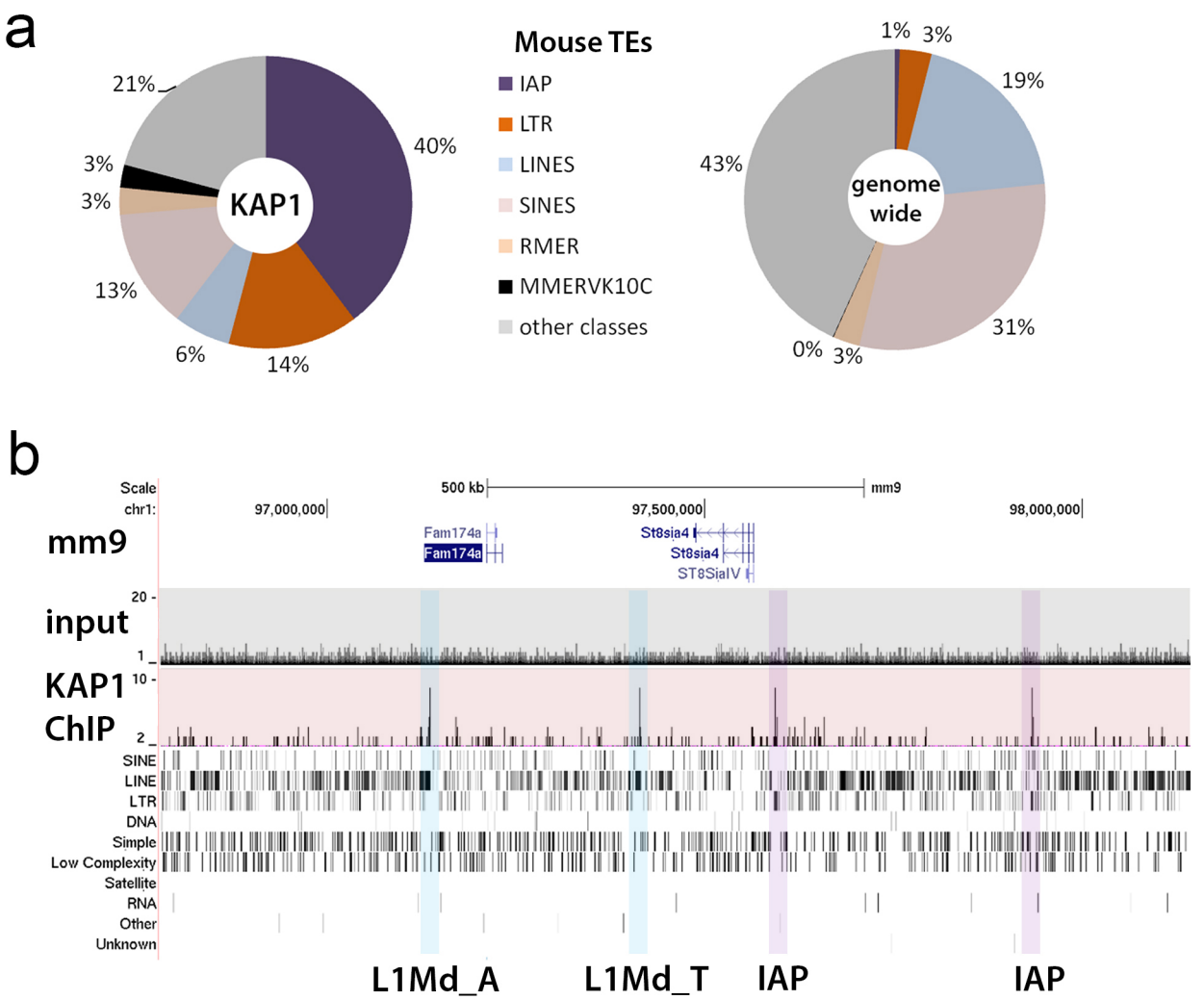
Determination of changes per million years for Fig. 4. For ZNF91 and ZNF93, we counted the numbers of zinc-fingers that have undergone structural changes that could affect DNA binding specificity for each of the evolutionary branchpoints, based on the multiple sequence analysis and ancestral reconstruction (see Methods sections 'Reconstructing the evolutionary history of ZNF91' and 'Reconstructing the evolutionary history of ZNF93'). Changes in DNA binding residues, zinc-finger

deletions or zinc-finger duplications/gains were all weighted equally and counted as '1' because it is unpredictable how each of these changes may change target DNA recognition. The number of changes from one branchpoint to another was divided by the number of million years of that timeframe to determine the number of zinc-fingers that changed per million years. For zinc-fingers in ZNF93 that were different between macaques and gibbons, but conserved between gibbons and great apes, we lacked an outgroup species necessary to determine when the changes occurred. Therefore, to get a rough estimate, we divided the total number of changes between macaques and gibbons, by the amount of time on each of these lineages. From the point of divergence of Old-World monkeys to present-day macaques is 25 Myr, from the point of divergence of Old-World monkeys to the LCA of gibbon and great apes is 7 Myr (25–18 Myr). Therefore we estimated that about 75% of the observed changes happened on the macaque lineage and 25% of the changes on the lineage to the LCA of gibbons and great apes. Similarly, for L1PA elements the consensus sequences of each L1PA element was compared to its direct predecessor and successor, and base-pair substitutions, deletions or insertions were all counted as '1' . The number of base-pair changes per site within the 5′ UTR (1,000 bp) from one L1PA element and its successor was divided by the number of years within the time-frame each L1PA-subfamily was dominant[9]. (See Methods section 'Phylogenetic analysis and calculation of evolutionary divergence of L1PA3-6030 and L1PA3-6160 subclasses') to get the base-pair changes per site per Myr values. For SVA, the percentage of VNTR increase per Myr between SVA-subfamilies is indicated for the timeframe from the emergence of one SVA subfamily to the successor. The average VNTR size for SVA-subtypes as determined in this study (Extended Data Fig. 10c) and the estimated time-points of emergence previously reported for SVA-subfamilies[12] were used to calculate the percentage increase of VNTR size per Myr.

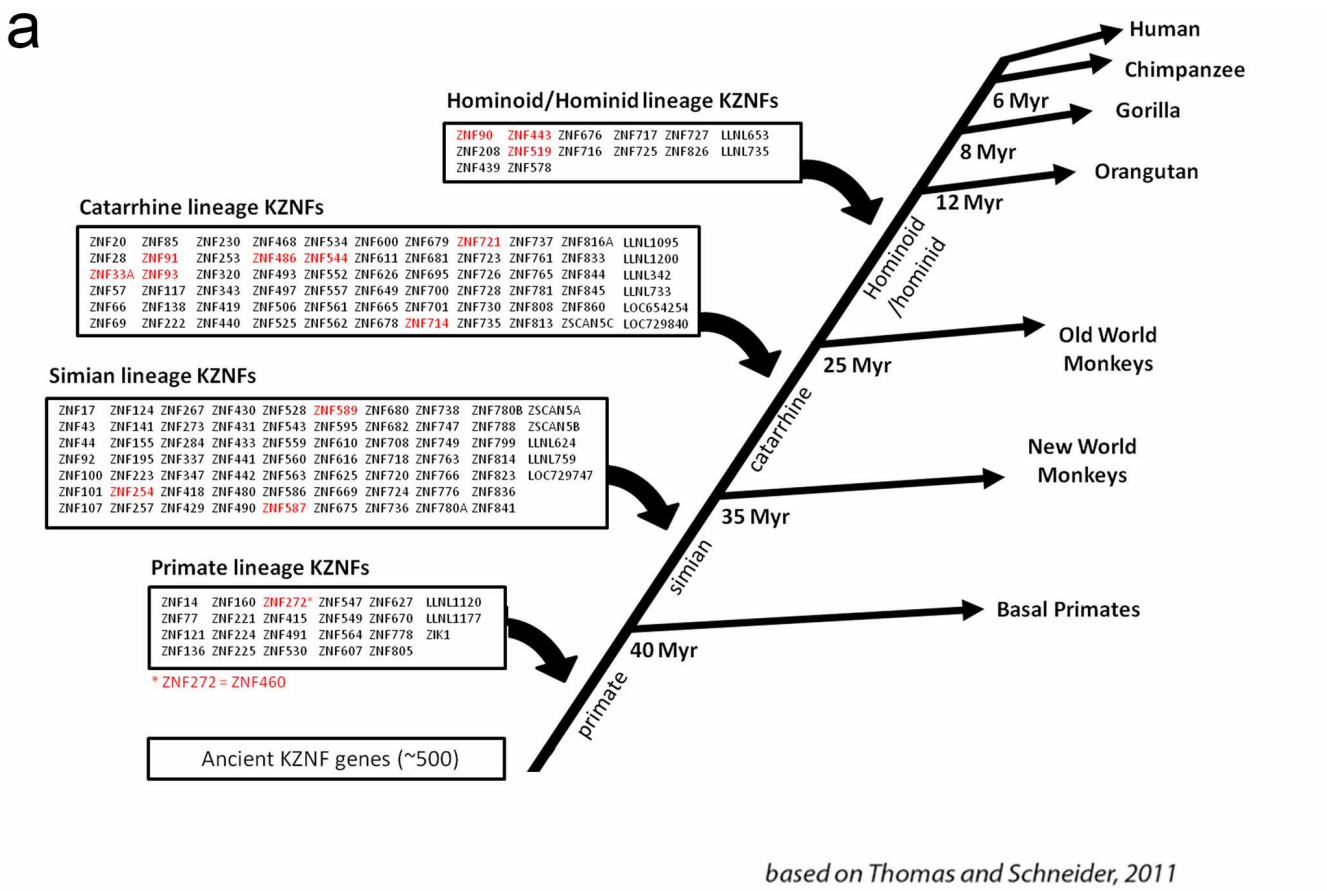
26. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37,** e123 (2009).
27. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
28. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
29. Hsu, F. *et al.* The UCSC known genes. *Bioinformatics* **22,** 1036–1046 (2006).
30. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
31. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
32. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
33. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
34. Onodera, C. S. *et al.* Gene isoform specificity through enhancer-associated antisense transcription. *PLoS ONE* **7,** e43511 (2012).
35. Ying, Q.-L., Stavridis, M., Griffiths, D., Li, M. & Smith, A. Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nature Biotechnol.* **21,** 183–186 (2003).
36. Hancks, D. C., Mandal, P. K., Cheung, L. E. & Kazazian, H. H. The minimal active human SVA retrotransposon requires only the 5′-hexamer and Alu-like domains. *Mol. Cell. Biol.* **32,** 4718–4726 (2012).
37. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
38. Löytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320,** 1632–1635 (2008).
39. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
40. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
41. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).
42. Rzhetsky, A. & Nei, M. A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.* **9,** 945–967 (1992).
43. Tamura, K. *et al.* Estimating divergence times in large molecular phylogenies. *Proc. Natl Acad. Sci. USA* **109,** 19333–19338 (2012).
44. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
45. Naas, T. P. *et al.* An actively retrotransposing, novel subfamily of mouse L1 elements. *EMBO J.* **17,** 590–597 (1998).

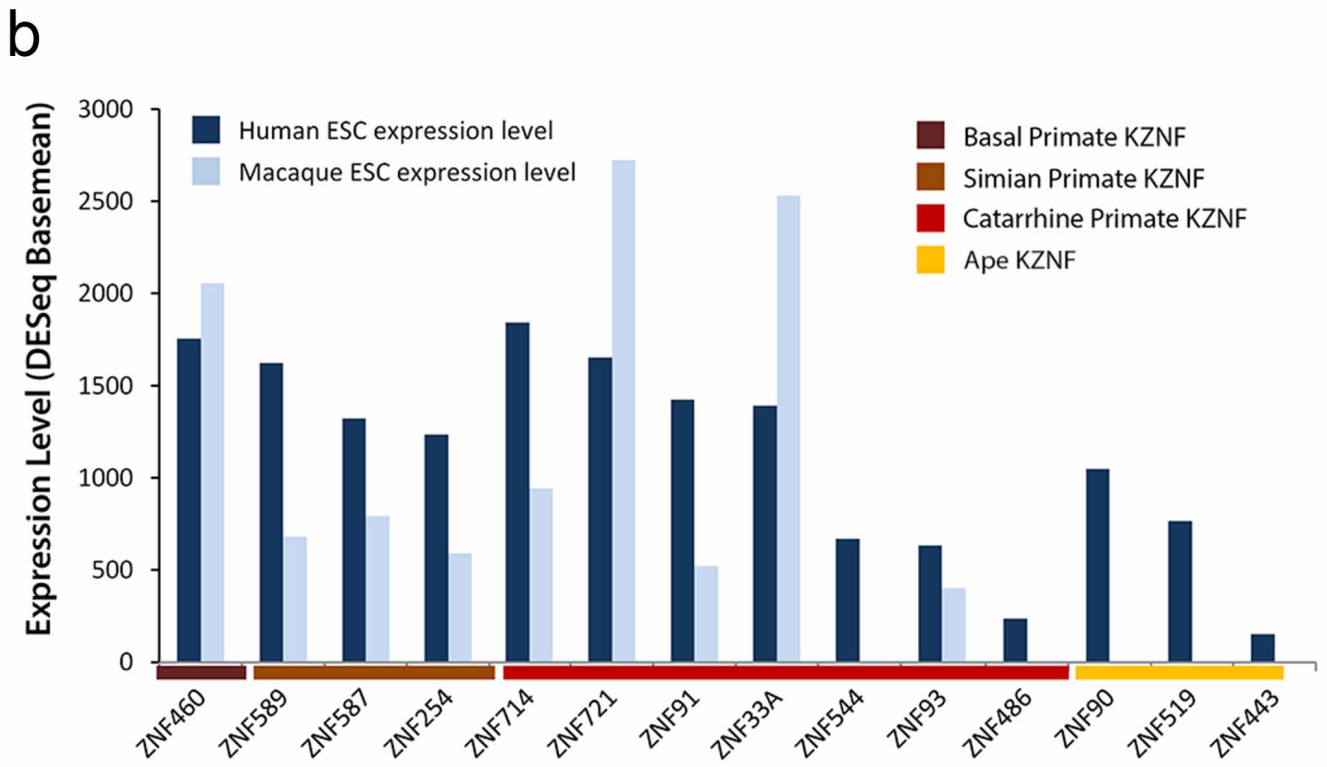


Extended Data Figure 1 | KAP1 associates with recently emerged transposable elements. **a**, Immunoblot incubated with anti-KAP1 antibody loaded with 1% input and eluates of KAP1-ChIP or IgG-ChIP derived from hESC lysates. **b**, Diagram showing numbers of KAP1 peaks identified in two independent biological replicates and common peaks. **c**, Distribution of 9,174 KAP1-ChIP-seq peaks over various DNA elements. **d**, Distribution of retrotransposon classes among KAP1-ChIP peaks from hESCs (left) or genome-wide (right). **e**, KAP1 and H3K4me3 ChIP-seq and RNA-seq coverage tracks for a representative region on human chromosome 11 in hESCs (white- or grey-shaded) and TC11-mESCs (yellow-shaded). Blue arrows, derepressed retrotransposons; black arrows, re-activated transcription; red vertical shading, reactivated SVAs; orange shading, reactivated LTR12C. Blue and tan in RNA-seq tracks indicate positive and negative strand transcripts, respectively. Note that while the majority of SVAs display aberrant H3K4me3 signal, for unclear reasons not all SVAs display aberrant transcription in TC11-mESCs. Rep, biological replicate; sup, supernatant; TSS, transcription start site.

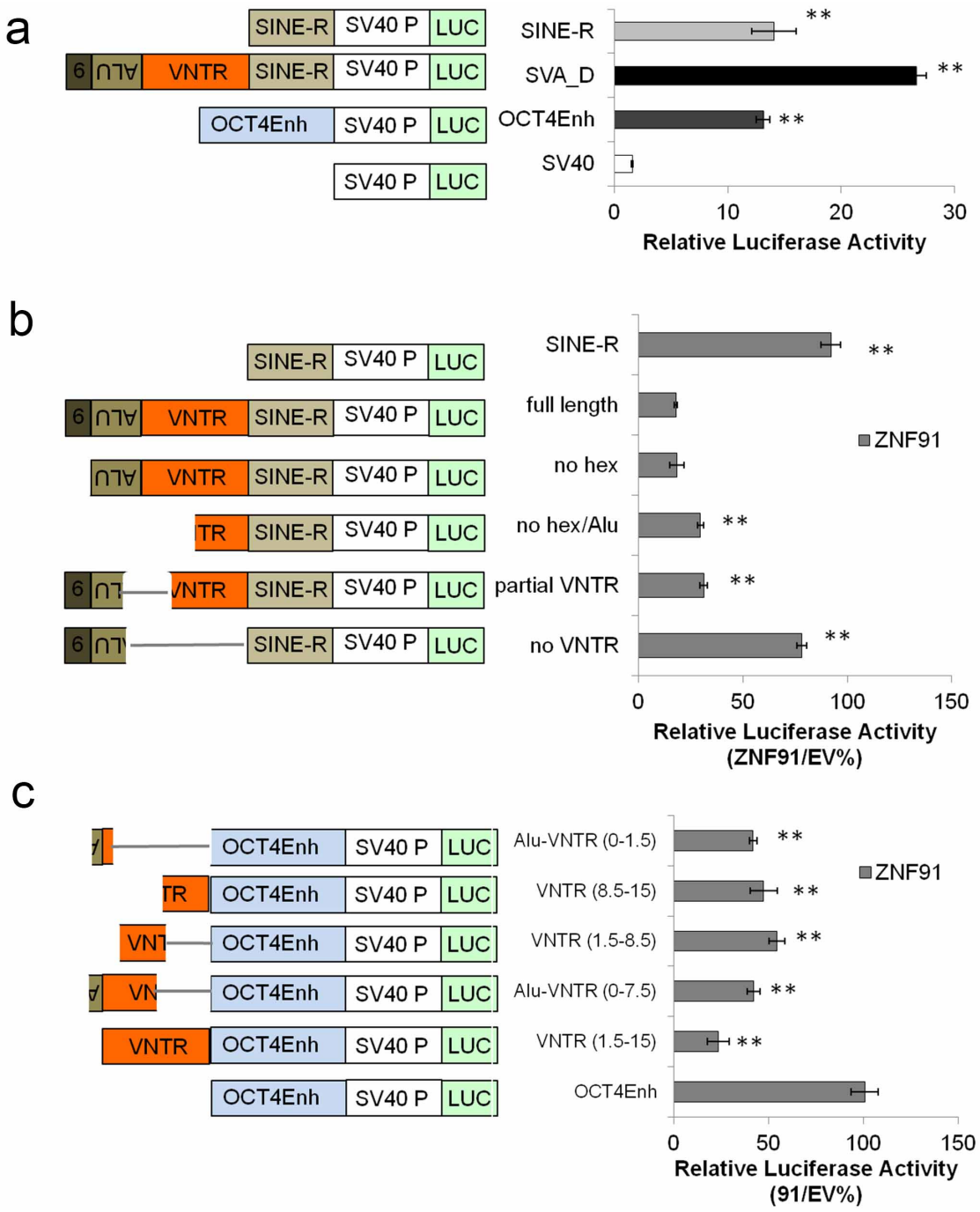


Extended Data Figure 2 | Mouse KAP1 associates with mouse-specific retrotransposons in mouse ESCs. a, Distribution of KAP1-ChIP-Seq reads from mESCs (left) and the mouse genome (right) for retrotransposon families as defined by RepeatMasker (http://www.repeatmasker.org/). **b**, UCSC Browser image displaying ChIP-seq tracks for input (grey shading) and KAP1 (red shading) as well as gene annotation and repeat element tracks for a region on mouse chromosome 1. Blue shading, KAP1-positive active mouse L1-subtypes[45]; purple shading, KAP1-positive active intracisternal A-particle (IAP) retrotransposons. LINES, long interspersed nuclear elements; LTR, long terminal repeat; MMERVK10C, mouse endogenous retrovirus subtype K10C; RMER, medium reiteration frequency repetitive sequence; SINES, short interspersed nuclear elements; TEs, transposable elements.

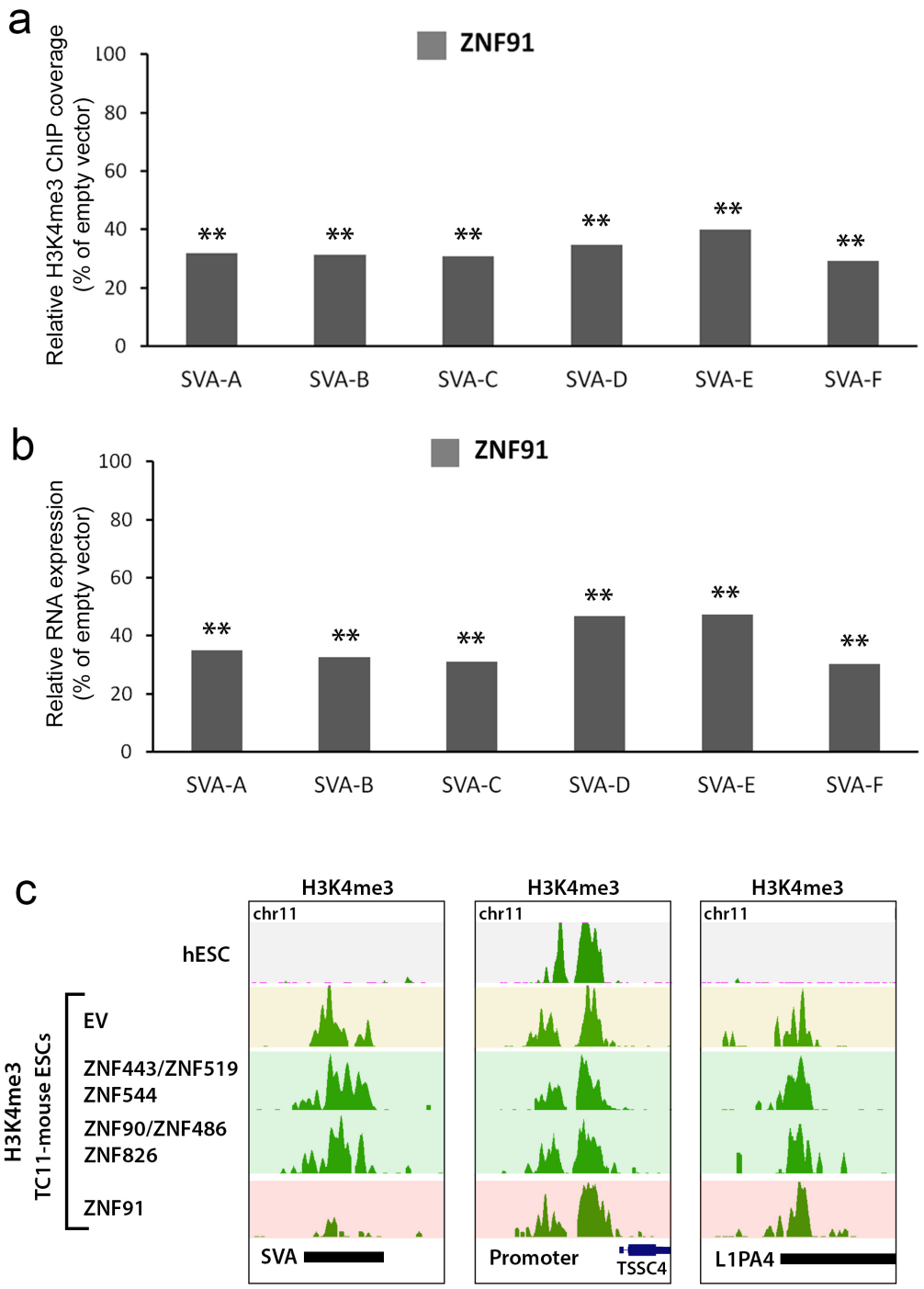




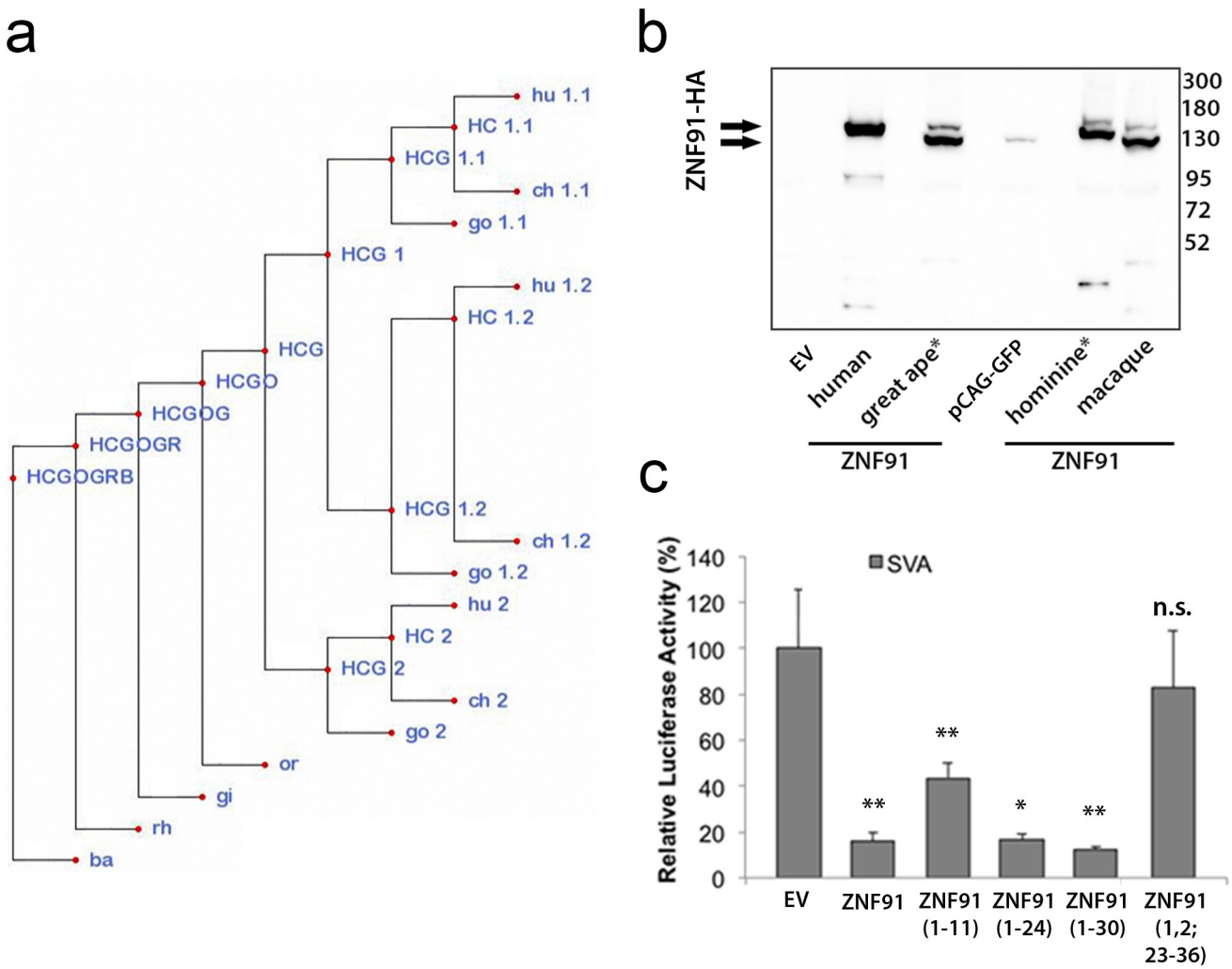
Extended Data Figure 3 | Selection of primate-specific KZNF genes with high expression in hESCs. a, Schematic of primate-specific KRAB zinc-finger genes subdivided in different clades based on previous analysis[7]. KZNFs shown in **b** are highlighted in red. **b**, DESeq-calculated gene expression levels for the 17 highest expressed KRAB zinc-finger genes in hESCs (dark blue) and macaque ESCs (light blue), subdivided by clades.



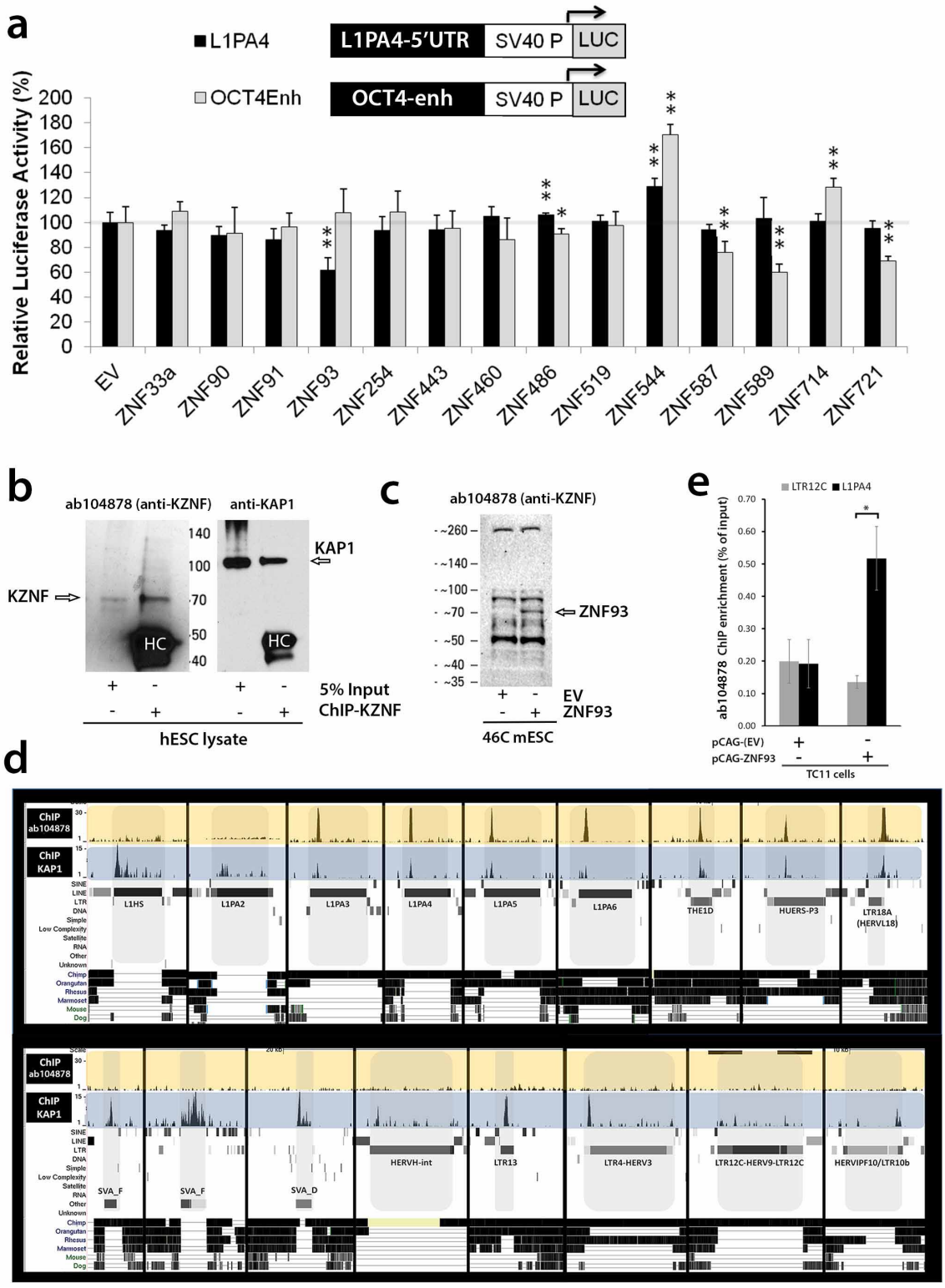
Extended Data Figure 4 | The SVA VNTR domain is necessary and sufficient for ZNF91-mediated repression of luciferase activity. **a–c**, Schematic of SV40–luciferase constructs used (left) and relative luciferase activity after transfection of the indicated constructs in mESCs (right). **a**, SVA and SINE-R are strong enhancers ($n = 6$ biological replicates). **b**, Deletion analysis reveals the VNTR of SVA is required for ZNF91-mediated reporter regulation. Luciferase activity in the presence of ZNF91 expressed as a ratio of that observed for empty vector with the same reporter. Biological replicates: no VNTR, $n = 9$; partial VNTR, $n = 3$; no hex/Alu, $n = 2$; no hex, $n = 2$; full length SVA, $n = 15$; SINE-R, $n = 3$. Empty vector is set to 100% for comparison. **c**, 1.5 VNTR repeats are sufficient to confer ZNF91-mediated regulation on an OCT4Enh–SV40–luciferase-reporter. $n = 3$ biological replicates. $^{**}P < 0.01$; error bars are s.e.m.



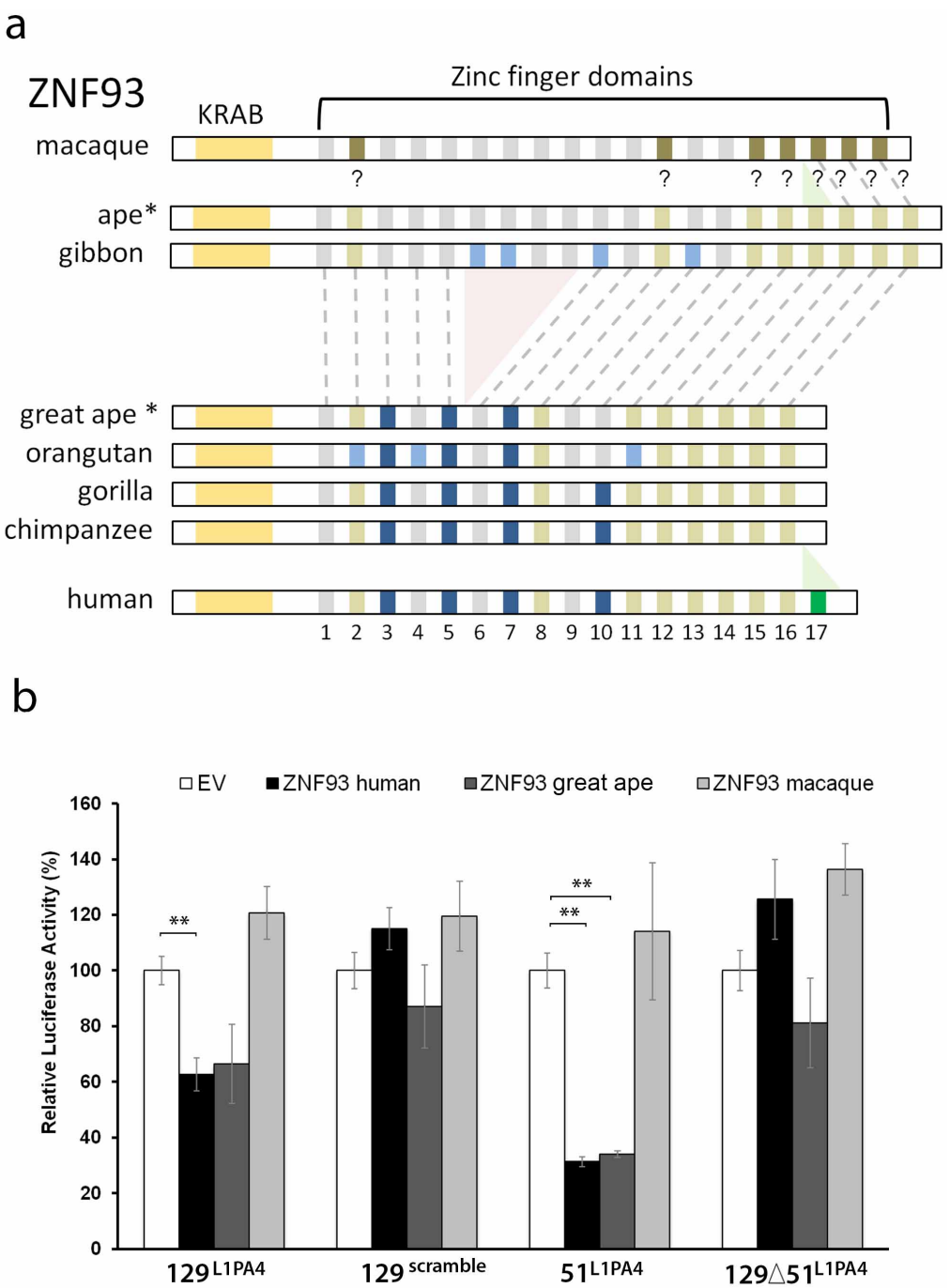
Extended Data Figure 5 | SVA is specifically repressed *in vivo* by ZNF91. **a**, **b**, Normalized DESeq basemean values for H3K4me3 ChIP-seq (**a**) and RNA-seq (**b**) for retrotransposon classes that showed a significant change in ZNF91-transfected TC11-mESCs relative to empty vector. SVAs were the only transposable elements that showed a significant decrease in H3K4me3 and RNA-seq values. **Benjamini–Hochberg adjusted-$P < 0.01$. **c**, UCSC browser images for a representative SVA element, promoter and L1PA4 element, showing H3K4me3 ChIP-seq signal for hESCs (grey), TC11-mESCs transfected with empty vector (yellow), pools of primate-specific KRAB zinc-fingers (green) and ZNF91 (red). TSSC4: tumor-suppressing subtransferable candidate 4.



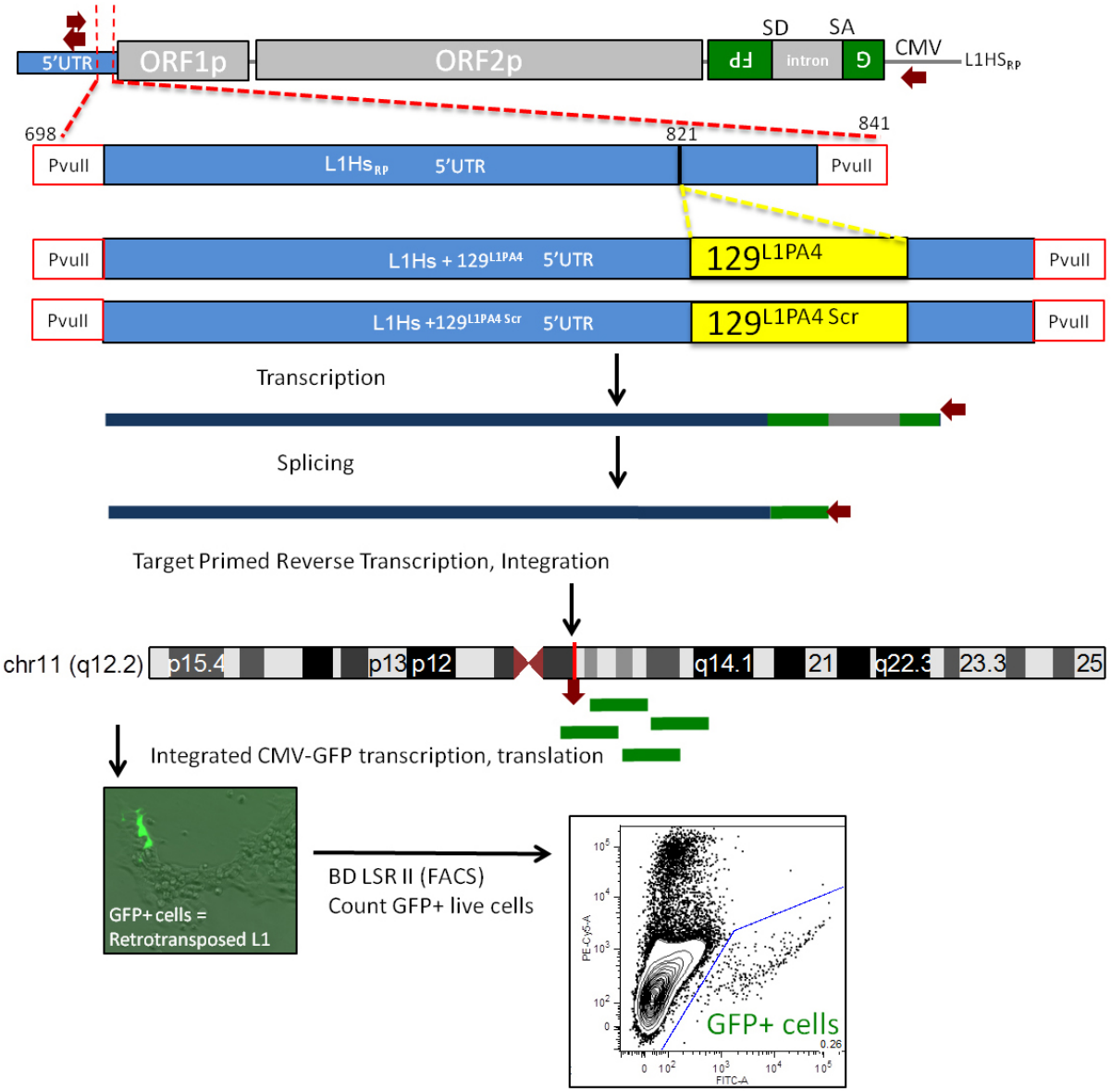
Extended Data Figure 6 | Evolutionary history of ZNF91. **a**, The phylogenetic tree used in multiple sequence alignment and ancestral reconstruction of ZNF91 (Supplementary Information File 3). 'hu 1.1', 'ch 1.1' and 'go 1.1' represent human, chimpanzee and gorilla domain 6, respectively, 'hu 1.2', 'ch 1.2', 'go 1.2' represent human, chimpanzee and gorilla domains 7–12, respectively, and 'hu 2', 'ch 2' and 'go 2' represent the ZNF91 sequence from start to domain 5, a breakpoint, and from domain 13 to the end (see Methods). Ancestors are labelled with first letters of leaf species below them, for example, HCG is a human–chimp–gorilla ancestor. **b**, Immunoblot incubated with anti-HA antibody on lysates of HEK293FT cells transfected with HA-tagged human, great ape, hominine and macaque ZNF91 proteins or lysates transfected with an empty vector and pCAG–GFP. Asterisks denote reconstructed ancestral proteins. **c**, ZNF91 domain deletion analysis showing relative luciferase activities on the SVA-D–SV40 luciferase reporter after transfection of empty vector or ZNF91 deletion constructs in mESCs. Error bars are standard deviation. Numbers in parenthesis indicate zinc-fingers present in the ZNF91 deletion construct. $*P < 0.05$; $**P < 0.01$. Biological replicates: empty vector, $n = 42$; ZNF91 (1–11), $n = 4$; ZNF91 (1–24), $n = 7$; ZNF91 (1–30), $n = 4$; ZNF91 (1, 2, 23–36), $n = 3$.



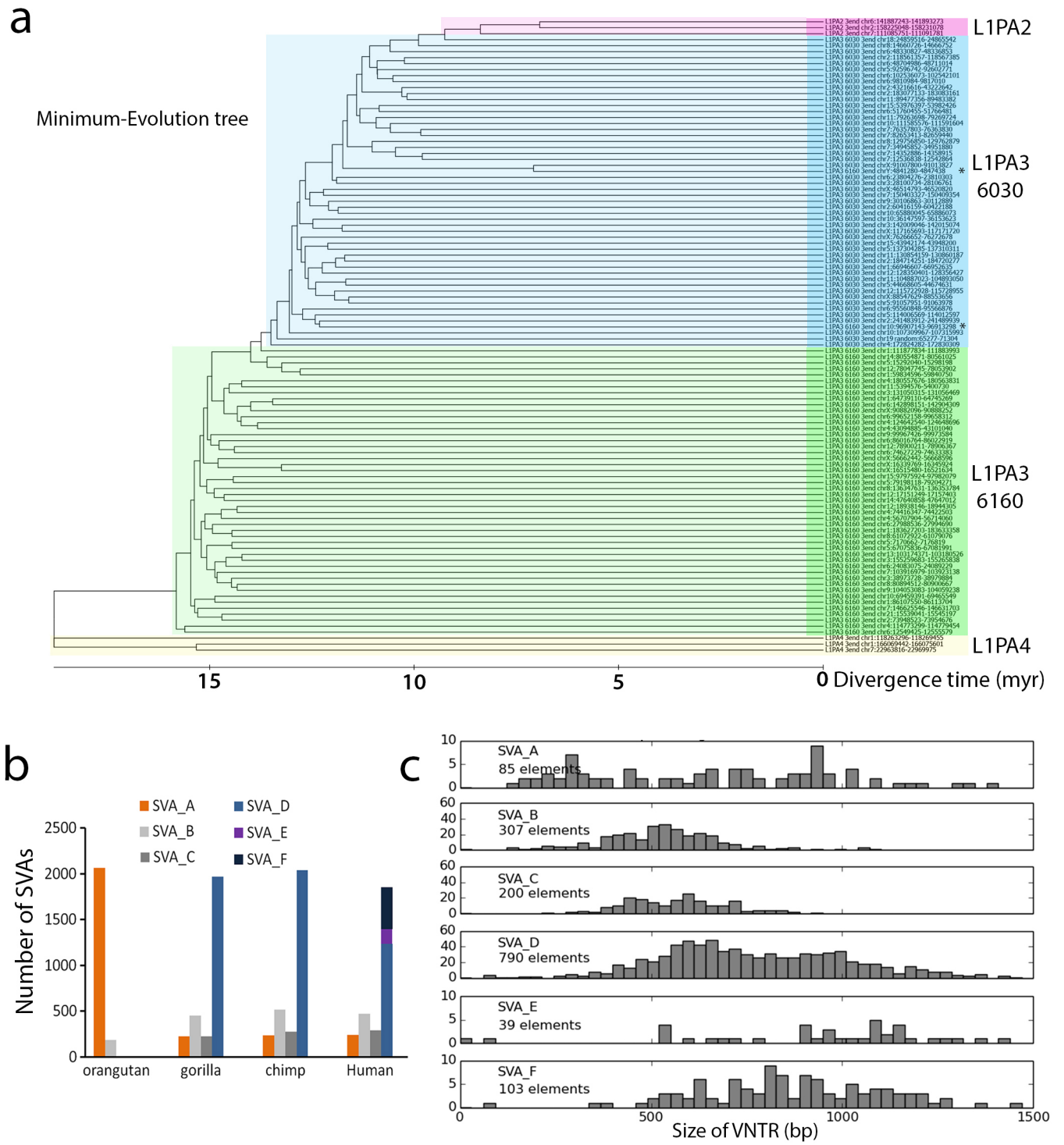
Extended Data Figure 7 | L1PA4 elements are repressed by primate-specific ZNF93. **a**, Relative luciferase activity on a L1PA4– and a OCT4-enhancer–SV40–luciferase-reporter after transfection of 14 KZNFs in mESCs. Significance measured relative to empty vector. $n = 3$ biological replicates; $^{*}P < 0.05$; $^{**}P < 0.01$; error bars are s.e.m. **b**, Immunoblot showing that ChIP with antibody ab104878 predominantly reacts with a protein of ~70 kDa (left panel) and co-immunoprecipitates KAP1 (right panel). HC, heavy chain of IgG. **c**, Immunoblot demonstrating that ChIP with ab104878 detects overexpressed ZNF93 in 46c mESCs as a ~70 kDa protein. **d**, Repeat Browser (see Methods) displaying ChIP-seq coverage tracks for ab104878 (ZNF93; yellow shading) and KAP1 (blue shading) for a selection of KAP1-bound retrotransposons. **e**, ChIP-qPCR for amplicons in L1PA4 and LTR12C elements on chromosome 11 in TC11-mESCs after transfection with an empty vector or ZNF93 and ChIP with ab104878. ChIP enrichment is plotted as percentage of input. $n = 3$ biological replicates; $^{*}P < 0.05$; error bars are s.e.m.



Extended Data Figure 8 | Reconstruction of the evolutionary history of ZNF93. **a**, Schematic based on the multiple sequence alignment of ZNF93 orthologues (Supplementary Information File 4). Red shaded area, deletion of zinc-fingers; green shaded area, gain of zinc-fingers; green stripes, gained zinc-fingers; dark blue stripes, zinc-fingers that changed contact residues in the lineage to humans; light blue stripes, changes in other lineages; brown stripes, zinc-fingers with different binding residues between macaques and gibbons, with gibbons sharing the great ape conformation. For this last group of zinc-fingers, it is unknown (represented with a ? symbol) whether the change happened in monkeys or in the LCA of gibbons and great apes after the divergence of Old-World monkeys (see Methods). Asterisks denote reconstructed ancestral proteins. **b**, Relative OCT4-enhancer–SV40p–luciferase activity for reporters with the indicated L1PA4-derived sequences after co-transfection of an empty vector or various ZNF93 constructs. $^{**}P < 0.01$; error bars are s.e.m.



Extended Data Figure 9 | Schematic of L1Hs retrotranspostion assay. **a**, Schematic of constructs tested indicating the site of 129^{L1PA4} transplant into L1Hs and concept of L1–GFP assay[24] in which GFP expression marks cells where a transfected L1 episome has retrotransposed into a HEK293 cell's chromosomes. ORF, open reading frame; CMV, cytomegalovirus promoter; SD, splice donor; SA, splice acceptor; PvuII, restriction enzyme site.



Extended Data Figure 10 | Evolutionary history of L1PA3-6030, L1PA3-6160 and the VNTR size in SVA. **a**, Phylogenetic tree, rooted on L1PA4, generated using the Minimum Evolution method[42] for fifty 3′-end sequences of L1PA3-6030 and L1PA3-6160, and three 3′-end sequences for L1PA2 and L1PA4. **b**, Bar graphs showing the number of SVA-_A through SVA_F insertions in each great ape genome. **c**, Distribution of VNTR size for untruncated SVA elements in the human genome plotted for each SVA-subfamily. The number of untruncated elements identified for each subtype is indicated.